Working Paper 01-16
Statistics and Econometrics Series 10
March 2001

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

WEATHER MODELLING USING A MULTIVARIATE LATENT GAUSSIAN MODEL

Durbán, M. and Glasbey, C.A.*

Abstract

We propose a vector autoregressive moving average process as a model for daily weather data. For the rainfall variable a monotonic transformation is applied to achieve marginal normality, thus defining a latent variable, with zero rainfall data corresponding to censored values below a threshold. Methodology is presented for model identification, estimation and validation, illustrated using data from Mynefield, Scotland. The new model, a VARMA(2,1) process, fits the data and produces more realistic simulated series than existing methods dur to Richardson (1981) and Peiris and McNicol (1996).

Keywords: Autocorrelation; Likelihood; Rainfall; Simulation; Vector autoregressive moving average process.

*Durbán, Department of Statistics and Econometrics; Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés (Madrid); e-mail: mdurban@est-econ.uc3m.es; Glasbey, Biomathematis & Statistics Scotland, JCMB, King´s Buildings, Edinburgh, EH9 3JZ, Scotlsnd, UK.

Weather modelling using a multivariate latent Gaussian model

M. Durban* and C.A. Glasbey
Biomathematics and Statistics Scotland
JCMB, King's Buildings, Edinburgh, EH9 3JZ, Scotland

February 23, 2001

Abstract

We propose a vector autoregressive moving average process as a model for daily weather data. For the rainfall variable a monotonic transformation is applied to achieve marginal normality, thus defining a latent variable, with zero rainfall data corresponding to censored values below a threshold. Methodology is presented for model identification, estimation and validation, illustrated using data from Mylnefield, Scotland. The new model, a VARMA(2,1) process, fits the data and produces more realistic simulated series than existing models due to Richardson (1981) and Peiris and McNicol (1996).

Key words: Autocorrelation, Likelihood, Rainfall, Simulation, Vector autoregressive moving average process.

1 Introduction

Weather variables have a significant influence on crop growth, and therefore, it is of interest to have meteorological variables as inputs in most agricultural models. However, long daily records are rare at most agricultural sites and many scientists solve this lack of historical data by using of weather generators, such as WGEN (Richardson and Wright, 1984), LARS-WG (Racsko and Semenov, 1995) or SIMMETEO (Geng et al., 1986). Existing daily weather generators treat rainfall differently from other weather variables, either by simulating it first and then conditionally simulating the remaining variables, or conversely by simulating other variables and then conditionally simulating rainfall (Peiris and McNicol, 1996). In particular, Richardson (1981) simulates rainfall as a Markov chain-exponential model and then the other variables are

*now at: Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Leganés, Madrid 28911, Spain

generated depending on whether the day is wet or dry, whereas Peiris and McNicol (1996) first simulate all variables but rainfall and then use logistic regression to model the probability of a dry day conditional on the other variables.

Many models have been proposed for rainfall, including those based on point processes for the onset of storms (Le Cam, 1961; Rodriguez-Iturbe et al., 1988), those constructed in two stages, first a binary rain/no-rain process and then a gamma distribution applied to wet periods (Richardson, 1981; Stern and Coe, 1984; Katz and Parlange, 1995), and those that apply a monotonic transformation to rainfall data to achieve marginal normality (Bell, 1987; Hutchinson, 1995; Glasbey and Nevison, 1997; Sanso and Guenni, 1999). This last approach defines a latent Gaussian variable, with zero rainfall data corresponding to censored values below a threshold, and simplifies the joint modelling of rainfall and other weather variables. Here we extend this method to joint modelling of rainfall and other weather variables, such as temperature, radiation, wind speed and relative humidity.

In §2 we fit a vector autoregressive moving average process to daily weather at a single site, estimating the parameters by minimising the sum of squares of differences between the expected and sample cross-correlations at a range of time lags (Glasbey and Nevison, 1997; Glasbey et al., 1998). Then, in §3 we simulate from the model and compare the results with those from the original data and from simulations of the models in Richardson and Wright (1984) and Peiris and McNicol (1996). Finally, we discuss the results in §4.

2 Model identification and estimation

We followed Peiris and McNicol (1996) in modelling six daily weather variables at Mylnefield, Scotland. A detailed description of the data is given in Peiris and McNicol (1996). There are 20 years of data, but we omitted the last 3 years from our analysis because of abnormalities in radiation measurements. We use $y_k(t)$ to denote the value of variable k at time t, for $k = 1, \ldots, K$ and $t = 1, \ldots, T$, where, in our case, $K = 6$ and $T = 365 \times 17$. The variables were: y_1 for maximum temperature, y_2 for minimum temperature, y_3 for log-transformed solar radiation, y_4 for wind speed, y_5 for relative humidity, and y_6 for rainfall, after a transformation. A log-transformation was sufficient to normalise the distribution of solar radiation, but rainfall needed something more complicated.

