

# Spatial disaggregation of rainfall using a latent Gaussian Markov random field

David Allcroft<sup>1</sup> and Chris Glasbey<sup>1</sup>

<sup>1</sup> Biomathematics & Statistics Scotland, King's Buildings, Edinburgh, EH9 3JZ, Scotland

**Abstract:** A spatio-temporal model for rainfall is described, involving the transformation of rainfall to a thresholded Gaussian process which we represent as a Gaussian Markov random field. Gibbs sampling is then used to disaggregate rainfall from coarser spatial resolutions, to produce fine-scale realisations that are consistent with the observed rainfall totals.

**Keywords:** Disaggregation; Gaussian Markov random field; Latent Gaussian model; Rainfall.

## 1 Introduction

There is a growing interest in spatio-temporal models, simply because many datasets are both spatial and temporal. The application we consider here is the modelling of rainfall. Information on rainfall is frequently required as input into agricultural or hydrology models, however rainfall data are typically collected at coarser scales than needed for these purposes. The aim here is therefore to model rainfall data in order to perform spatial disaggregation, i.e. by modelling the spatio-temporal autocorrelation structure, we are subsequently able to simulate plausible realisations of the data at a fine scale that are consistent with the observed rainfall at the coarse scale. Spatio-temporal models typically combine geostatistical approaches in space with time series models and are Gaussian (see, for example, Mardia et al. 1998). In Gaussian Markov random fields (GMRFs), variables at non-adjacent locations are further assumed to be conditionally independent. Rainfall is non-Gaussian, but by suitable transformation can be modelled as a thresholded Gaussian variable. We formulate the process as a spatio-temporal GMRF and simulate disaggregations by Gibbs sampling.

For illustration, we consider 12 hours of hourly data from the Arkansas-Red Basin River Forecast Center website, <http://www.srh.noaa.gov/abrfc>. Chandler et al. (2000) modelled the data at  $8\text{km} \times 8\text{km}$  resolution, then aggregated to  $5 \times 5$  blocks and proceeded to disaggregate. We do likewise. Figure 1 shows one hour of data at both fine ( $75 \times 150$  values) and coarse ( $15 \times 30$ ) resolutions. The approach of Chandler et al. (2000) was limited to binary presence/absence data and was complex, involving a combination of transition probabilities, regression relationships and Markov random fields. Subsequently, Mackay et al. (2001) extended this by simulating values for

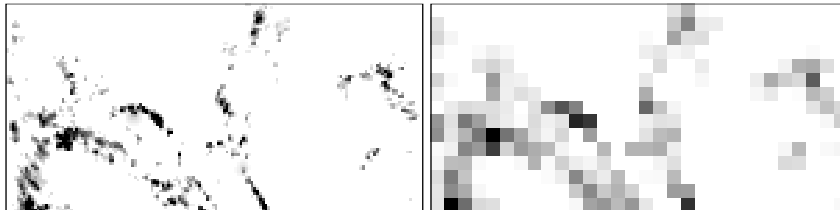


FIGURE 1. One of 12 hours of data, at fine and coarse resolution. (Rainfall intensities are displayed as shades of grey of increasing darkness.)

rainfall intensity and allocating them to wet locations. Our approach is able to directly model rainfall intensity and is more elegant.

## 2 Model

There are three stages in fitting the model: firstly the transformation of the rainfall to a thresholded Gaussian variable, secondly the estimation of the spatio-temporal autocorrelation of the resulting Gaussian field, and lastly the fitting of a GMRF to this autocorrelation structure.

