

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

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 vector second-order autoregressive first-order moving average (VARMA(2,1)) process, fits the data better, 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. Many scientists solve this lack of historical data by using 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. In particular, Richardson (1981) simulates rainfall as a Markov chain, exponential model and then the other variables are generated depending on whether the day is wet or dry, whereas Peiris and McNicol (1996) first simulate all variables

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

†Corresponding author: FAX - +(44) 131 650 4901, email - chris@bioss.ac.uk

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 daily 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. 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 (1981) 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 $k = 1, \dots, K$ and $t = 1, \dots, T$, 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 the latent variable underlying rainfall. 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

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

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 values of rain over 43mm. However, we are not overly concerned, as this value was exceeded only on four occasions in 17 years. Figure 2 illustrates the relationship between the latent variable and rainfall.

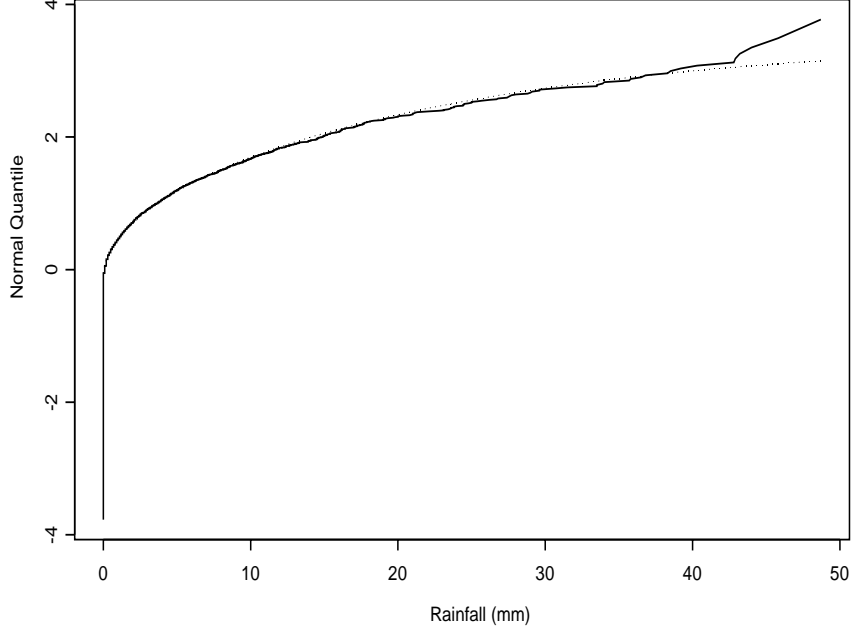


Figure 1: *Normal probability plot for 17 years of daily rainfall data for Mylnfield, Scotland (—), and the fitted curve, a quadratic function of power-transformed rainfall (⋯).*

All six weather variables exhibited trends due to annual cyclic patterns, which we accounted for by finite Fourier series:

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

We estimated parameters β , θ and σ^2 by maximum likelihood, assuming independence, and used likelihood-ratio tests to select the value of J . In the case of the rainfall latent variable, 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. For all variables, for these data, $J = 1$ or 2 . Table 1 shows the results.

We denote the detrended weather variables by z , where

$$z_k(t) = \frac{y_k(t) - \hat{\mu}_k(t)}{\hat{\sigma}_k}.$$

We used a vector autoregressive moving average (VARMA) to model the temporal dependence between these variables. The general form of a VARMA process of order (p, q) is:

$$z(t) = A_1 z(t-1) + \dots + A_p z(t-p) + e_t - M_1 e(t-1) - \dots - M_q e(t-q), \quad (3)$$