Daily UK rainfall is clearly a non-Gaussian variable as its distribution has a peak at zero and a long upper tail. We followed the approach of Glasbey and Nevison (1997) and used the quadratic power relationship

$$y_K(t) = \begin{cases} \alpha_0 + \alpha_1 r(t)^\gamma + \alpha_2 r(t)^{2\gamma} & \text{if } r(t) > 0 \\ * & \text{otherwise} \end{cases} \qquad \text{for} \quad t = 1, \ldots, T. \tag{1}$$

as an analytically-invertible monotonic transformation, where $r(t)$ denotes rainfall on day t and '$*$' denotes a censored value when rainfall is zero. Figure 1 shows the normal probability plot for the 17 years of daily rainfall data. Superimposed in Figure 1 is the least squares fit of (1), with $\hat{\alpha} = (-0.053, 0.529, -0.027)$, $\hat{\gamma} = 0.597$. This transformation fits the data well, except for

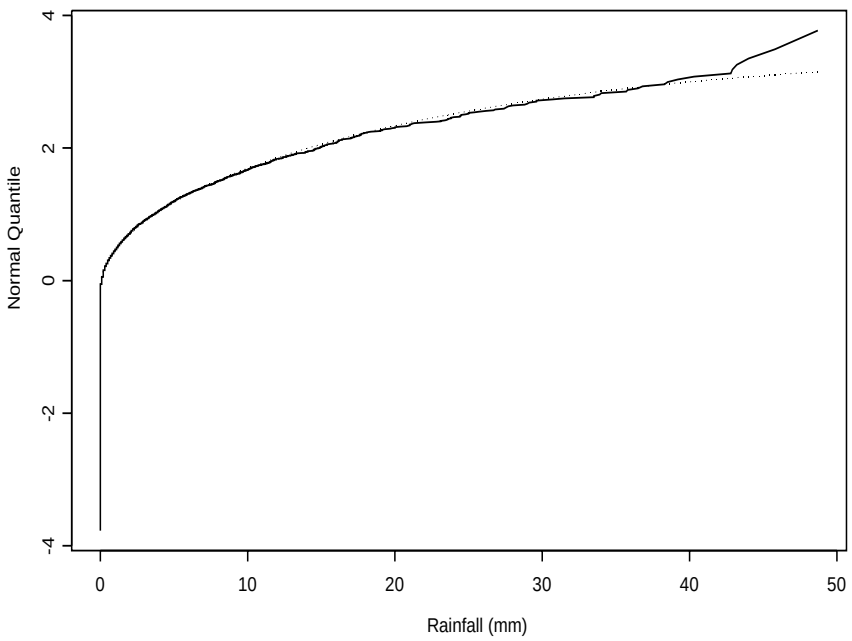

Figure 1: *Normal probability plot for 17 years of daily rainfall data for Mylnefield, Scotland (- - -), and the fitted curve, a quadratic function of power-transformed rainfall (—).*

values of rain over 42.8mm, however this value was exceeded only on four occasions in 17 years. Figure 2 illustrates the transformation.

All six weather variables presented a trend due to annual cyclic patterns, which we accounted for by finite Fourier series:

$$y_k(t) \sim N(\mu_k(t), \sigma_k^2), \quad \text{where} \quad \mu_k(t) = \beta_{k0} + \sum_{j=1}^{J} \beta_{kj} \cos\left(\frac{2\pi j t}{365\frac{1}{4}} + \theta_{kj}\right) \quad k = 1, \ldots, K; t = 1, \ldots, T. \tag{2}$$

We estimated parameters β and σ^2 by maximum likelihood and used likelihood-ratio tests to select the value of J. In all cases, for these data, $J = 1$ or 2. Table 1 shows the results. In the case of rainfall, we do not know the value of the variable below the threshold, therefore, ordinary likelihood cannot be used to estimate the parameters in the equation above. A modified version of the likelihood was used instead, as described in Appendix A.

