

Rainfall Modelling Using a Latent Gaussian Variable

C.A. Glasbey and I.M. Nevison

*Biomathematics and Statistics Scotland
United Kingdom*

ABSTRACT A monotonic transformation is applied to hourly rainfall data to achieve marginal normality. This defines a latent Gaussian variable, with zero rainfall corresponding to censored values below a threshold. Autocorrelations of the latent variable are estimated by maximum likelihood. The goodness of fit of the model to Edinburgh rainfall data is comparable with that of existing point process models. Gibbs sampling is used to disaggregate daily rainfall data, to generate typical hourly data conditional on daily totals.

Key words and phrases: Gibbs sampling, normalising transformation, rainfall disaggregation, time series.

1 Introduction

Temporal and spatio-temporal models of rainfall, either univariate or in combination with other climatic variables, are needed for many reasons, including simulation, forecasting and disaggregation. Monthly rainfall data are often modelled as Gaussian processes, but daily and hourly rainfall data typically have distributions with a singularity at zero and a long upper tail. Many models have been proposed, based either on point processes [9, 11] or constructed in two stages – first a binary rain/no-rain process and then a rainfall distribution applied to the wet periods [8, 13]. These models are far more difficult than Gaussian ones to study analytically, to combine with models of other weather variables [10], or to make use of in forecasting or disaggregation [6]. In particular, the need for disaggregation arises if rainfall estimates are required at a finer temporal or spatial resolution than are available in recorded data.

In this paper we develop an alternative approach, previously considered by Bell [2] and Hutchinson [7]: we apply a monotonic transformation to hourly rainfall data to achieve marginal normality, an approach akin to trans-Gaussian kriging [4]. This defines a latent Gaussian variable, with

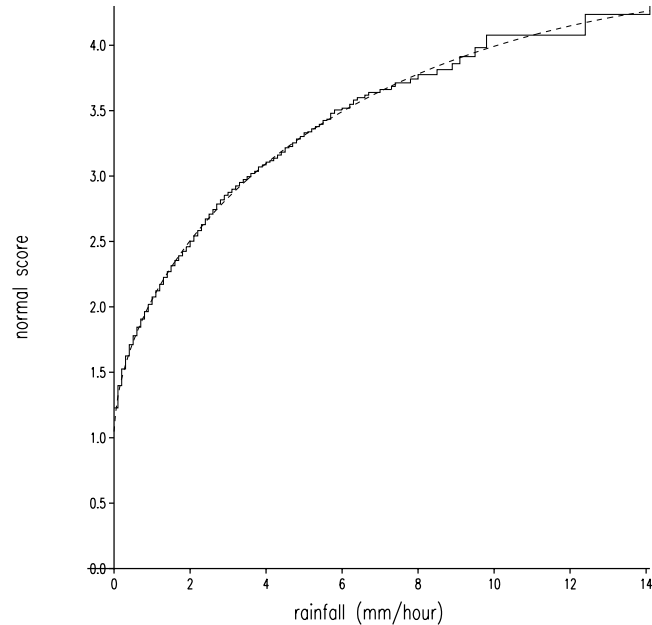


FIGURE 1. *Normal probability plot for 10 years of hourly rainfall data (—), and the fitted curve, a quadratic function of power-transformed rainfall (- - -).*

zero rainfall corresponding to censored values below a threshold, which we summarise by its autocorrelation function. Latent variable models are well established in categorical data analysis [1]. In §2, we identify and estimate the parameters in this model, using data from Edinburgh for illustration. Then, in §3 we check the validity of the model by comparing summary statistics from the data with those from simulations of the model and of a point process model [6, 11]. In §4, we show how Gibbs sampling can be used to disaggregate rainfall data. Finally, in §5 we critically evaluate the model.