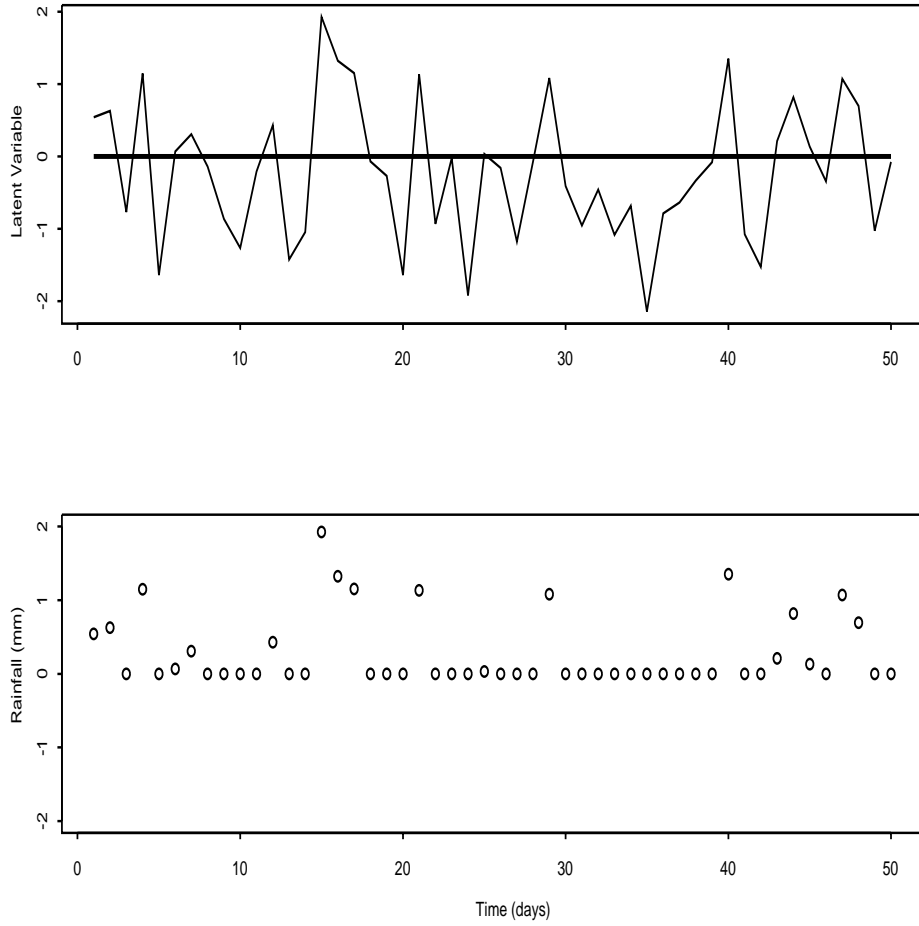


Figure 2: *Plot of the latent variable with threshold at zero (top) and the corresponding values for rainfall (bottom).*

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	$\log_e(\text{Radiation}+1)$	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	Transformed rainfall	0.00	0.13	0.31	-0.04	2.42	1.00

Table 1: *Parameter estimates for annuals trends*

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

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

where $z(t) = (z_1(t), \dots, z_K(t))'$ is a K -dimensional random vector, A_1, \dots, A_p and M_1, \dots, M_q are $K \times K$ matrices of parameters and e is a K -dimensional white noise process with $e \sim N(0, \Sigma)$. As the rainfall variable is latent, we cannot follow the standard route and estimate parameters by full maximum likelihood. Instead we adopt an ad hoc procedure, previously studied by Glasbey and Nevison (1997) and Glasbey et al. (1998), of estimating the parameters by minimising the sum of squares of differences between estimated and expected correlations. We numerically minimise

$$\sum_{i=1}^K \sum_{k=1}^K \sum_{l=0}^L (c_{ik}(l) - \rho_{ik}(l))^2, \quad (4)$$

with respect to the parameters in the VARMA process, for a specified number of lags, L . Here, $\rho_{ik}(l)$ denotes the expected auto- or cross-correlation between $z_i(t)$ and $z_k(t-l)$, i.e., between series i and k at time lag l . This is complicated functions of A , M and Σ (see, for example, Lütkepohl, 1991). Similarly, $c_{ik}(l)$ denotes an estimated auto- or cross-correlation. Those involving the rainfall latent variable cannot be estimated directly, but instead we use the method given in Appendix B.