A vector autoregressive moving average (VARMA) was used to model jointly all detrended weather variables, denoted z, where

$$z_k(t) = \frac{y_k(t) - \mu_k(t)}{\sigma_k}.$$

The general form of a VARMA process of order (p, q) is:

$$z(t) = A_1 z(t-1) + \ldots + A_p z(t-p) + e_t - M_1 e(t-1) - \ldots - M_q e(t-q) \tag{3}$$

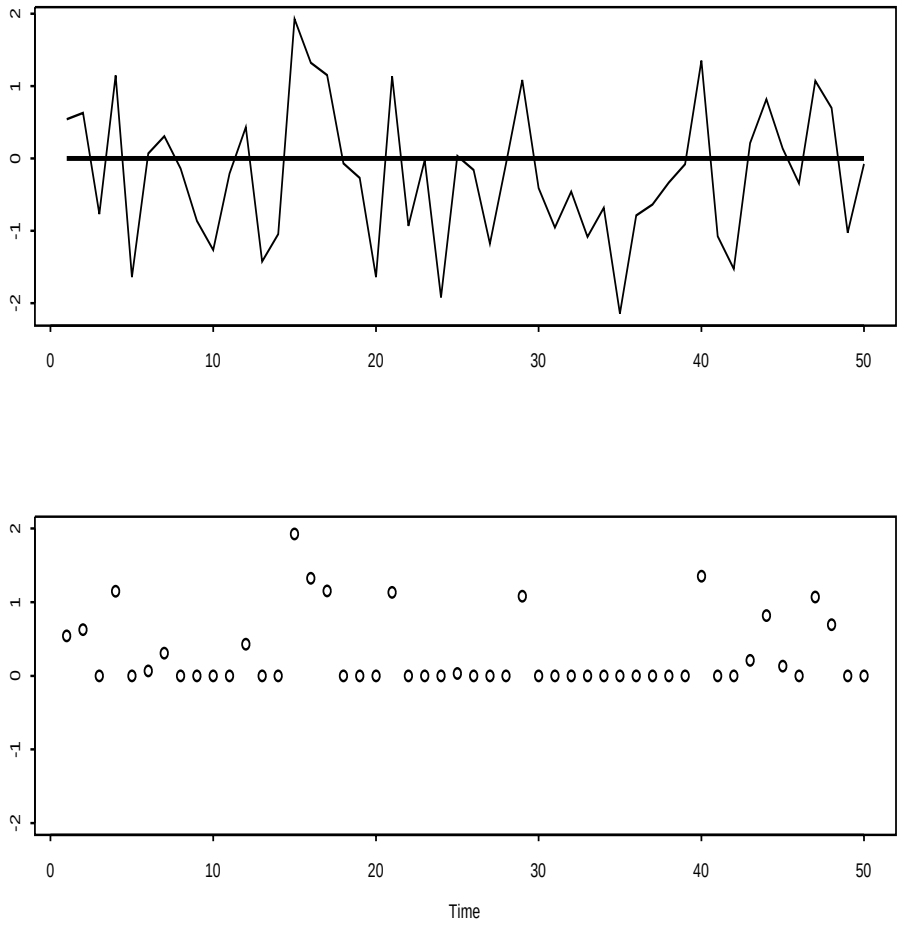

Figure 2: *Plot of the latent variable with threshold at zero (top) and the corresponding values for rainfall (bottom).*

k	variable	$\hat{\beta}_{k0}$	$\hat{\beta}_{10}$	$\hat{\theta}_{1k}$	$\hat{\beta}_{20}$	$\hat{\theta}_{2k}$	$\hat{\sigma}_k$
1	Daily max. temp.	11.72	6.72	2.75			2.87
2	Daily min. temp.	5.03	−5.28	−0.48	−0.55	1.84	2.98
3	$\log_e$(Radiation+1)	1.16	0.79	9.63	−0.05	−0.45	0.75
4	Relative humidity	78.55	7.32	0.21	1.75	−1.06	10.79
5	Wind speed	3.37	0.43	−0.61			2.22
6	Rainfall	0.0	0.13	0.31	−0.04	2.42	1.00

Table 1: *Parameter estimates for annuals trends*

L	$10^5\times$ m.s.e.
3	412
4	421
5	416
6	279
7	**195**
8	213
9	312
10	307
11	418
12	522
13	491
14	521

Table 2: *Mean square error of parameter estimates in VARMA(2,1) model when L correlation terms are used.*

where $z(t) = (z_1(t), \ldots, z_K(t))'$ is a $(K \times 1)$ random vector, A_i and M_i are fixed $(K \times K)$ coefficient matrices and e is a K-dimensional white noise process with $e \sim N(0, \Sigma)$. The parameters in the latent Gaussian model are estimated by an ad hoc procedure previously used by Glasbey and Nevison (1997): we minimise the sum of squares,

$$\sum_{i=1}^{K}\sum_{k=1}^{K}\sum_{l=0}^{L}(\hat{\rho}_{ik}(l) - \rho_{ik}(l))^2, \tag{4}$$

where $\hat{\rho}_{ik}(l)$ and $\rho_{ik}(l)$ are the sample and expected cross-correlations between series i and k at lag l (for details on the calculation of the expected cross-correlations, see Lütkepohl, 1991). The sample cross-correlations between rainfall and the other weather variables cannot be calculated directly by ordinary maximum likelihood. In Appendix B we explain how these were obtained. Parameters were estimated using different numbers of lags, L, and a simulation study was performed to choose the optimal number of lags for a particular model. The lag chosen was the one that minimised the mean square error of parameter estimates. Table 2 shows the mean square errors for the VARMA(2,1) model, indicating that $L = 7$ is optimal in this case.