### 2.1 Rainfall transformation

For locations at which rain is observed, the latent Gaussian variable takes a value above a threshold ( $\alpha_0$ ), and when no rain is observed, the variable takes a censored value below the threshold. For each location  $(i, j)$  in space at time  $t$ , we transform the rainfall,  $r$ , to the Gaussian variable,  $y$ , by

$$y_{ijt} = \begin{cases} \alpha_0 + \alpha_1 r_{ijt}^\gamma + \alpha_2 r_{ijt}^{2\gamma} & \text{if } r_{ijt} > 0, \\ * & \text{otherwise,} \end{cases} \quad (1)$$

as in Glasbey and Nevison (1997). Parameters are estimated by numerically minimising the sum of squares of differences between  $y$  and expected normal scores. Figure 2 shows the normal probability plot of the twelve hours of data, with the least squares fit of (1), with  $\alpha = (1.236, 0.04990, -0.0001610)$  and  $\gamma = 0.4411$ . The fit is seen to be good, except for rain levels above 140mm, but as only five of the 135000 values were in this range, this is not a cause for concern. We also note that the transformation (1) is a quadratic one, however with the maximum occurring at  $r = (-\alpha_1/2\alpha_2)^{1/\gamma}$ , here corresponding to 923mm rain, the function is monotonic for the range we consider.

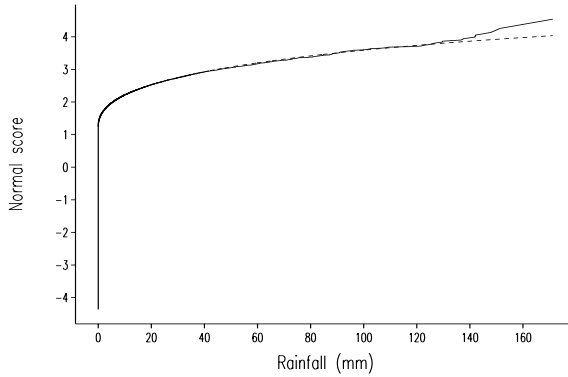


FIGURE 2. Normal probability plot for 12 hours of data (—), and fitted values (---).

## 2.2 Autocorrelation estimation

We estimate the autocorrelation of the GMRF at lag  $(k, l, s)$ , denoted  $\hat{\rho}_{kls}$ , by maximisation of  $\prod_{ij,t} p(y_{ijt}, y_{i-k, j-l, t-s})$ . For simplicity, we assume a torus wrap-round of the  $n_i \times n_j \times n_t$  spatio-temporal locations. Pairwise probabilities,  $p$ , take one of three forms depending on whether neither, one or both of the locations is dry and hence the latent Gaussian variable is censored:

$$p(y_{ijt}, y_{i-k, j-l, t-s}) = \begin{cases} \Phi_2(\alpha_0, \alpha_0, \rho) & \text{if } y_{ijt} = y_{i-k, j-l, t-s} = * \\ \phi(y_{ijt}) \Phi\left(\frac{\alpha_0 - \rho y_{ijt}}{\sqrt{1-\rho^2}}\right) & \text{if only } y_{i-k, j-l, t-s} = * \\ \phi_2(y_{ijt}, y_{i-k, j-l, t-s}, \rho) & \text{otherwise,} \end{cases}$$

where  $\phi$  and  $\phi_2$  are standard univariate and bivariate Gaussian probability densities, and  $\Phi$  and  $\Phi_2$  are the corresponding cumulative distribution functions (Durban and Glasbey, 2001).

The resulting pattern was judged to be spatially isotropic and so estimates at common distances were averaged for simplicity. Table 1 gives estimates at the first few lags in space for time lags  $s = 0$  and  $s = 1$ .

## 2.3 GMRF estimation

GMRF parameters are estimated by weighted least squares, as proposed by Rue and Tjelmeland (2002), by minimising

$$\sum_k \sum_l \sum_s \frac{1}{k^2 + l^2 + s^2} (\rho_{kls} - \hat{\rho}_{kls})^2, \quad (2)$$

4				.4907	4				.4419			
3			.5731	.5249	3		.5042	.4695				
2		.6771	.6155	.5574	2		.5678	.5326	.4923			
1	.8314	.7313	.6472	.5787	1	.6345	.5941	.5496	.5062			
0	1.	.8904	.7541	.6573	.5860	0	.6809	.6538	.6043	.5546	.5103	
	0	1	2	3	4		0	1	2	3	4	
	Time lag $s = 0$						Time lag $s = 1$					

TABLE 1. Estimates of autocorrelation for spatio-temporal lags up to (4, 4, 1).

