Generalized additive modelling of mixed distribution Markov models with application to Melbourne's rainfall

Rob J. Hyndman[1] and Gary K. Grunwald[2]

16 September 1999

Abstract: We consider modelling time series using a generalized additive model with first-order Markov structure and mixed transition density having a discrete component at zero and a continuous component with positive sample space. Such models have application, for example, in modelling daily occurrence and intensity of rainfall, and in modelling the number and size of insurance claims.

We show how these methods extend the usual sinusoidal seasonal assumption in standard chain-dependent models by assuming a general smooth pattern of occurrence and intensity over time. These models can be fitted using standard statistical software. The methods of Grunwald & Jones (1998) can be used to combine these separate occurrence and intensity models into a single model for amount. We use 36 years of rainfall data from Melbourne, Australia, as a vehicle of illustration, and use the models to investigate the relationship between the Southern Oscillation Index and Melbourne's rainfall.

[1] Department of Econometrics and Business Statistics, Monash University, Clayton VIC 3168, Australia.

[2] Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center, Denver, CO 80262, USA.

1 Introduction

Time series with a mixed density composed of a discrete component at zero and a continuous component on the positive real line commonly occur with meteorological and environmental data where there may be no recordable level of precipitation or pollutant at some times. They also occur in some business contexts such as insurance claims and non-recurrent expenditure.

Most of the previous discussion about modelling such data has concentrated on modelling daily rainfall occurrence and amounts. We will also use some rainfall data as a vehicle of illustration, although our methods are generally applicable to all such time series with mixed density.

One approach, developed by Stern & Coe (1984) uses GLMs (Generalized Linear Models; see McCullagh & Nelder, 1989) to model rain occurrence (probability) and intensity (amount when it rains). These methods are effective in describing typical rainfall patterns throughout the year, but they assume the same seasonal pattern for each year and thus are not capable of modelling droughts, trends, or other non-seasonal effects. They also result in separate models for occurrence and intensity rather than a single model for rainfall amount.

Recent developments in statistical methodology have made formulation and estimation of more complex models possible. In this paper we use the Generalized Additive Models (GAMs) of Hastie & Tibshirani (1990) to relax the assumption that each year follows the same seasonal pattern, and the Markov models for mixed distributions of Grunwald & Jones (1998) to combine the separate occurrence and intensity models into a single model for amount.

To illustrate our model and estimation methods, we use daily rainfall data from Melbourne, Australia (the Melbourne city station, 86071) for the period 1 January 1963 to 30 September 1998. During this period rainfall was recorded on 39.8% of the 13,057 days. We will show that for this series, rainfall is influenced by several factors including seasonality, drought, and rainfall occurrence and intensity the preceding day.

2 The model

Let Y_t be a random variable denoting the series at time t, $t = 1, \ldots, n$. This is referred to as the *amount process*. The distribution of Y is assumed to be a mixture comprising a discrete component at $y = 0$ and a continuous component for $y > 0$.

Let $q_t(y \,|\, \boldsymbol{X}_t = \boldsymbol{x}_t)$ denote the transition 'density' for Y_t where $\boldsymbol{X}_t$ denotes a vector of covariates including Y_{t-1}, and possibly other explanatory variables. (The term 'density' is used loosely here as the distribution is a mixture of discrete and continuous components.) Following Stern & Coe (1984) and others, we introduce the occurrence and intensity processes to simplify expressions for q.

The *occurrence process* is $J_t = 1$ if $Y_t > 0$ and $J_t = 0$ otherwise. Thus it is an indicator process of whether Y_t is positive. The *intensity process* is defined to be $W_t = Y_t$ if $Y_t > 0$ and missing otherwise. The bivariate random variable (Y_t, J_t) is assumed to be Markov order p.

Thus, J_t has conditional Bernoulli distribution with $\pi_t(\boldsymbol{x}_t) = \Pr(J_t = 1 \,|\, \boldsymbol{X}_t = \boldsymbol{x}_t)$. We use the logit link function so that

$$\pi_t(\boldsymbol{x}_t) = \ell(\mu_t(\boldsymbol{x}_t)) \qquad \text{where} \qquad \ell(u) = e^u/(1+e^u),$$

$$\mu_t(\boldsymbol{x}_t) = \alpha_0 + \sum_{k=1}^{p} (\alpha_k j_{t-k} + g_k(y_{t-k})) + \sum_{i=p+1}^{r} g_i(x_{i-p,\, t}) + g_{r+1}(t),$$

$\boldsymbol{X}_t = (J_{t-1}, \ldots, J_{t-p}, Y_{t-1}, \ldots, Y_{t-p}, X_{1,\, t}, \ldots, X_{r-p,\, t})'$ is a vector of covariates, and each g_i ($i = 1, 2, \ldots, r+1$) is a smooth function. We can generalize this model further by allowing some interaction between the covariates.

Let the intensity process W_t have continuous conditional density $f_t(w \,|\, \boldsymbol{X}_t)$ for $w > 0$ and 0 otherwise. We assume that f_t is a gamma density with mean $\nu_t(\boldsymbol{x}_t)$ and log link function so that

$$\log(\nu_t(\boldsymbol{x}_t)) = \beta_0 + \sum_{k=1}^{p} (\beta_k j_{t-k} + h_k(y_{t-k})) + \sum_{i=p+1}^{r} h_i(x_{i-p,\,t}) + h_{r+1}(t)$$

where each h_i ($i = 1, 2, \ldots, r+1$) is a smooth function. The shape parameter of the density f_t is assumed to be constant for all t and $\boldsymbol{x}_t$. Note that we could replace the gamma density assumption by some other appropriate density such as the log-normal which was used by Katz & Parlange (1995). Or, more generally, we could estimate f_t nonparametrically using the methods of Hyndman, Bashtannyk & Grunwald (1996) and Hyndman & Yao (1998). However, in this paper we shall use the gamma density.

The transition density of Y_t can now be written as

$$q_t(y \mid \boldsymbol{X}_t = \boldsymbol{x}_t) = [1 - \pi_t(\boldsymbol{x}_t)]\delta_0(y) + \pi_t(\boldsymbol{x}_t) f_t(y \mid \boldsymbol{x}_t) \qquad (2.1)$$

where $\delta_0(y)$ denotes a Dirac delta function with support zero. Properties of Y_t such as moments conditional on Y_{t-1} can be found as in Aitchison (1955). We give such results as we use them below.

Following Grunwald & Jones (1998), we shall assume that $\pi_t(\boldsymbol{x}_t)$ and $f_t(w \mid \boldsymbol{x}_t)$ have no common model terms, so that the likelihood admits a simple factorization.

The above model generalizes the model of Grunwald and Jones in several ways. We allow the dependence of Y_t on t and $\boldsymbol{X}_t$ and of J_t on t and $\boldsymbol{X}_t$ to be non-linear and estimate it non-parametrically using a GAM. In particular, we do not assume the same seasonal patterns recur every year, thus providing the facilities to model unusual events (such as droughts if Y_t denotes rainfall).