The order of the process (p, q) was determined by comparing the autocorrelations with 95% confidence intervals obtained by simulation, for a range of values of p ($p = 1, 2$) and q ($q = 0, 1$). Figure 3 shows the results for $(p, q) = (2, 1)2$, the model of lowest order which was acceptable on this criterion. This order of VARMA process was also chosen in Peiris and McNicol (1996), but with rainfall omitted. The simulation study chose $L = 7$ as the optimal number of lags to be used in the estimation of the parameters in the VARMA(2,1) model for these data. The estimated matrices A_1, A_2, M_1 and Σ in model (3) are given in Table 3.

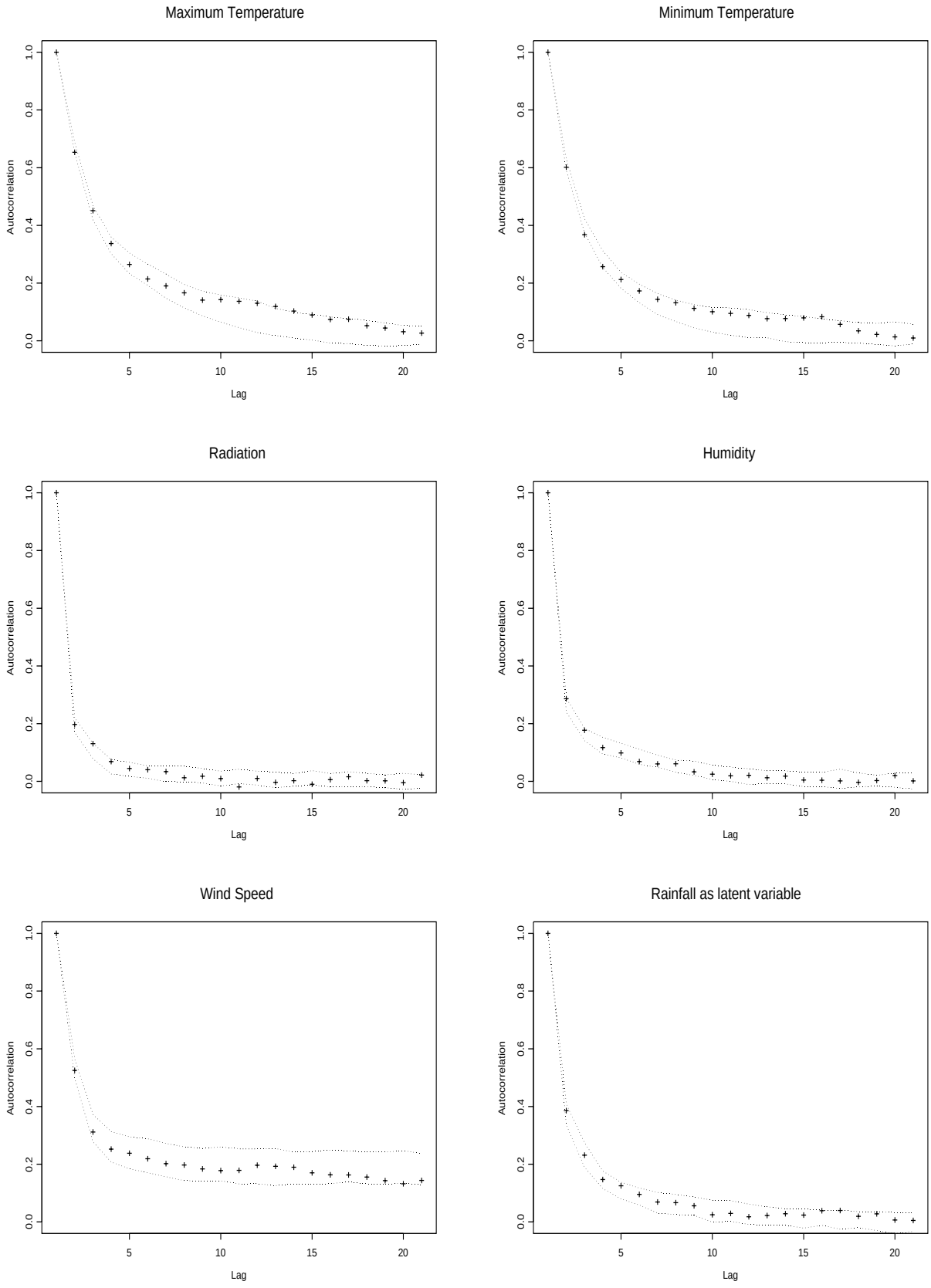

Figure 3: *Autocorrelations for the six weather variables (—) and simulation-based 95% confidence interval based on the fitted VARMA(2,1) model.*