2 Estimation

Fig 1 shows, as a continuous line, the normal probability plot for 10 years of hourly rainfall data from Turnhouse airport, Edinburgh. Note, all normal scores exceed 1.0 because 89% of hours were dry. An analytically-invertible monotonic fit to the plot was sought, to transform the rainfall data to marginal normality (with zero mean, unit variance). A power-transformation alone was not quite adequate, but a quadratic in power-

transformed rainfall,

$$y = \beta_0 + \beta_1 r^\gamma + \beta_2 r^{2\gamma} \quad \text{for } r > 0,$$

gave a good fit, where r denotes hourly rainfall and y denotes the latent variable. Parameters were estimated by non-linear least squares, giving $\beta = (1.05, 1.10, -0.09)$ and $\gamma = 0.6$. (With these parameter values, the function is monotonic only for $r < 20.3\text{mm}$, which places an upper limit of 4.4 on y , but this has a very small probability of 6×10^{-6} of being exceeded, i.e. once in 19 years.) The estimated transformation is displayed as the dashed line in Fig 1.

The autocorrelation function is sufficient to characterise fully a stationary Gaussian process. Autocorrelation coefficients of the latent variable cannot be estimated directly because, although y is known when $r > 0$, in rain-free hours y is censored, we know only that $y < 1.05$. However, it is relatively straightforward to estimate the autocorrelation coefficient at a particular time lag, by numerically maximising the likelihood of the observed bivariate histogram of the censored latent variable. Autocorrelations for lags up to twenty days were each estimated separately, and are shown as the plotted points in Fig 2. The rate of decay is not exponential: both short-term effects of a few hours duration, and persistent correlations lasting several days, typical of cyclonic weather systems, are apparent. No diurnal pattern was found (although this could easily be included in this model), so the fluctuations in autocorrelation were attributed to sampling variation. To smooth out these fluctuations and estimate the autocorrelation function, mixtures of exponential curves,

$$\text{cor}(y_t, y_{t+l}) = \sum_{i=1}^m \alpha_i e^{-\lambda_i |l|} \quad \text{subject to} \quad \sum_{i=1}^m \alpha_i = 1,$$

were fitted by non-linear least squares. The constraint that all coefficients are positive is sufficient for the function to be valid. A mixture of four exponentials, with $\alpha = (0.21, 0.51, 0.20, 0.08)$, $\lambda = (0.48, 0.17, 0.049, 0.0077)$ hour⁻¹, was judged to be adequate and is plotted as the line in Fig 2. One way to interpret this function is that the latent variable is a weighted sum of four Markov processes, and therefore the rainfall model is a hidden-Markov process. More sophisticated methods, such as Markov chain Monte Carlo [12], could have been used to estimate the autocorrelation function, perhaps simultaneously with parameters in the normalising transformation. However, because the number of observations is in excess of 87000, we do not think that efficiency of estimation is of critical consideration in this application.