Note that intervention effects such as changes in measurement or relocation of a recording station, which in Grunwald & Jones (1998) needed to be modelled explicitly using dummy variables, can now be modelled by $g_{r+1}(t)$ and $h_{r+1}(t)$, and need not be included in the model separately. However, these effects will now be included in a smooth form, so if the effect is of real interest in its own right, or if it is expected to be discontinuous in effect, including a separate term may be useful.

One by-product of our model is a natural method for producing seasonally adjusted estimates of probabilities of occurrence and mean intensity.

3 Estimation

Fitting this model requires estimating α_j and β_j ($j = 0, \ldots, p$), and the functions g_i and h_i, ($i = 1, \ldots, r+1$). Since the mixed transition density is not of a standard form, standard methods and software are not available for doing this. However, Grunwald & Jones (1998) show that for GLMs, if it is assumed that there are no common parameters in the occurrence and intensity models, the Markov likelihood function for $\{y_2, \ldots, y_n\}$ conditional on $Y_1 = y_1$, as found from (2.1), factors into separate parts for the occurrence and intensity models. Thus, the overall likelihood is maximized by the estimates of α_j, β_j, g_i and h_i which maximize the occurrence and intensity models separately. The same argument holds for GAMs. Since the occurrence and intensity models do have standard transition densities (binary and gamma respectively) the standard methods and software of GAMs can be used.

Estimation of the functions and parameters in the separate models can be done using GAMs with any nonparametric smoothing method including moving averages, locally weighted polynomials such as loess (Cleveland, Grosse & Shyu, 1992), smoothing splines (Green & Silverman, 1994) or penalized regression splines (Eilers & Marx, 1996). The present implementation of generalized additive modelling in S-Plus allows loess or spline smoothing, and we choose the latter since it is faster computationally. Repetition of various aspects of the analyses using loess does not show any notable differences.

We also need to select the $r+1$ smoothing parameters for each of the occurrence and intensity models. As a guide to selecting these smoothing parameters, we shall use Akaike's Information Criterion (AIC), defined by AIC $=$ deviance $+\ 2k$ where k is the total number of degrees of freedom in the model. We have had mixed success in using the AIC as a bandwidth selection method. In Grunwald & Hyndman (1998) we show that the AIC is optimal or nearly so for selecting the smoothing parameter when smoothing non-Gaussian time series, whereas the Bayesian Information Criterion (BIC) gives extreme oversmoothing (selecting very small degrees of freedom). For fitting GAMs, the BIC also tends to give extreme oversmoothing, and the AIC often suggests reasonable smoothing parameters. However, occasionally the AIC is minimized with smoothing parameters which do not appear to highlight the effect being modelled. Consequently, we use it as a guide rather than as an automatic bandwidth selector. When the AIC suggests reasonable smoothing parameters we use them, otherwise we subjectively select the smoothing parameters to provide a model which highlights the effect of interest.

4 Modelling rainfall occurrence in Melbourne

To simplify the analysis of seasonality, we omitted the 9 leap days from the series, although the leap day data were used as the lagged regressors on March 1 when it followed a leap day. As in Grunwald & Jones (1998), we use the log of previous rainfall values to improve the fit. Specifically, we use $\log(y_{t-j} + c)$ for some $c > 0$. (Without this transformation, a variable bandwidth would be necessary due to the extreme skewness of Y_{t-1}.) For the GLMs fitted by Grunwald and Jones, c was chosen by maximum likelihood to be equal to 0.2. To facilitate comparisons between models, we shall also use $c = 0.2$ in this paper.

We include as a covariate the value of the Southern Oscillation Index (SOI), the standardized anomaly of the Mean Sea Level Pressure (MSLP) between Tahiti and Darwin. Let T_k denote the Tahiti MSLP for month k, D_k denote the Darwin MSLP for month k, and let $\Delta_k = T_k - D_k$ denote the difference. Then the monthly value of the SOI is calculated as $I_k^* = 10(\Delta_k - \mu_k)/\sigma_k$ where μ_k denotes the mean of $\{\Delta_i; i(\text{mod}\,12) = k(\text{mod}\,12)\}$ and σ_k denotes the standard deviation of $\{\Delta_i; i(\text{mod}\,12) = k(\text{mod}\,12)\}$. This is known as the Troup SOI. Figure 1 shows the monthly values between January 1963 and September 1998. There is clearly a lot of random variation in the measurement. We have highlighted

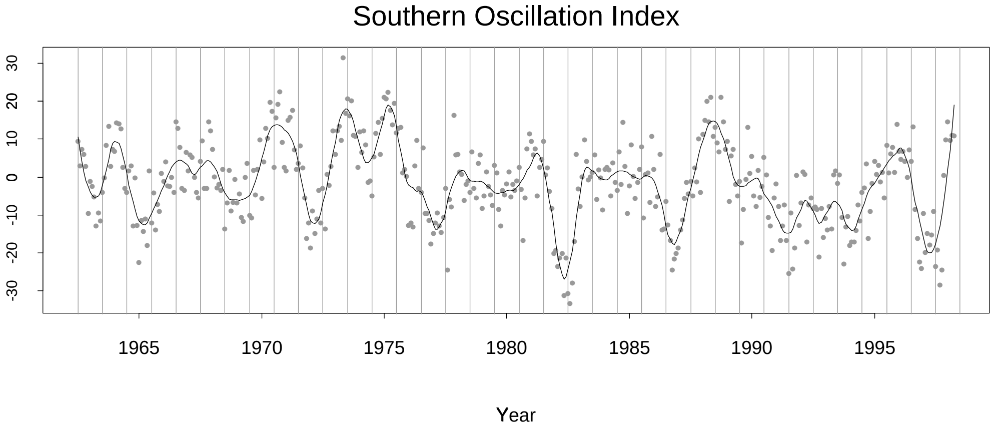


Figure 1: *Monthly Southern Oscillation Index with smooth line highlighting the pattern. The smooth line was computed using a loess curve of degree 2 with span of 6%.*

the underlying trend with a loess curve of degree 2 and span 6%. Negative values of I_k^* indicate 'El Niño' episodes and are usually accompanied by sustained warming of the central and eastern tropical Pacific Ocean, a decrease in the strength of the Pacific Trade Winds, and a reduction in rainfall over eastern and northern Australia. Positive values of I_k^* are associated with stronger Pacific trade winds and warmer sea temperatures to the north of Australia (a 'La Niña' episode). Together these are thought to give a high probability that eastern and northern Australia will be wetter than normal. It should be noted that the effect of the Southern Oscillation is greater in Queensland and New South Wales than Victoria (Allan, Lindesay & Parker, 1996). We define $x_{1,t} = I_t$ to be the value of the fitted loess curve at day t. (Almost identical results are obtained if I_t is calculated by linearly interpolating the raw values of I_k^*.)

We also include the covariate $x_{2,t} = S_t = t(\text{mod}\, 365)$ to model the seasonal variation. The function g_{p+2} is constrained to be periodic; that is, we constrain $g_{p+2}(S_t)$ to be smooth at the boundary between $S_t = 365$ and $S_t = 1$.