$$\hat{A}_1 = \begin{pmatrix} 1.212 & -0.425 & 0.084 & 0.097 & -0.161 & -0.169 \\ 0.516 & 0.324 & -0.215 & 0.146 & -0.156 & 0.150 \\ 0.314 & -0.215 & 0.286 & -0.033 & 0.098 & -0.290 \\ -0.562 & 0.266 & 0.782 & 0.959 & -0.230 & 0.719 \\ 0.131 & -0.203 & -0.179 & 0.064 & 1.290 & -0.072 \\ 0.846 & -0.565 & -0.869 & 0.055 & -0.352 & 0.996 \end{pmatrix}$$

$$\hat{A}_2 = \begin{pmatrix} -0.200 & 0.209 & -0.012 & -0.014 & 0.082 & 0.156 \\ -0.052 & 0.037 & 0.027 & -0.007 & 0.067 & -0.038 \\ -0.102 & 0.088 & 0.049 & 0.008 & -0.017 & 0.181 \\ 0.257 & -0.091 & -0.188 & -0.082 & 0.132 & -0.435 \\ -0.036 & 0.094 & 0.004 & -0.009 & -0.289 & -0.024 \\ -0.193 & 0.073 & 0.026 & -0.019 & 0.392 & -0.162 \end{pmatrix}$$

$$\hat{M}_1 = \begin{pmatrix} -0.030 & -5.072 & -1.483 & 1.930 & -0.188 & -2.560 \\ 0.708 & 0.367 & -0.907 & 2.211 & -0.515 & -2.130 \\ 1.583 & 4.410 & 1.030 & -0.543 & 0.006 & 0.751 \\ -2.873 & -6.825 & -0.406 & 1.779 & -0.010 & -0.801 \\ 0.448 & 0.108 & -0.383 & 0.276 & 0.909 & -0.503 \\ 3.403 & 7.300 & 0.421 & -0.589 & -0.346 & 1.989 \end{pmatrix}$$

$$\hat{\Sigma} = \begin{pmatrix} 0.293 & & & & & \\ -0.114 & 0.056 & & & & \\ 0.218 & -0.153 & 0.849 & & & \\ -0.035 & 0.009 & -0.154 & 0.358 & & \\ 0.017 & 0.005 & -0.055 & 0.059 & 0.615 & \\ -0.042 & 0.033 & -0.441 & 0.346 & 0.052 & 0.521 \end{pmatrix}$$

Table 3: *Parameter estimates in VARMA(2,1) model.*

3 Model validation

The multivariate latent Gaussian model was used to simulate 100 runs of 18 years of weather data using the parameter values for the VARMA process given above. The first 365 days in each run were discarded to ensure independence from the starting values. Weather variables were then obtained by reversing the transformations applied in §2. For comparison, series were also simulated from the models of Richardson (1981) and Peiris and McNicol (1996), again with parameters estimated from the Mylnefield data.

Comparisons were based on monthly means of weather variables (see Figure 4), number of wet days and total amount of rain per month (see Figure 5). The maximum and minimum temperature of the generated data from the different models did not differ significantly from the observed data. However, for radiation, the model of Richardson (1981) overestimated values

during summer, and relative humidity was also overestimated. The model of Peiris and McNicol (1996) performs poorly in the aspects related to rainfall, confirming what they pointed out in their paper. The amount of rainfall was underestimated significantly for some months. The data generated from the latent model are generally consistent with the observed data.

Figure 3 shows a histogram of durations of wet periods per year for historical and simulated data. The model of Peiris and McNicol (1996) does not cope with wet periods of 5 or more days and it also underestimate the number of wet periods of duration between 2 and 4 days. The latent and Richardson (1981) models perform well overall, although they both fail to reproduce wet periods over 10 days.

4 Discussion

The model we have proposed provides a unified approach to the simulation of weather data: all variables are generated simultaneously by means of a multivariate latent Gaussian process which assumes that rainfall is a latent variable with threshold at zero. This general approach avoids a two-stage model where some variables are simulated conditional on others. The simplicity of the model facilitates the introduction of new weather variables and the Gaussian nature of the model makes easier the extension to a spatio-temporal framework where data can be interpolated between different locations.

Our model improves on the one proposed by Richardson (1981) in that the adequacy of the simulation of the weather variables does not depend so strongly on the proper description of the sequence of wet and dry days. The simulation study showed how the monthly average radiation did not compare well with the observed data for some months. This might has been a result of a poor fit of the series of wet and dry days. Another possible reason is the fact that Scottish rainfall is not Markovian.