The order of the process, (p, q) , was determined by comparing the estimated autocorrelations, c , with 95% probability intervals obtained by simulation, for a range of values of p and q . Figure 3 shows the results for $(p, q) = (2, 1)$, 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 VARMA process was estimated using $L = 7$, a value also chosen by simulation, based on Table 2, which shows the mean square errors of parameter estimates for a range of values of L . Estimated values of A_1 , A_2 , M_1 and Σ are given in Table 3.

3 Model validation

The multivariate latent Gaussian model obtained in §2 was used to simulate 100 runs of 18 years of weather data. The first 365 days in each run were discarded to ensure independence from the starting values. 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, Richardson's model overestimated values during summer, and relative humidity was also overestimated. Peiris and McNicol's model performs

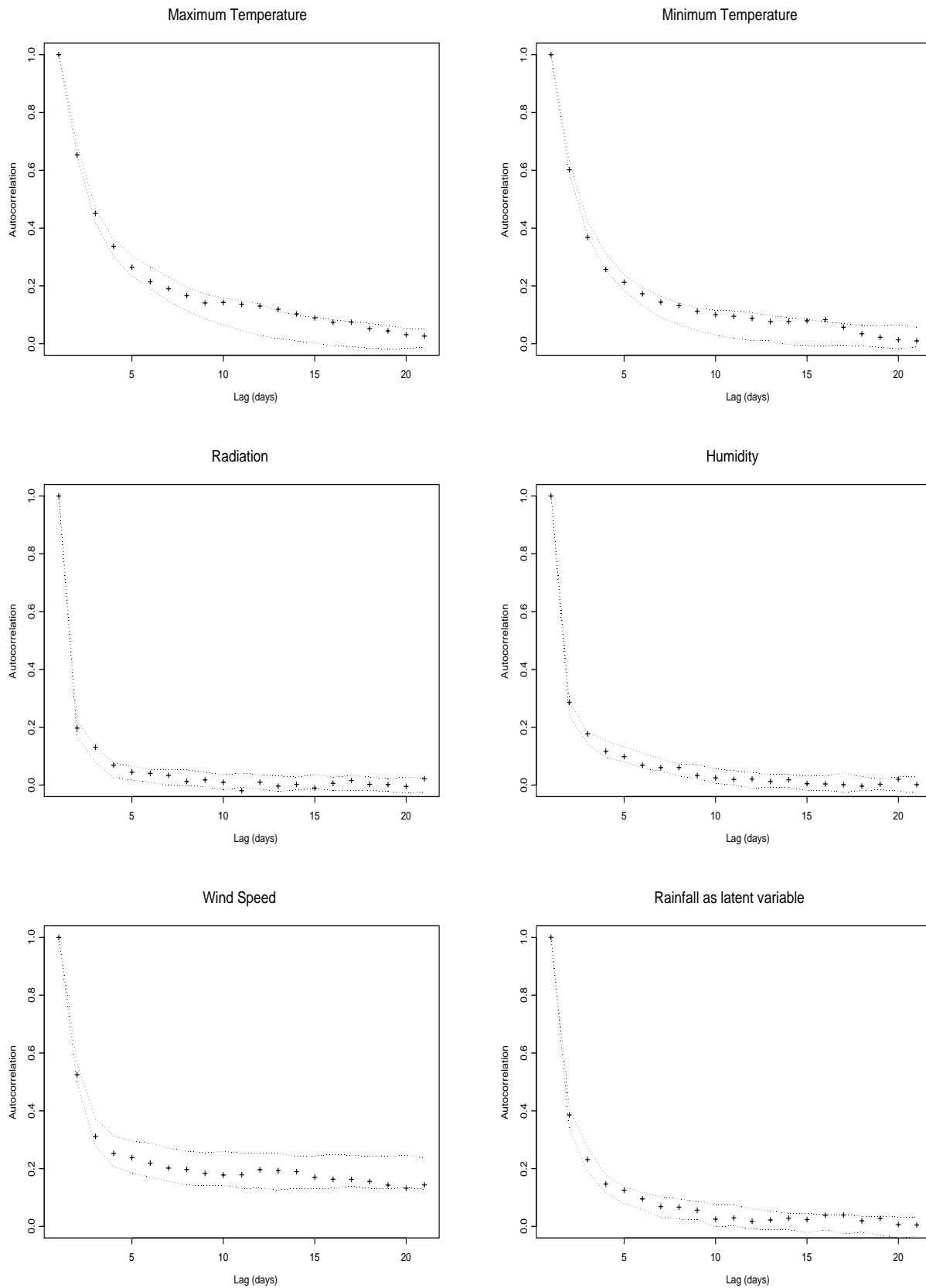


Figure 3: *Estimated autocorrelations, c , for the six weather variables (+) and simulation-based 95% probability intervals ($\cdot\cdot\cdot$) based on the fitted VARMA(2,1) model.*

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

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

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

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

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

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. In Figure 4, we also compare the distributions of values within a month simulated by the latent model with those in the data. We have used interquartile limits to summarise distributions, denoting the 25th and 75th percentiles in a set of values. We see reasonable consistency.

As a further evaluation and comparison of models, we compared runlengths of rainy days and of warm, humid days. In particular, we focussed on warm, humid days as a way of illustrating one of the many possibilities for considering auto- and cross-covariances of multiple variables. Smith (1956) identified the critical threshold for the onset of potato blight as being *a temperature in excess of 10°C and relative humidity above 90% for 11 or more hours in each of two or more consecutive days*. For daily data, we modified the temperature threshold to be that the