Thus our occurrence model has

$$\mu_t(\boldsymbol{x}_t) = \alpha_0 + \sum_{k=1}^{p}(\alpha_k j_{t-k} + g_k(\log(y_{t-k} + c)) + g_{p+1}(I_t) + g_{p+2}(S_t) + g_{p+3}(t).$$

Models with $p = 1, 2, 3$ and 4 were fitted. The results were very similar for all p so we selected $p = 1$ as it simplifies the interpretation.

The smooth term involving I_t was not significantly different from a linear function and so $g_2(I_t)$ was restricted to the linear function $g_2(z) = \alpha_2 z$. Because $g_3(S_t)$ is a periodic function, we model it using a Fourier function of the form

$$g_3(S_t) = \sum_{k=1}^{m} \left[\alpha_{i,s} \sin(2\pi k S_t/365) + \alpha_{i,c} \cos(2\pi k S_t/365)\right],$$

and select the value of m using the AIC. (An alternative approach would be to use a periodic smoother.) The smooth terms $g_1(Y_{t-1}+c)$ and $g_4(t)$ were fitted using smoothing splines. The final model had $\hat{\alpha}_1 = 0.26$ (se 0.07), $\hat{\alpha}_2 = 0.0088$ (se 0.0023), $m = 3$ in $g_3(S_t)$ and smoothing parameters $\text{df}_1 = 4.9$ and $\text{df}_4 = 50$ where df_i denotes the degrees of freedom for the smooth function g_i. The value of df_1 was chosen by minimizing the AIC, while the smoothing parameter for $g_4(t)$ was selected to allow sufficient flexibility to model changes in the probability of occurrence over a period of two or three years.

The value of α_2 is significant (using a t-test at the 5% level). However, if the SOI term is omitted from the model and the other terms re-estimated, the deviance of the model does not change significantly (using a χ^2 test at the 5% level). This anomaly occurs because, if SOI is omitted, the $g_4(t)$ term can model the variation in SOI. We choose to include SOI because we are interested in assessing its effect on rainfall. The term $g_1(y)$ is significantly different from linear ($P = 0.009$ using a χ^2 test). We discuss the significance of the $g_4(t)$ term later.

Figure 2 shows some results for the fitted model. The lower solid line is the estimate of the probability of rain following a dry day ($y_{t-1} = 0$):

$$\Pr(J_t = 1 \mid Y_{t-1} = 0) = \ell(\alpha_0 + g_1(\log c) + g_2(I_t) + g_3(S_t) + g_4(t)).$$

The upper solid line is the estimate of the probability of rain following a day of median intensity (2mm):

$$\Pr(J_t = 1 \mid Y_{t-1} = 2) = \ell(\alpha_0 + \alpha_1 + g_1(\log(2+c)) + g_2(I_t) + g_3(S_t) + g_4(t)).$$

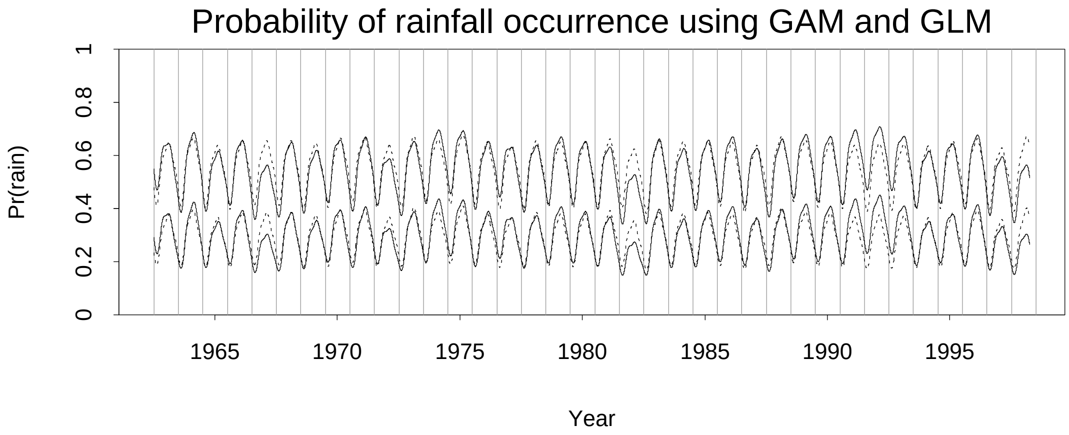


Figure 2: *Lower solid line: estimated probability of rain following a dry day. Upper solid line: estimated probability of rain following a day of median intensity (2mm). These estimates are based on the GAM; dashed lines show analogous curves for the GLM.*

For comparison, analogous curves for a GLM are shown as dashed lines. This model had

$$\begin{aligned}\mu_t(\boldsymbol{x}_t) &= \alpha_0 + \alpha_1 j_{t-1} + \alpha_1^* \log(y_{t-1}+c) + \alpha_2 I_t \\ &\quad + \sum_{k=1}^{3} \left[\alpha_{i,s}\sin(2\pi k S_t/365) + \alpha_{i,c}\cos(2\pi k S_t/365)\right].\end{aligned}$$

Again, the AIC was used to select the number of sinusoidal terms in the seasonal pattern.

Higher order AR models were tried but gave very similar results. Note that the GAM allows the modelling of non-seasonal temporal variation whereas the GLM does not.

We can also look at the probability of rainfall occurrence as a function of the rainfall intensity of the previous day. Figure 3 shows this relationship with all other variables held fixed at the levels observed on two days in the period of the data. The lower curves are for 17 February 1982 (when $g_2(I_t) + g_3(S_t) + g_4(t)$ was minimized). The upper curves are for 20 August 1992 (when $g_2(I_t) + g_3(S_t) + g_4(t)$ was maximized). The solid lines represent the probabilities calculated using the GAM, conditioning on the value of t. The dashed lines show the analogous probabilities as calculated using the GLM.

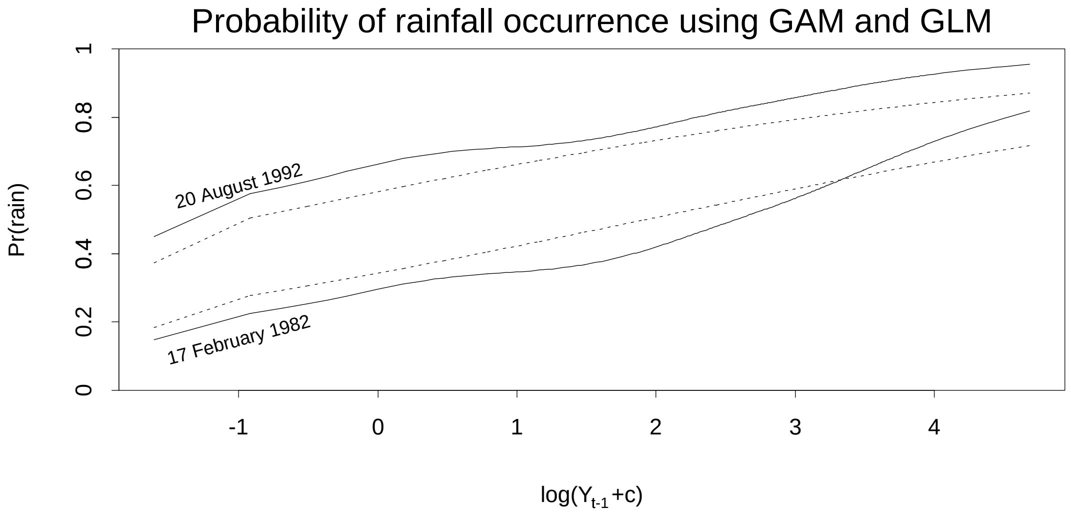


Figure 3: *Lower solid line: estimated probability of rain vs log*$(Y_{t-1}+c)$ *with all other variables held at the levels observed on 17 February 1982. Upper solid line: estimated probability of rain vs log*$(Y_{t-1}+c)$ *with all other variables held at the levels observed on 20 August 1992. Dashed lines show analogous curves for the GLM.*