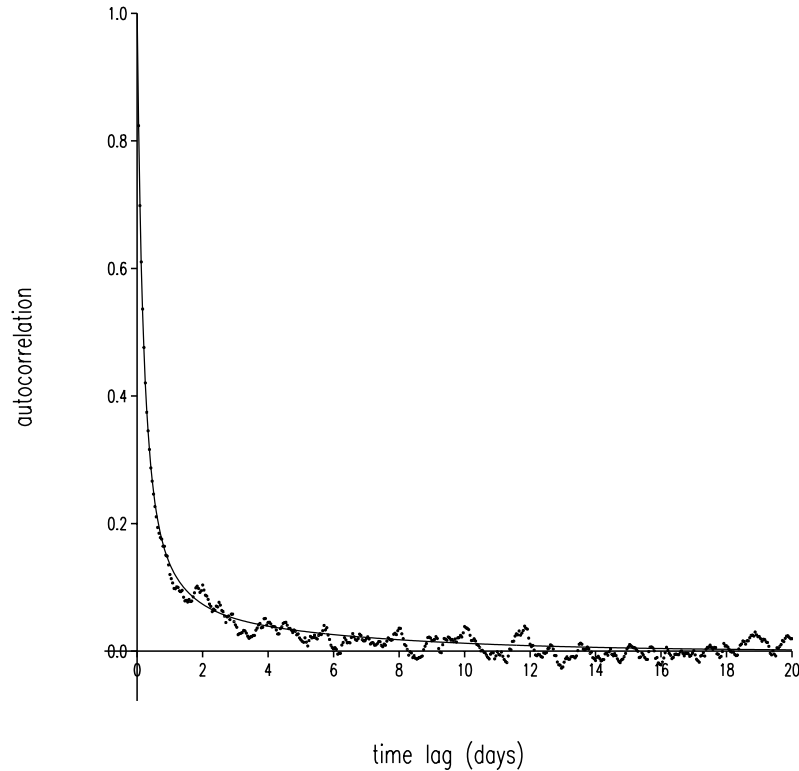


FIGURE 2. Autocorrelations of latent variable at a range of time lags, estimated from censored data (\cdot) and a fitted mixture of four exponentials ($—$).

3 Validation

To check the adequacy of the model, 100 years of data were simulated, using an autoregressive process derived from the estimated autocorrelation function, with lag terms up to twenty days and y truncated at 4.4. Partial autocorrelations beyond this lag were assumed to be negligible. (Alternatively, an exact representations of the autocorrelation function could have been achieved using a fourth-order autoregressive third-order moving-average process.) A large number of summary statistics were compared with those from the data. The most critical of these statistics are given in Table 1, together with results from a particular clustered-point-process model [11] fitted by the method of moments [6]. (Some analytic results are available for both models, but for simplicity we have relied on simulations.) There is close agreement in the first three statistics because of the estimation

	Edinburgh data	Latent Gaussian model	Point process model [11]
Hourly data			
proportion wet	0.11	0.11	0.11
mean rainfall (mm)	0.068	0.070	0.067
standard deviation	0.33	0.34	0.32
autocorrelation lag 1	0.55	0.66	0.53
lag 2	0.38	0.46	0.31
lag 3	0.28	0.34	0.23
mean duration of dry run	22.	21.	28.
standard deviation	43.	33.	35.
mean duration of wet run	2.7	2.6	3.4
standard deviation	3.1	2.7	2.9
Aggregated daily data			
proportion wet	0.53	0.58	0.52
mean rainfall (mm)	1.6	1.7	1.6
standard deviation	3.6	4.0	3.4
autocorrelation lag 1	0.19	0.21	0.20
lag 2	0.08	0.08	0.06
lag 3	0.02	0.04	0.03
mean duration of dry run	2.8	2.2	2.1
standard deviation	2.9	1.7	1.5
mean duration of wet run	3.1	3.0	2.3
standard deviation	2.7	2.7	1.8

TABLE 1. *Summary statistics for 10 years of Edinburgh’s rainfall data and 100 years of simulated data for two models.*

procedures. Agreement for other statistics is not quite so good, but is still reasonable. The point process model produces a closer match to hourly and daily autocorrelations, whereas the latent Gaussian model agrees better with observed run-lengths of dry and wet periods.

To explore further the discrepancies in rainfall autocorrelations, the bivariate distributions of rainfall in two hours, a fixed time lag apart, were examined. Fig 3(a) shows the bivariate histogram of observed rainfall at lag one hour, with counts displayed as shades of grey. The symmetry in Fig 3(a) about the main diagonal supports the assumption of time reversibility which is implicit in the latent Gaussian model. Fig 3(b) shows the expected counts, at lag one hour, from the latent Gaussian model. To compare the distributions, standardised residuals ($\{\text{observed count} - \text{predicted count}\} / \sqrt{\text{predicted count}}$) are shown in Fig 3(c). Groups of positive and negative residuals can be seen in the bottom-left of Fig 3(c), showing lack of fit of the model. In particular, there is a sequence of negative then positive residuals in both the first column and last row of the