The main difference between our model and that of Peiris and McNicol (1996) is the link between rainfall and the other variables. In the model of Peiris and McNicol (1996), rainfall is a function of the non-rainfall variable. This is a disadvantage because these variables may not capture well the correlation structure of rainfall. The latent model estimates the correlation structure of all variables simultaneously and generates more accurate results.

The latent Gaussian model needs to be fitted to data from other locations to check for variability of the parameters. It would also be of interest to study the influence of the number of years used on the parameter estimation. The programs used to fit the latent model and generate the simulated data were written in Fortran90 and they are available on request.

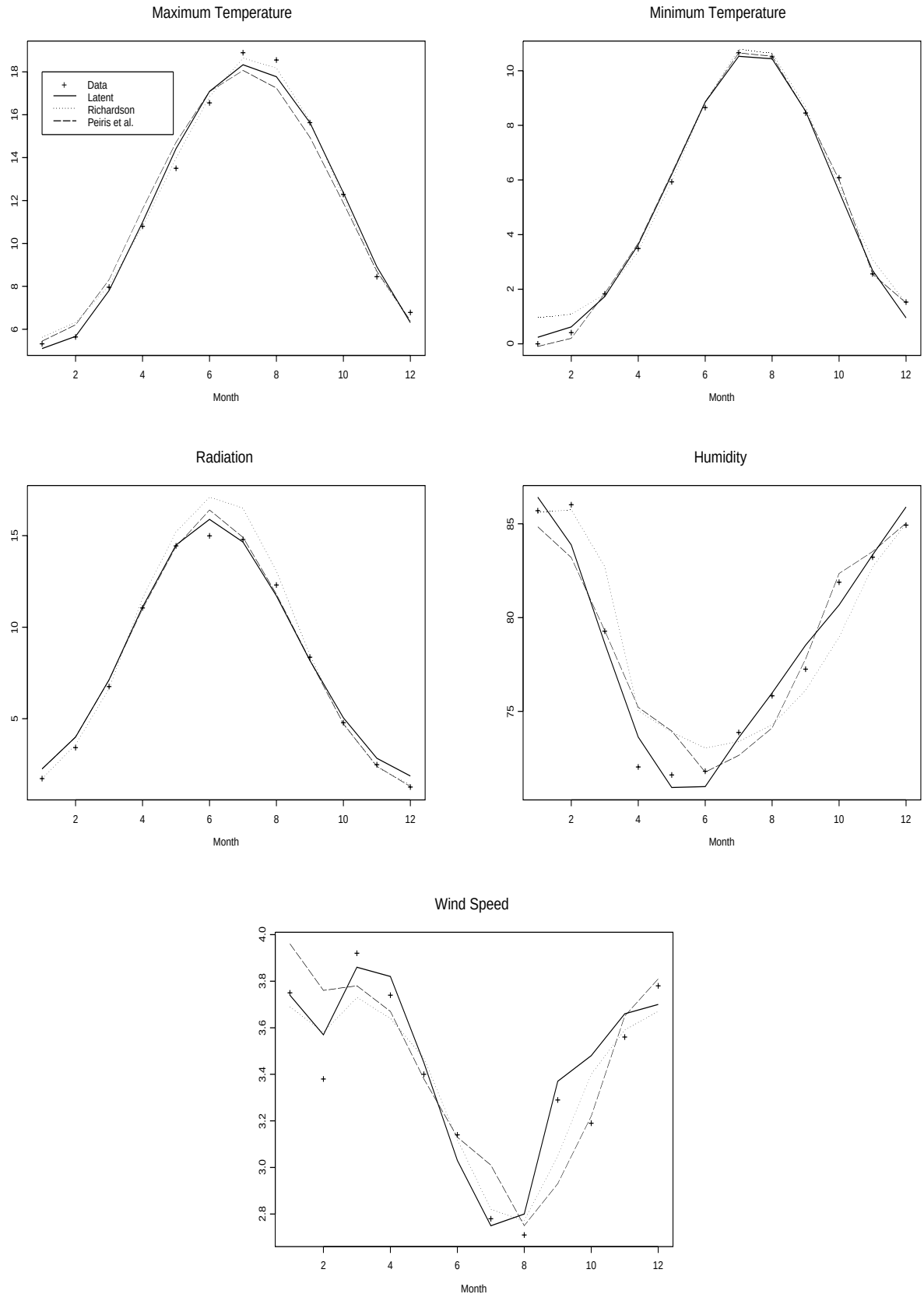

Figure 4: *Comparison of monthly means for non-rain variables: Data (+), Latent model (—), Richardson and Wright (· · ·), Peiris and McNicol (- - -).*