4.1 Seasonally adjusted occurrence effects

One objective of the GAM analysis is to highlight unusual periods of occurrence, relative to 'typical' annual occurrence patterns. For instance, comparing the GLM and GAM fits in Figure 2 suggests that 1982 had unusually low occurrence and 1990–1993 had unusually high occurrence. To facilitate and quantify such comparisons, we can apply a simple method of seasonal decomposition to decompose $\mu_t(\boldsymbol{x}_t)$ into a seasonal term $s_t(\boldsymbol{x}_t)$ that repeats each year and represents a 'typical' year, and a remainder term $r_t(\boldsymbol{x}_t)$ that represents deviations from this regular pattern. Let $\mu_t(\boldsymbol{x}_t) = s_t(\boldsymbol{x}_t) + r_t(\boldsymbol{x}_t)$ where $s_t(\boldsymbol{x}_t) = s_{t+365k}(\boldsymbol{x}_t)$ for $k = 1, 2, \ldots$. These effects can be interpreted in terms of odds of rain, so that

$$\frac{\Pr(J_t = 1 \,|\, \boldsymbol{X}_t = \boldsymbol{x}_t)}{\Pr(J_t = 0 \,|\, \boldsymbol{X}_t = \boldsymbol{x}_t)} = \exp(\mu_t(\boldsymbol{x}_t)) = \exp(s_t(\boldsymbol{x}_t)) \exp(r_t(\boldsymbol{x}_t)) \,.$$

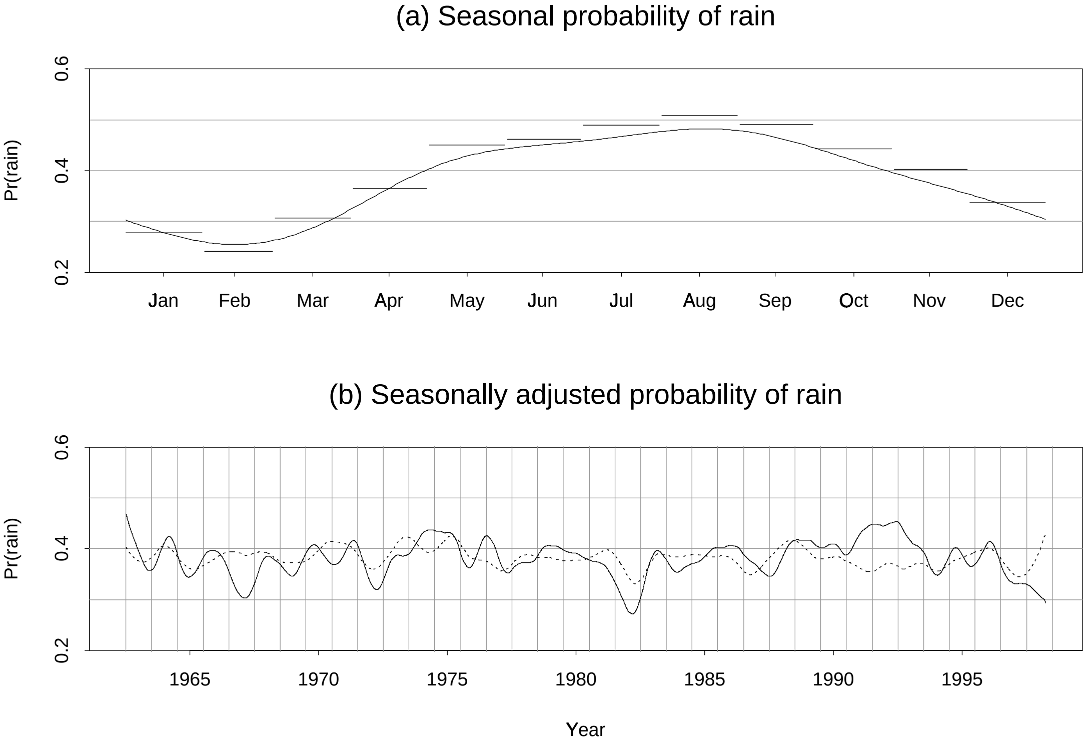


Figure 4: *(a) Estimated seasonal probability of rain, π_t^s. The horizontal bars show the proportion of rainy days for each month during the data period. (b) seasonally adjusted estimated probability of rain, π_t^a. The dashed lines shows the estimated probability of rain further adjusted to show the effect of the SOI.*

Thus $\exp(r_t(\boldsymbol{x}_t))$ represents the factor deviation of the odds of rain from the odds in a typical year, at time t. The seasonally adjusted probability of rain is

$$\pi_t^a(\boldsymbol{x}_t) = \ell(\bar{s}(\boldsymbol{x}_t) + r_t(\boldsymbol{x}_t))$$

where $\bar{s}(\boldsymbol{x}_t) = \frac{1}{365}\sum_{t=1}^{365} s_t(\boldsymbol{x}_t)$, and the seasonal probability of rain is

$$\pi_t^s(\boldsymbol{x}_t) = \ell(s_t(\boldsymbol{x}_t)) .$$

Our model provides a convenient estimate of $s_t(\boldsymbol{x}_t)$. We let

$$\hat{s}_t(\boldsymbol{x}_t) = \hat{\alpha}_0 + \hat{\alpha}_1 \bar{\jmath}_{t-1} + \bar{g}_1 + \hat{\alpha}_2 \bar{I} + \hat{g}_3(S_t) + \bar{g}_4$$

where $\bar{\jmath}$ denotes the mean of j_t, $\bar{I}$ denotes the mean of I_t, $\bar{g}_1$ denotes the mean of

$\hat{g}_1(\log(y_{t-1}+c))$ and $\bar{g}_4$ denotes the mean of $\hat{g}_4(t)$, $t=1,\dots,n$.

Figure 4 shows estimates of the seasonal probability of rain, π_t^s, and the seasonally adjusted probability of rain, π_t^a, plotted against time t. The most striking periods of low occurrence are in 1967, 1972, 1982 and 1998. Apart from the most recent drought, these are exactly the droughts in areas encompassing Melbourne, as reported by Keating (1992). The period of highest probability of occurrence is 1992 (which had the greatest number of wet days of any year in the period studied).