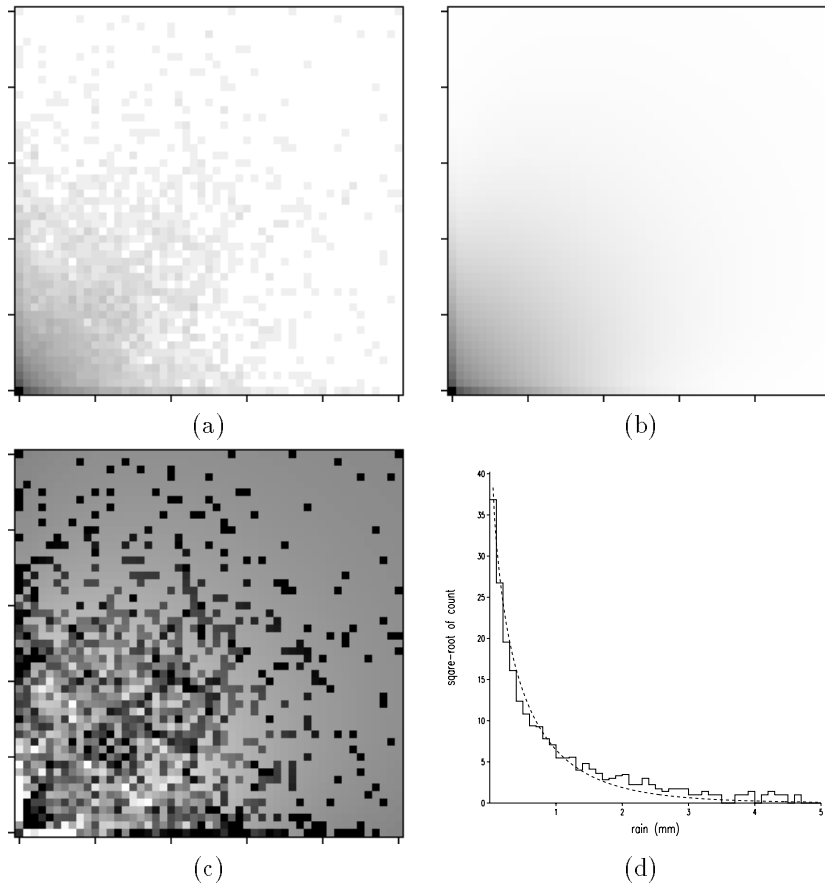


FIGURE 3. *Bivariate distribution of rainfall in consecutive hours (for rainfall in the range 0 - 5 mm): (a) histogram of observed data (zero counts are displayed as white squares, and progressively darker squares denote larger counts), (b) predicted counts from Latent Gaussian model, (c) standardised residuals (zeroes are displayed as mid-greys, negative residuals as lighter shades of grey and positive residuals as darker shades of grey), (d) histogram of rainfall in one hour when the previous hour was rain-free (observed: —, fitted: - - -).*

display. These are pairs of hours, one of which is rain-free and the other wet. Fig 3(d) shows the histogram for the last row of Figs 3(a), together with predicted values from the latent Gaussian model. Some lack of fit is evident, particularly for rainfall 2–5mm, but it represents very few hours. A similar pattern was observed in bivariate distributions at other short time lags.

We conclude that the latent Gaussian model fits the data reasonably well in comparison with the clustered-point-process model. The worst fit occurs

in the rainfall autocorrelations, but examination of bivariate distributions at a range of time lags has shown this to be small. If, in a particular application, a discrepancy in the autocorrelations was thought to be critical, this could be overcome by choosing values of the autocorrelation coefficients of the latent variable to give an exact match [7], which is easy to do because there is a 1-1 correspondence between the autocorrelations in the two variables. Discrepancies in the bivariate histograms would still exist, of course. In §5, we discuss possible ways to overcome these.