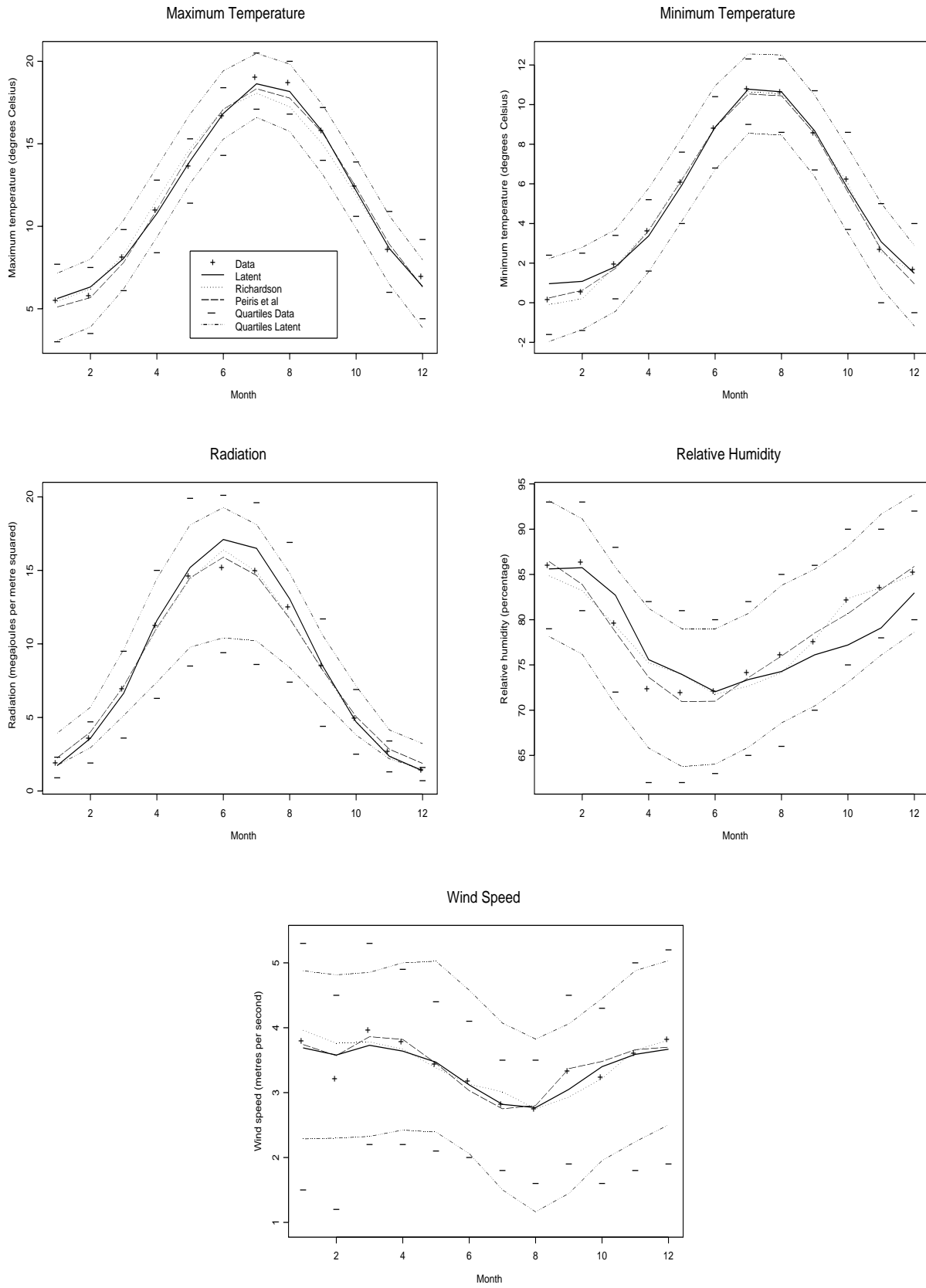


Figure 4: Comparison of monthly means for non-rain variables: Data (+), Latent model (—), Richardson's model (⋯⋯), Peiris and McNicol's model (- - -).

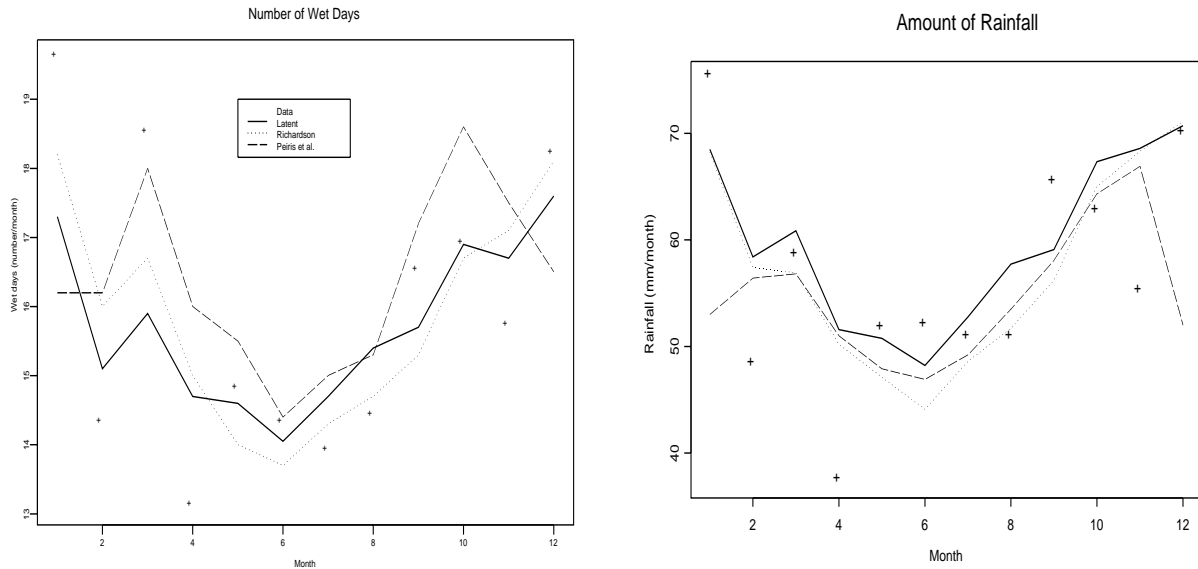


Figure 5: Comparison of number of wet days and total amount of rain per month: Data (+), Latent model (—), Richardson's model (⋯), Peiris and McNicol's model (- -).

average of maximum and minimum temperature exceeded 10°C . Figure 3 shows histograms of durations of wet and warm, humid periods year for historical and simulated data. Peiris and McNicol's model underestimates the frequency of wet periods, severely so for those of longer duration. Otherwise, the models perform reasonably well, though they all have a tendency to underestimate the persistence of weather conditions.