Our model attempts to separate the non-seasonal temporal variation, $g_2(I_t)+g_4(t)$, into two parts: one due to the SOI and one which is unaffected by the SOI. To help visualize the effect of this separation, the dashed lines in the bottom plot of Figure 4 show the probability of rain predicted by the model after seasonal adjustment and removing the effect of $g_4(t)$. That is, we plot

$$\ell(\hat{s}_t(\boldsymbol{x}_t)+\hat{\alpha}_2(I_t-\bar{I})).$$

The resulting curve shows the effect of the SOI on rainfall probability.

The differences between the solid and dashed curves are of interest. For example, in 1967, the solid curve is substantially lower than the dashed curve. This was a period of drought (reflected by the dip in the solid curve) which was not associated with a corresponding low in SOI. The drought of 1982 was associated with the SOI (hence the trough in the dashed curve), but it was more severe than the SOI suggested. Thus, the solid line dips further than the dashed line. The period 1991–1993 is one with unusually high rainfall occurrence that was not associated with a corresponding high in the SOI.

Much of the non-seasonal temporal variation in rainfall probability is being modelled by $g_4(t)$ rather than $g_2(I_t)$. So while the SOI appears to have some effect on the rainfall occurrence it is not a strong predictor and extreme values of the SOI do not always translate into extreme values of rainfall probability in the Melbourne area.

It should be noted that not all of the wiggles in the seasonally adjusted probabilities are

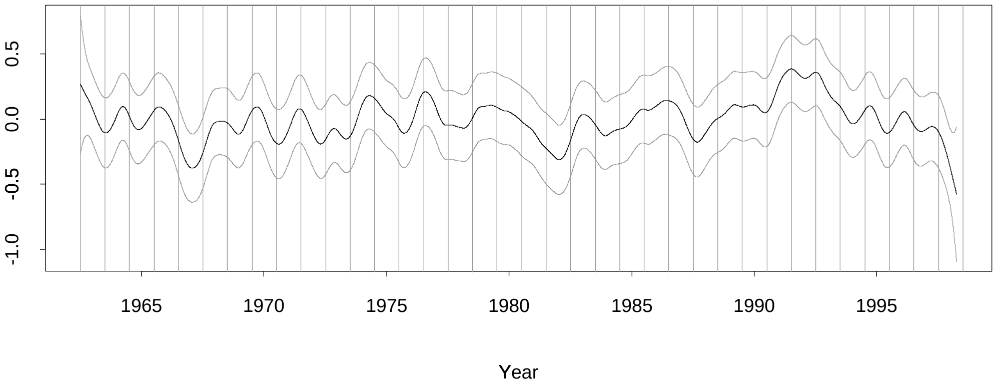


Figure 5: *The curve $\hat{g}_4(t)$ with pointwise 95% confidence intervals calculated using plus and minus twice the standard error.*

likely to be 'real'. To assess the importance of the features of the curve, we calculate pointwise confidence intervals around $g_4(t)$, the greatest contributor to the fluctuations in the seasonally adjusted probabilities. These are shown in Figure 5. The confidence intervals were calculated using plus and minus twice the pointwise standard error of $g_4(t)$, calculated as described in Hastie & Tibshirani (1990, p.60). It seems that the major troughs such as those in 1967, 1982 and 1998 are genuine drops in rainfall probability, whereas some of the smaller wiggles such as those in 1973 and 1995 are probably artifacts of the data.

5 Modelling intensity

Following the same sorts of modelling procedures as we used for the occurrence process, we can construct GLMs and GAMs for rainfall intensity W_t. Recall that $W_t = Y_t$ if $Y_t > 0$, and that we assume it has distribution $G(\nu_t(\boldsymbol{x}_t), r)$ where $G(\nu, r)$ denotes a Gamma distribution with mean $\nu > 0$ and shape parameter $r > 0$. The fitted model had conditional mean

$$\nu_t(\boldsymbol{x}_t) = \exp\left(\beta_0 + \beta_1 j_{t-1} + h_1(y_{t-1} + c) + \beta_2 I_t + h_3(S_t) + h_4(t)\right).$$

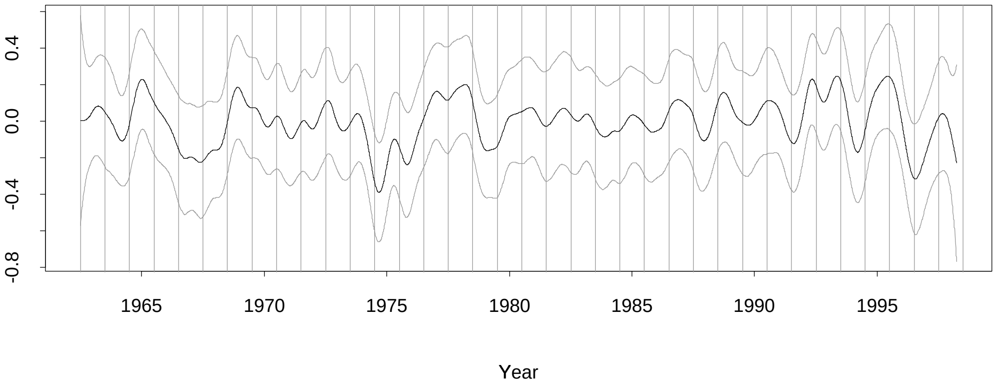
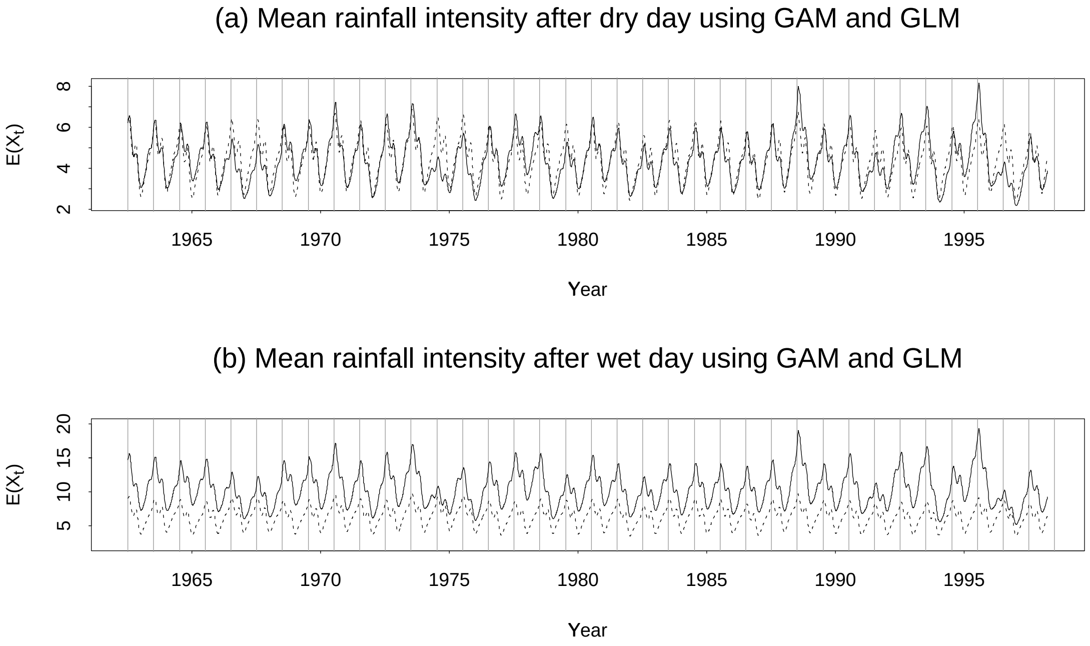