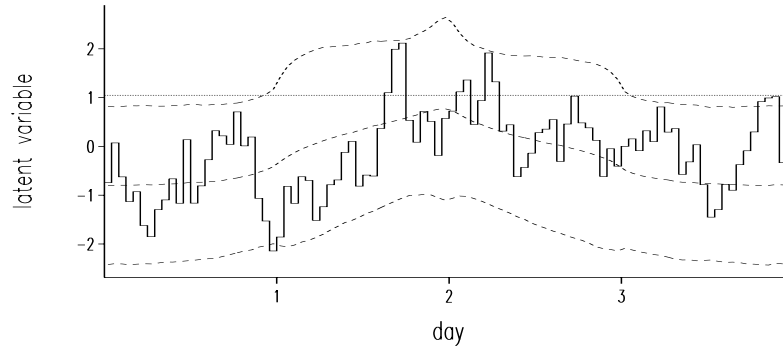
4 Disaggregation

A particular strength of the latent Gaussian approach to rainfall modelling is that it is relatively easy to either disaggregate or forecast, using the Gibbs sampler. We will illustrate this by considering disaggregation of daily Edinburgh data, i.e. simulation of stochastically-representative hourly data, conditional on daily totals. In practice, we would not usually know the appropriate parameter values at a site if rainfall were only recorded daily, but we could use estimates from the nearest hourly site, or spatially interpolate over several sites.

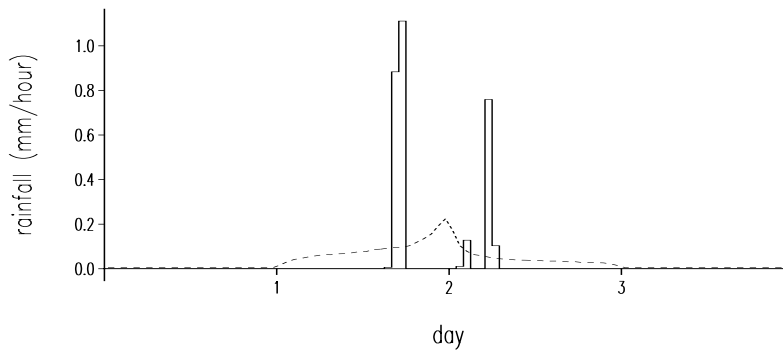
We initially allocate rainfall uniformly over each 24 hour period, to match the daily data, transform to latent variables, and set $y = 0$ on rain-free days. Then, the Gibbs sampler simulates in turn each day's set of latent variables, conditional on the latent variables on all other days. Initialising this simulation involves solving 24 sets of 960 simultaneous linear equations. By repeating this procedure over the full time series many times, a Markov chain of rainfall sequences is simulated from the appropriate distribution.

The hourly values of the latent variable on a single day, conditional on latent variables on all other days, are multivariate normally distributed. It is a standard result that their variances are derivable from the autocorrelation function (we used 20 days before and after the day under consideration, and assumed that partial autocorrelations were negligible outside this range), and their means also depend on the past and future latent variables. For disaggregation, we need to simulate from this distribution, subject to the constraint that, after transforming to the rainfall scale, the total amount of rain for that day matches the known value. There are many possible ways to achieve this, such as adaptive rejection sampling [5], but we took the simple approach of repeated sampling from the multivariate normal distribution until the rainfall total fell within either 0.05mm or 1% of the target. We then rescaled the latent variable in order to get an exact match for that day. This method proved to be efficient enough, requiring an average of 170 simulations per day, although considerably larger numbers were sometimes needed on very wet days.

The Markov chain appears to converge rapidly and to have good mixing properties: summary statistics converged to stable values within five iter-



(a)



(b)

FIGURE 4. *Illustration of disaggregation, when day 1 has a total of 2mm of rain, day 2 has 1mm of rain, and all preceding and succeeding days are rain-free. (a) A single realisation of the latent variable (—), and expected values and 95% confidence envelopes (- - -) estimated from 10000 simulations. (The threshold of 1.05, above which rainfall occurs, is also shown (···).) (b) The single realisation of rainfall resulting from (a) (—), and expected rainfall (- - -).*

ations, and the average correlation between consecutive simulations was only 8%. However, correlations were as high as 55% at the beginning and end of very wet days. Therefore, after a burn-in period of ten iterations, we took every tenth sequence as an independent disaggregation. Summary statistics (not presented) agree well with those in Table 1. Fig 4 illustrates the results obtained for two wet days during a long dry period. Note that the expected values and variances of the latent variable are not constant within a day. Perhaps surprisingly, we see that the expected rainfall reaches

a clear peak on the final hour of day one. However, a similar pattern was subsequently found in the data themselves.