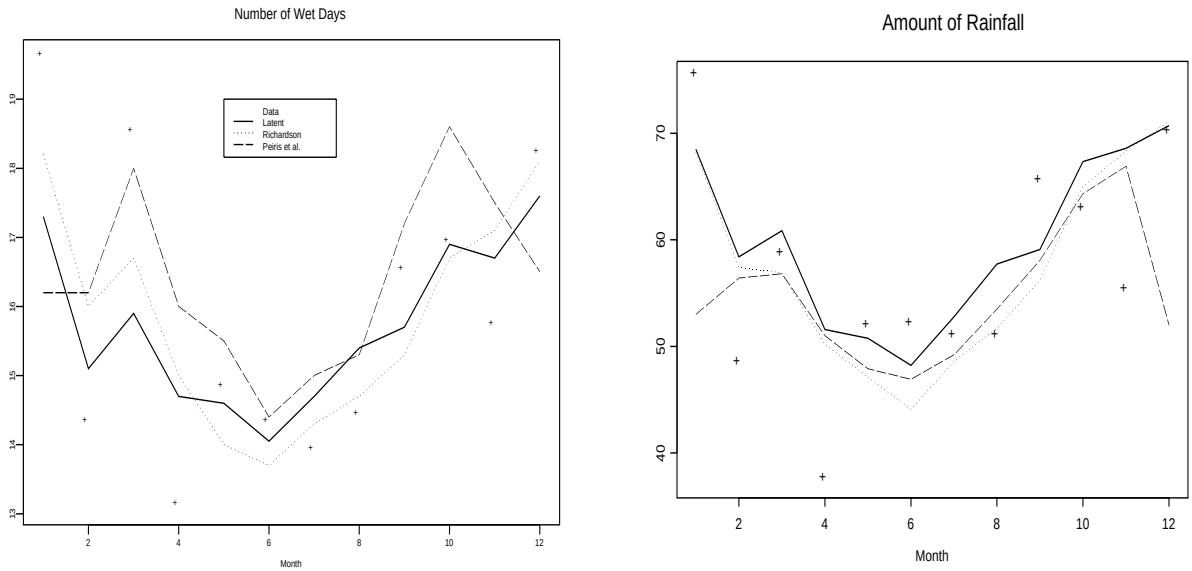

Figure 5: *Comparison of number of wet days and total amount of rain per month: Data (+), Latent model (—), Richardson and Wright (· · ·), Peiris and McNicol (- - -).*

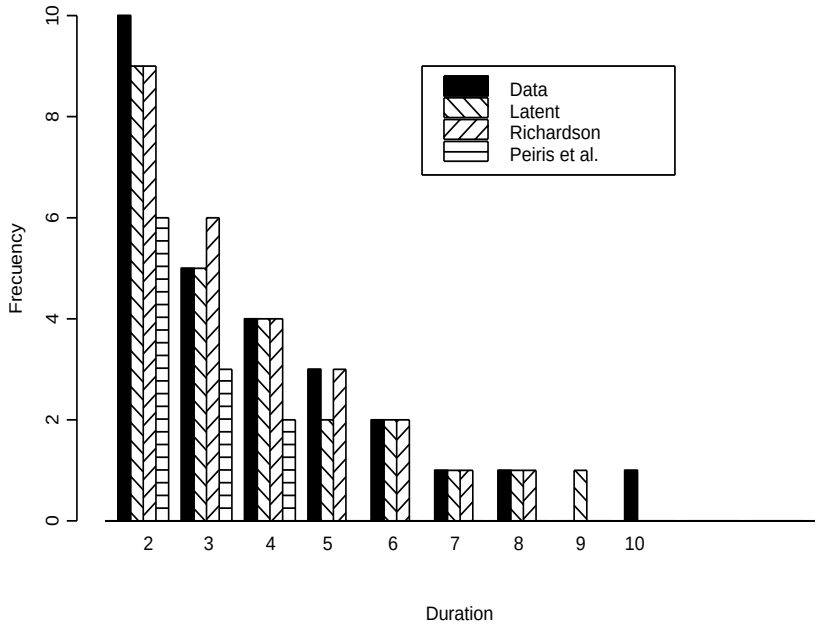

Figure 6: *Frequency of wet periods of duration two or more days (days/year).*

Acknowledgements

We are grateful to Jim McNicol and Ian Jolliffe for advice and encouragement on this work, which was supported by funds from the Scottish Executive Rural Affairs Department.