Figure 6: *The curve $\hat{h}_4(t)$ with pointwise 95% confidence intervals calculated using plus and minus twice the standard error.*

(As with occurrence, we also tried higher order autoregressive terms but they made little difference to the fitted models.) The seasonal term h_3 had $m = 4$ and the bandwidths for $h_1(y_{t-1})$ and $h_4(t)$ were 14 and 50 respectively (df$_1$ chosen by minimizing the AIC). The estimated coefficients were $\hat{\beta}_1 = -0.18$ (se 0.07) and $\hat{\beta}_2 = 0.0094$ (se 0.0023). The curve h_1 was significantly different from linear ($P = 0.018$ using an F-test). Figure 6 shows pointwise 95% confidence intervals around $h_4(t)$, indicating that the major 'dips' in rainfall intensity in 1967, 1975–76 and 1996 are significant.

Interestingly, the drought of 1982 does not appear to have affected rainfall intensity—it apparently was an event mainly involving the frequency of rain, not the amount of rain when it did rain. The summer of 1996–1997 had unusually low rainfall intensity, whereas it was not unusual in the frequency of rain (compare Figure 2). While both years were associated with an El-Niño event (indicated by low values of the SOI, see Figure 1), the effect appears to have been different.

Residual analysis for this model is possible using the scaled data $R_t = W_t/\nu_t(\boldsymbol{x}_t)$. Then, if the model is correct, $R_t \overset{d}{=} G(1, r)$. So plots of $\hat{R}_t$ against predictors and QQplots of $\hat{R}_t$ against the quantiles of a $G(1, \hat{r})$ distribution (as discussed in Grunwald & Jones, 1998), will enable some assessment of the fitted model. For the model fitted here, these residual plots did not reveal any model inadequacies.



(a) Mean rainfall intensity after dry day using GAM and GLM

(b) Mean rainfall intensity after wet day using GAM and GLM

Figure 7: *(a) Estimated mean intensity of rain following a dry day. Solid lines calculated from the GAM, dashed lines calculated from the GLM. (b) estimated mean intensity of rain following a day with 30.2mm of rain. Solid lines calculated from the GAM, dashed lines calculated from the GLM. Note that the vertical scales on these two plots are not the same.*

The GLM we used had

$$\begin{aligned}\nu_t(\boldsymbol{x}_t) &= \exp\Big(\beta_0 + \beta_1 j_{t-1} + \beta_1^* \log(y_{t-1}+c) + \beta_2 I_t \\ &\quad + \sum_{k=1}^{4}\Big[\beta_{k,s}\sin(\tfrac{2\pi k S_t}{365}) + \beta_{k,c}\cos(\tfrac{2\pi k S_t}{365})\Big]\Big).\end{aligned}$$

The sinusoidal terms describe the seasonal pattern in rainfall intensity. The number of sinusoidal terms was chosen using the AIC.

Figure 7 shows the mean rainfall intensity $\nu_t(\boldsymbol{x}_t)$ plotted against t for two different values of y_{t-1}. The top plot shows the curve following a dry day ($y_{t-1} = 0$). For comparison, the analogous curve from the fitted GLM is shown. The bottom plot shows the curve where the previous day has rainfall $y_{t-1} = 30.2$mm. This value of y_{t-1} provides the maximum value of $\hat{h}_1$. The GLM curve in the lower plot is clearly biased downwards due to the assumption of a linear relationship with $\log(y_{t-1}+c)$.

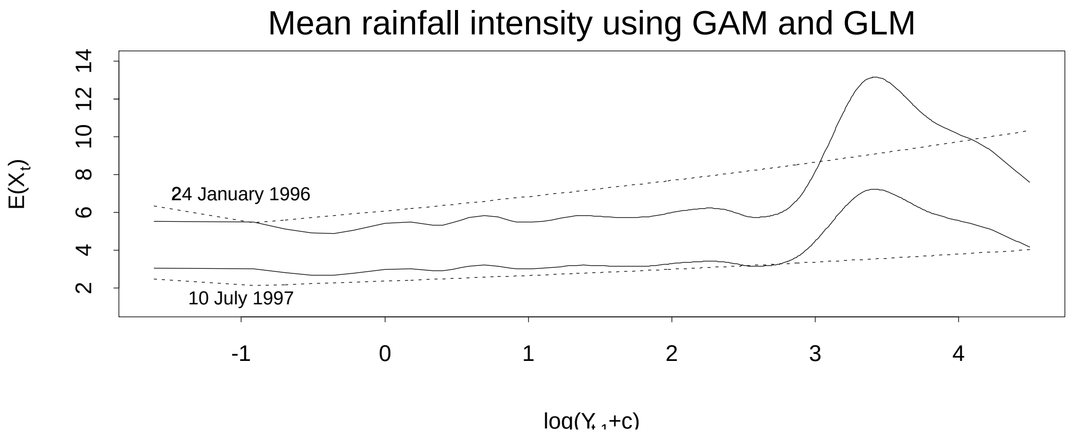


Figure 8: *Mean rainfall intensity (calculated from the GAM) vs log$(Y_{t-1} + c)$ with all other variables held at the levels observed on two days: 24 January 1996 and 10 July 1997. These days were at the maximum and minimum of $\hat{h}_2(I_t) + \hat{h}_3(S_t) + \hat{h}_4(t)$ respectively. Solid lines calculated from the GAM; dashed lines calculated from the GLM.*