The latent Gaussian model can be used in a similar way for forecasting from hourly data. In this case, the Gibbs sampler is used to simulate values of the latent variable for rain-free hours in the past, and then to simulate the latent variable into the future. Alternatively, an analytic approach may be possible.

5 Discussion

One inelegant feature of the latent Gaussian model is that, unlike point process models [11], it is not scale independent. For example, if hourly data arise from such a model, then daily data cannot also do so, although it may be possible to formulate the model in continuous time. Also, we share some of Sir David Cox's reservations, expressed in a review of time series in 1981 [3], that 'by pointwise transformation of a Gaussian series it is possible to produce series with any required marginal distribution. Such a construction will, however, often be artificial, especially for discrete distributions.' However, others maintain that point processes are themselves unnatural models for rainfall and many of the processes in the high atmosphere which generate rainfall are Gaussian. A further drawback of the latent Gaussian model is that it is difficult to see how to embed it in a broader class of models, so as to be able to generalise the basic model in situations where multivariate normality is not satisfied. It may be possible to model the latent variable as a probabilistic mixture of two or more Gaussian processes, or to simultaneously transform two or more consecutive rainfalls to multivariate normality, but both these approaches seem to encounter technical difficulties.

On the positive side, the model is simple to understand and to fit to data, it agrees well with Edinburgh's rainfall data, and is comparatively easy to use for simulation, disaggregation and forecasting. Also, the latent Gaussian model generalises elegantly to encompass both spatio-temporal data and multivariate situations involving other climatic variables, provided that distributional assumptions prove to be reasonable. We would expect both forecasting and disaggregation to be more appropriate in a multivariate, spatial setting. We are pursuing these ideas further, together with improving the estimation procedure and disaggregation algorithm.

ACKNOWLEDGEMENTS

The work was supported by funds from the Scottish Office Agriculture, Environment and Fisheries Department.

6 REFERENCES

- [1] Andersen, E.B. (1990). *The Statistical Analysis of Categorical Data*. Springer-Verlag, Berlin.
- [2] Bell, T.L. (1987). A space-time stochastic model of rainfall for satellite remote-sensing studies. *Journal of Geophysical Research*, **92**, 9631-9643.
- [3] Cox, D.R. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, **8**, 93-115.
- [4] Cressie, N.A.C. (1991). *Statistics for Spatial Data*. Wiley, New York.
- [5] Gilks, W.R., Best, N.G., Tan, K.K.C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, **44**, 455-472.
- [6] Glasbey, C.A., Cooper, G., McGechan, M.B. (1995). Disaggregation of daily rainfall by conditional simulation from a point-process model. *Journal of Hydrology*, **165**, 1-9.
- [7] Hutchinson, M.F. (1995). Stochastic space-time weather models from ground-based data. *Agricultural and Forest Meteorology*, **73**, 237-264.
- [8] Katz, R.W., Parlange, M.B. (1995). Generalizations of chain-dependent processes: applications to hourly precipitation. *Water Resources Research*, **31**, 1331-1341.
- [9] Le Cam, L. (1961). A stochastic description of precipitation. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (ed. J. Neyman) **3**, 165-186.
- [10] Richardson, C.W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, **17**, 182-190.
- [11] Rodriguez-Iturbe, I., Cox, D.R., Isham, V. (1988). A point process model for rainfall: further developments. *Proceedings of the Royal Society, London, Series A*, **417**, 283-298.
- [12] Sanso, B., Guenni, L. (1997). Venezuelan rainfall data analysed using a Bayesian space-time model. *Applied Statistics*, (in press).
- [13] Stern, R.D., Coe, R. (1984). A model fitting analysis of daily rainfall data (with discussion). *Journal of the Royal Statistical Society, Series A*, **147**, 1-34.