Appendix A: Detrending a latent Gaussian variable

We estimate the trend for the latent Gaussian variable, y_K, by numerically maximising the log-likelihood:

$$L = \sum_t \log p(t) \quad \text{where } p(t) = \begin{cases} \Phi\{C(t)\} & \text{if } y_K(t) = * \\ \frac{1}{\sigma_K}\phi\left(\frac{y_K(t)-\mu_K(t)}{\sigma_K}\right) & \text{otherwise.} \end{cases}$$

Here, ϕ is the Gaussian probability density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-x^2}{2}\right],$$

Φ is the Gaussian integral

$$\Phi(u) = \int_{-\infty}^{u} \phi(x)\ dx,$$

which is needed when y_K is censored, and the censoring limit for $z_K(t)$ is

$$C(t) = \frac{\alpha_0 - \mu_K(t)}{\sigma_K},$$

where α_0 is the censoring limit for y_K, given in (1).

Appendix B: Estimation of sample auto- and cross-correlations

We estimate the auto- or cross-correlation between the detrended variables z_i and z_k, at time lag l, denoted $\rho_{ik}(l)$, by numerically maximising a log-likelihood:

$$L = \sum_t \log p_{ik}(t, t-l).$$

If $i, k \neq K$, then

$$p_{ik}(t, t-l) = \phi_2\{z_i(t), z_k(t-l), \rho_{ik}(l)\},$$

where ϕ_2 is the Gaussian bivariate probability density function

$$\phi_2(w, x, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[\frac{-1}{2(1-\rho^2)}(w^2 + x^2 - 2\rho wx)\right].$$

If $i = K$ and $k \neq K$, then

$$p_{Kk}(t, t-l) = \begin{cases} \phi(z_k(t-l))\Phi\left\{[C(t) - \rho_{Kk}(l)z_k(t-l)]/\sqrt{1-\rho_{Kk}^2(l)}\right\} & \text{if } y_K(t) = * \\ \phi_2\{z_K(t), z_k(t-l), \rho_{Kk}(l)\} & \text{otherwise.} \end{cases}$$

We need not consider the case $i \neq K$ and $k = K$, because $\rho_{Kk}(l) = \rho_{kK}(-l)$. Finally, if $i = k = K$, then

$$p_{KK}(t, t-l) = \begin{cases} \Phi_2\{C(t), C(t-l)\} & \text{if } y_K(t) = y_K(t-l) = * \\ \phi\{z_K(t-l)\}\Phi\left\{[(t) - \rho_{KK}(l)z_K(t-l)]/\sqrt{1-\rho_{KK}^2(l)}\right\} & \text{if only } y_K(t) = * \\ \phi_2\{z_K(t), z_K(t-l), \rho_{KK}(l)\} & \text{otherwise.} \end{cases}$$

Here, Φ_2 is the bivariate Gaussian integral

$$\Phi_2(u, v, \rho) = \int_{-\infty}^{u} \int_{-\infty}^{v} \phi_2(w, x, \rho)\, dw\, dx.$$

References

Bell, T. L. (1987). A space-time stochastic model of rainfall for satellite remote-sensing studies. *Journal of Geophysical Research*, 92:9631–9643.

Geng, S., Devries, F. W. T. P., and Suppit, I. (1986). A simple method for generating daily rainfall data. *Agricultural and Forest Meteorology*, 36:363–376.

Glasbey, C. A. and Nevison, I. M. (1997). Rainfall modelling using a latent Gaussian variable. In Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek-Cohen, E., Warren, W. G., and Wolfinger, R. D., editors, *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, number 122 in Lecture Notes in Statistics, pages 233–242. Springer, New York.

Glasbey, C. A., Nevison, I. M., and Hunter, A. G. M. (1998). Parameter estimators for Gaussian models with censored time series and spatio-temporal data. In Payne, R. and Green, P., editors, *COMPSTAT98 Proceedings in Computational Statistics*, pages 323–328, Heidelberg. Physica-Verlag.

Hutchinson, M. F. (1995). Stochastic space-time weather models from ground-base data. *Agricultural and Forest Meteorology*, 73:237–264.

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

Le Cam, L. (1961). A stochastic description of precipitation. In J., N., editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 165–186.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Springer-Verlag.

Peiris, D. R. and McNicol, J. W. (1996). Modelling daily weather with multivariate time series. *Agricultural and Forest Meteorology*, 79:219–231.

Racsko, S. L. and Semenov, M. (1995). A serial approach to local stochastic weather models. *Ecological Modelling*, 57:27–41.

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17:182–190.

Richardson, C. W. and Wright, D. A. (1984). WGEN: a model for generating daily weather variables. Technical Report ARS-8, U.S. Dep. of Agric. Res. Service.

Rodriguez-Iturbe, I., Cox, D. R., and Isham, V. (1988). A point process model for rainfall: further developments. *Proceedings of the Royal Society, London, Series A*, 417:283–298.

Sanso, B. and Guenni, L. (1999). A stochastic model for tropical rainfall at a single location. *Journal of Hydrology*, 214:64–73.

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