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 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 for Richardson's model did not compare well with the observed data for some months. This might have been a result of a poor fit of the series of wet and dry days. Another possible reason is the possibility that Scottish rainfall is not Markovian.

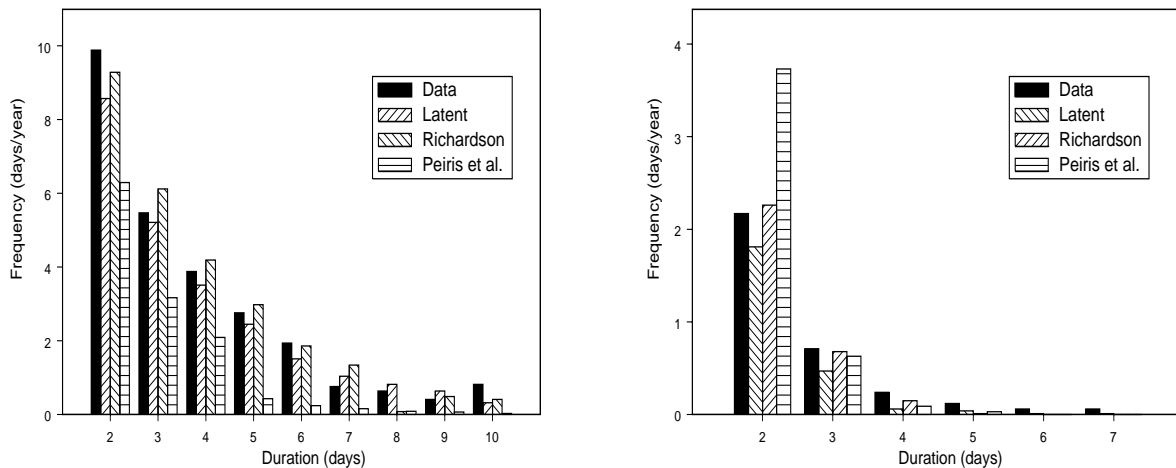


Figure 6: *Frequency of wet periods and of warm, humid periods (average temperature $> 10^{\circ}C$, relative humidity $> 90\%$) of duration two or more days.*

The main difference between our model and that of Peiris and McNicol is the link between rainfall and the other variables. Their model rainfall is a function of the non-rainfall variables. This is a disadvantage, because these variables may not capture well the auto-correlation structure of rainfall, whereas the latent model estimates the auto- and cross-correlation structure of all variables simultaneously.

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 Fortran90 program used to generate the simulated data is available on request.

5 Conclusions

We have used a vector autoregressive moving average process as a model for daily weather data. In the case of rainfall, this involved the use of a monotonic transformation to achieve marginal normality, thus defining a latent variable, with zero rainfall data corresponding to censored values below a threshold. We presented methodology for model identification, estimation and validation. We found that a vector second-order autoregressive first-order moving average process fitted data from Mylnfield, Scotland, better than existing models due to Richardson (1981) and Peiris and McNicol (1996), and produced more realistic simulated series of daily weather data.

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:

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

Here, '*' denotes a censored value when rainfall is zero, 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

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

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

Appendix B: Estimation of 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 $c_{ik}(l)$, by numerically maximising a log-likelihood:

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

If $i, k \neq K$, then

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

where ϕ_2 is the Gaussian bivariate probability density function

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

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

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

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

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

Here, Φ_2 is the bivariate Gaussian integral

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

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 Neyman, J., 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, Berlin.

- 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.
- Smith, L. P. (1956). Potato blight forecasting by 90% humidity criteria. *Plant Pathology*, 5:83–87.
- Stern, R. D. and Coe, R. (1984). A model fitting analysis of daily rainfall data. *J. R. Statist. Soc. A*, 147:1–34.