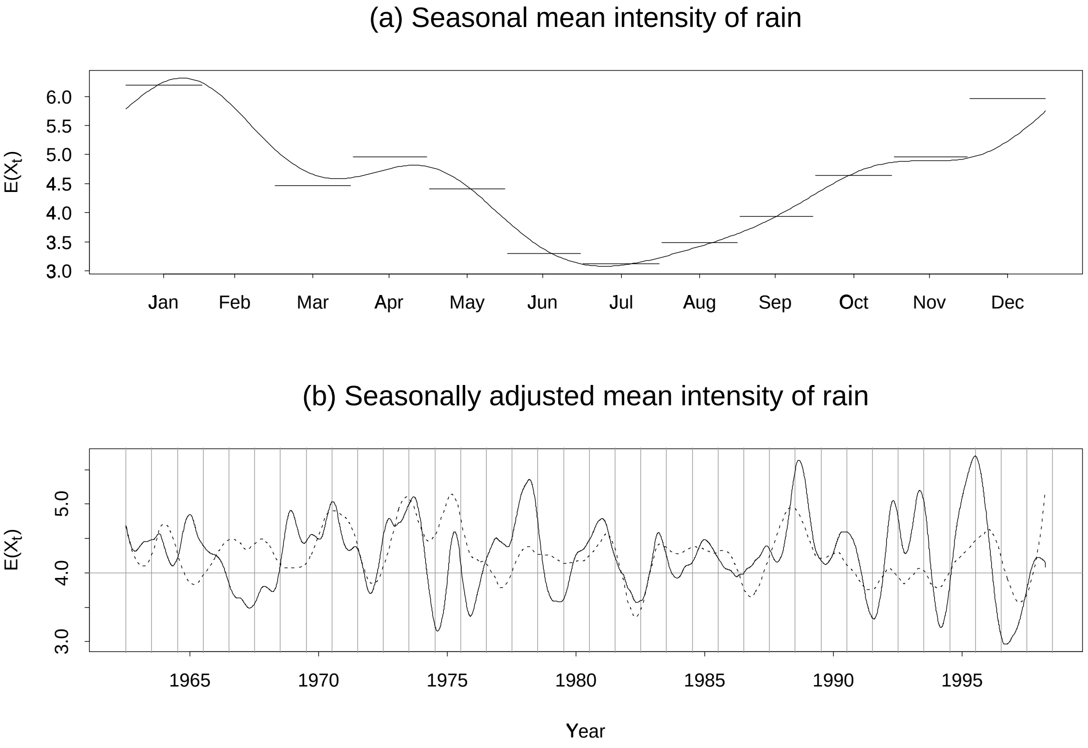


Figure 9: *(a) Seasonal mean intensity. The horizontal bars show the average rainfall intensity for each month during the data period. (b) Seasonally adjusted mean intensity. The dashed lines shows the probability of rain further adjusted to show the effect of the SOI.*

Figure 8 shows the mean rainfall intensity as a function of $\log(y_{t-1} + c)$. The values of the other variables were held fixed at the levels observed on day t where t was chosen to provide the minimum and maximum values of $\hat{h}_2(I_t)+\hat{h}_3(S_t)+\hat{h}_4(t)$. Clearly, the amount of rain on one day, y_{t-1}, has virtually no effect on the amount of rain on the subsequent day, y_t, unless $y_{t-1} > e^3 - c \approx$ 20mm. In other words, there is little autocorrelation in the intensity series unless there is a large rainstorm, in which case it will probably extend into the following day. The clearly non-linear relationship demonstrates why there is bias in the GLM estimate of intensity as seen in Figure 7.

The seasonally adjusted mean intensity is calculated in a similar way to that for probability of occurrence described in Section 4.1. The results are shown in Figure 9. Of the major droughts, not all are clearly identified by low intensity. The droughts of 1967 and 1982 were periods of low intensity but not the drought of 1972. Although the SOI is statistically significant, its effect is small. The major non-seasonal temporal variation in intensity is not associated with the SOI.

6 Markov Generalized Additive Models for rainfall

We now consider combining the occurrence and intensity models of previous sections to give a model for rainfall amount with a mixed density as given in (2.1). In some applications this combined model will be of most interest since it has units of mm/day while intensity has units of mm/wet day. The fitted amount model yields a mixed density which can be summarized in various ways. For instance, we can calculate the mean of Y_t directly from (2.1) as

$$\mathrm{E}(Y_t|\boldsymbol{X}_t = \boldsymbol{x}_t) = \pi_t(\boldsymbol{x}_t)\nu_t(\boldsymbol{x}_t).$$

We are also interested in the marginal mean

$$M_t = \mathrm{E}(Y_t|X_{1,t} = x_{1,t}, \ldots, X_{r-p,t} = x_{r-p,t})$$

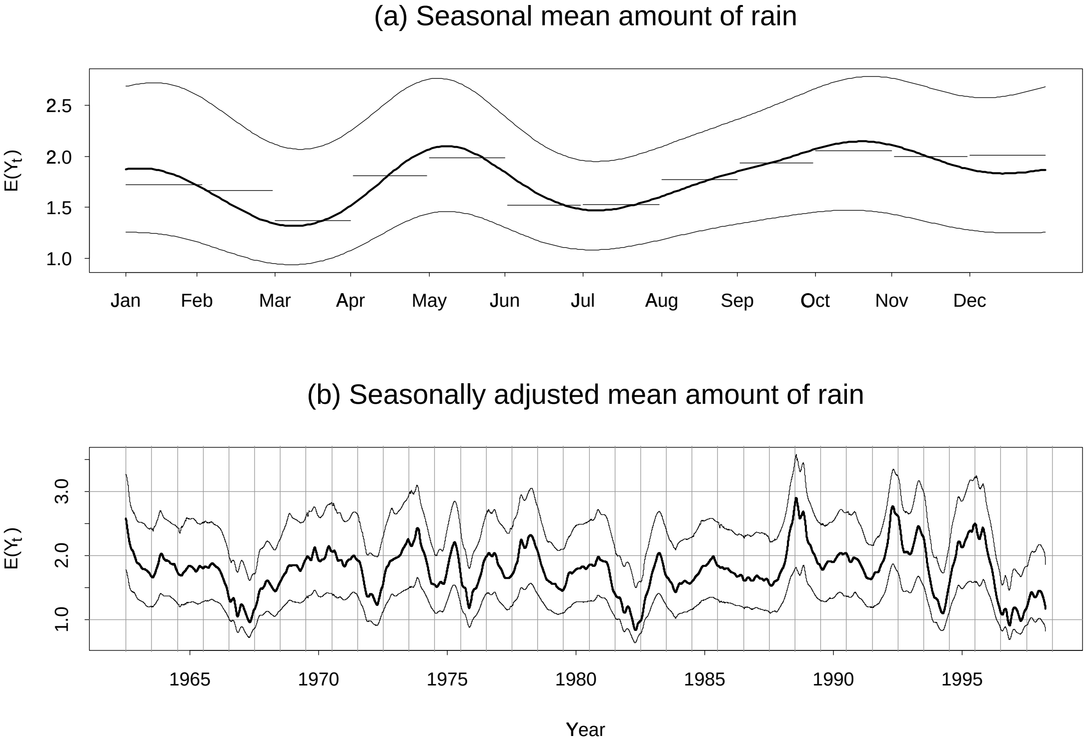


Figure 10: *(a) Seasonal mean amount. The lower curves shows mean conditional on previous day being dry; the upper curves shows mean conditional on previous day having median intensity (2mm); the centre (bold) curves show unconditional means. The horizontal lines show the average amount for each month in the data period. (b) Seasonally adjusted mean amount. Lines as for (a).*

although this is difficult to calculate analytically from the fitted model. Instead we simulate 10,000 sample paths from the model and average across the sample paths to calculate $\hat{M}_t$.

To find the seasonally adjusted mean of Y_t, we let $\hat{s}_t^* = \text{average}(\hat{M}_{t+365k})$ for $t = 1, \dots, 365$ and the average is taken over $k = 0, \pm1, \pm2, \dots$. Then we smooth $\hat{s}_t^*$ using a periodic smoother to obtain s_t, the average rainfall for day t. Finally, the seasonally adjusted rainfall on day t is $r_t = M_t - s_t + \bar{s}_t$. The results are shown in Figure 10 with the seasonal average (s_t) shown in the top plot and the seasonally adjusted values (r_t) shown in the bottom plot.

Note that there is far less variability in the seasonal mean amount than in the seasonal probability of occurrence or seasonal mean intensity. The probability of occurrence is highest in the winter months while the mean intensity is highest in the summer months.

These seasonal patterns largely cancel each other out to give a relatively flat daily mean amount across the year. However, the probability density of rainfall amount varies a lot throughout the year, even though the mean is relatively stable. This is seen, for example, in the conditional distribution functions shown in Grunwald & Jones (1998).

The seasonally adjusted values show that droughts in southern Victoria are more complex than may have previously been understood. Comparing Figures 4, 9 and 10, we note that the drought of 1994, for example, appears to have resulted from lower intensity than usual but that the occurrence was not particularly low for that year. However, the drought of 1982 appears to be more due to low occurrence than low intensity.

Crude estimates of the seasonally adjusted mean curves for amount, intensity and occurrence could be obtained by relatively simple smoothing techniques. However, the modelling approach we have proposed here has enabled us to go much further in estimating curves conditional on past observations, in estimating the effect of the Southern Oscillation Index on both occurrence and intensity, and in decoupling the effects of occurrence and intensity on rainfall amounts.

The purpose of our model has been to investigate the effect of temporal variation and of covariates such as the Southern Oscillation Index on rainfall. We have not attempted to provide an all-purpose model for investigating other aspects of rainfall. For example, our model does not perform well in modeling dry spells. Table 1 shows the actual number of dry spells of different lengths observed in the data with the number predicted by the model. These predictions were obtained by simulating 50 sample paths from the model and computing the number of dry spells of each length for each sample path. The numbers were then averaged to give the values in the table. The standard errors for these predicted values are less than 2.0 in all cases. Including higher order AR terms may lead to better dry spell prediction.

LDS	3	4	5	6	7	8	9	10	11	12	13	14	≥15
Obs	307	239	156	104	81	41	33	30	30	14	14	10	37
Exp	197	162	128	105	89	70	60	49	43	33	31	23	160

Table 1: *LDS = length of dry spell (days). Observed and expected numbers of dry spells over the period of the data.*

However, the model does have implications for long-term forecasting as it attempts to describe the relationship between precipitation and other explanatory variables over a long period of time. To produce such forecasts, we must estimate the conditional mean functions for the occurrence and intensity models for future times.

This can be done by simulating future sample paths from the model, and then averaging these at each time point. Where $x_{i-p,t}$ is not known in advance (such as with SOI), it must be replaced by a forecast or by a given value representing a situation of interest. The non-seasonal temporal variation must also be forecast as the functions $g_{r+1}(t)$ and $h_{r+1}(t)$ are not defined for t beyond the range of the historical data. These can be computed by fitting stationary AR models. This approach has the advantages of (1) being able to model the cyclic fluctuations seen in the models for Melbourne's rainfall, and (2) producing long-term forecasts which converge to the long-term mean (see Makridakis, Wheelwright & Hyndman, 1998).

Until recently, the Bureau of Meteorology used forecasts of the SOI to guide their long-term climate prediction. The analysis presented here suggests that this procedure is not going to yield good prediction for southern Victoria because the relationship between SOI and rainfall is not strong. The method is probably much better for locations in New South Wales and Queensland where the relationship between SOI and rainfall is stronger (Allan, Lindesay & Parker, 1996). However, even there the model presented here will probably lead to better long-term forecasts as it incorporates temporal variation not due to the SOI.

The current practice of the Bureau of Meteorology in long-term rainfall prediction (Drosdowsky & Chambers, 1998) is to use rotated principal components of Sea Surface Temperature anomalies recorded at sites in the Indian and Pacific Oceans. We have not attempted to use these data in our model, but the analysis presented here suggests a new methodology which could be used with these predictors to produce more accurate long-term climate prediction.

The model, as presented here, is less useful for short-term forecasting which would be better handled spatially rather than restricting attention to a single station, and would involve other meteorological covariates. The extension to a spatial network is an open problem.

Acknowledgments

Gary Grunwald and Rob Hyndman were supported by Australian Research Council grants. Much of the work by Gary Grunwald was completed when he was in the Department of Statistics, University of Melbourne. The work by Rob Hyndman was done while he was visiting the Department of Statistics, Colorado State University. We would like to thank Lynda Chambers of the Bureau of Meterology and two anonymous referees for their helpful comments.

References

AITCHISON, J. (1955) On the distribution of a positive random variable having a discrete probability mass at the origin. *J. Amer. Statist. Assoc.*, **50**, 901–908.

ALLAN, R., LINDESAY, J. and PARKER, D. (1996) *El Niño: Southern oscillation and climatic variability*. CSIRO Publications, Melbourne, Australia.

CLEVELAND, W.S., GROSSE, E. and SHYU, W.M. (1992) Local regression models, in J.M. Chambers and T.J. Hastie (Eds.), *Statistical models in S*, Wadsworth and Brooks: Pacific Grove.

DROSDOWSKY, W., and CHAMBERS, L. (1998) *Near global Sea Surface Temperature anomalies as predictors of Australian seasonal rainfall*. Research Report No. 65, Bureau of Meteorology Research Centre, Melbourne, Australia.

EILERS, P.H.C. and MARX, B.D. (1996) Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.*, **89**, 89–121.

GREEN, P.J. and SILVERMAN, B.W. (1994) *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall: London.

GRUNWALD, G.K. and HYNDMAN, R.J. (1998) Smoothing non-Gaussian time series with autoregressive structure. *Computational Statistics and Data Analysis*, **28**, 171–191.

GRUNWALD, G.K. and JONES, R.J. (1998) Markov models for time series with mixed distribution. *Environmetrics*, **10**, in press.

HASTIE, T.J. and TIBSHIRANI, C. (1990) Generalized additive models. Chapman and Hall: London.

HYNDMAN, R.J., BASHTANNYK, D.M. and GRUNWALD, G.K. (1996) Estimating and visualizing conditional densities. *J. Comp. Graph. Statist.* **5**(4), 315–336.

HYNDMAN, R.J. and YAO, Q. (1998) "Nonparametric estimation and symmetry tests for conditional density functions". Working paper, Department of Econometrics and Business Statistics, Monash University.

KATZ, R.W. and PARLANGE, M.B. (1995) Generalizations of chain-dependent processes: applications to hourly precipitation. *Water Resources Research*, **31**, 1331–1341.

KEATING, J. (1992) *The drought walked through: a history of water shortage in Victoria*. Department of Water Resources Victoria: Melbourne.

MAKRIDAKIS, S., WHEELWRIGHT, S. and HYNDMAN, R. (1998) *Forecasting: methods and applications*. John Wiley & Sons: New York.

MCCULLAGH, P. and NELDER, J. (1989) *Generalized linear models*. 2nd ed., Chapman and Hall: London.

STERN, R.D. and COE, R. (1994) A model fitting analysis of daily rainfall data (with discussion). *J. R. Statist. Soc. A*, **147**, 1–34.