where  $\rho$ , the expected autocorrelation, is a function of the GMRF parameters,  $\theta$ , specified by  $Q_{000, kls} = \theta_{kls}$ , where  $Q$  is the precision matrix and hence the likelihood of the process is proportional to  $\exp\{-\frac{1}{2}y'Qy\}$ . We compute  $\rho$  via two 3-D Fourier transforms: first obtaining eigenvalues

$$q_{ijt} = \sum_{k=0}^{n_i-1} \sum_{l=0}^{n_j-1} \sum_{s=0}^{n_t-1} Q_{000, kls} \exp \left\{ -2\pi i \left( \frac{ik}{n_i} + \frac{jl}{n_j} + \frac{ts}{n_t} \right) \right\},$$

then

$$\gamma_{kls} = \frac{1}{n_i n_j n_t} \sum_{i=0}^{n_i-1} \sum_{j=0}^{n_j-1} \sum_{t=0}^{n_t-1} \frac{1}{q_{ijt}} \exp \left\{ 2\pi i \left( \frac{ik}{n_i} + \frac{jl}{n_j} + \frac{ts}{n_t} \right) \right\},$$

and finally,  $\rho_{kls} = \gamma_{kls}/\gamma_{000}$ . The minimisation of (2) is constrained, as we need all  $q_{ijt} > 0$  to ensure that the variance matrix is positive definite and hence describes a valid process. We follow the approach of Rue and Tjelmeland (2002) and use an unconstrained simplex-based algorithm, adding a penalty if any eigenvalues are non-positive.

The parameters can be estimated on any sufficiently large torus, and are then valid for other grid sizes. To ensure spatial isotropy, we fit on a torus for which  $n_i = n_j (= 256)$ . As the correlation in time dies away more quickly, taking  $n_t = 64$  is sufficiently large. Several neighbourhood sizes were investigated; we found  $5 \times 5 \times 3$  to be the most suitable, requiring the estimation of 11 parameters. This produced a substantially better fit than a  $3 \times 3 \times 3$  neighbourhood, but a larger one ( $7 \times 7 \times 5$ ) proved unstable to fit (requiring 29 parameters) and offered no noticeable improvement. Figure 3 therefore illustrates the fit obtained at time lags 0 and 1 with neighbourhood size  $5 \times 5 \times 3$ , fitting to sample autocorrelation estimates up to lag (20, 20, 3). Due to the weighting coefficient used in (2), inclusion of higher lags in the sum of squares made negligible difference to the estimates obtained.

Table 2 shows the parameter estimates for the illustrated fit, converted to the quantities  $1/\sqrt{\theta_{000}}$  and  $-\theta_{kls}/\theta_{000}$ , corresponding to the conditional standard deviation and conditional autocorrelations, respectively.  $\theta_{000}$  is

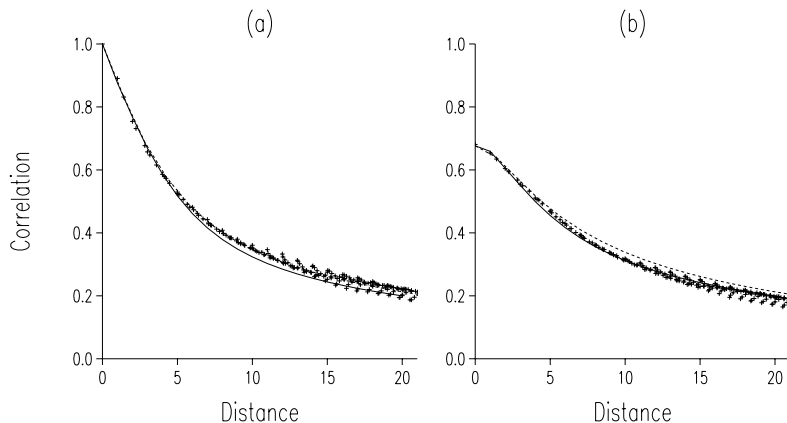


FIGURE 3. Sample and fitted autocorrelation of GMRF; (a) at time lag 0, (b) at time lag 1; (·) sample autocorrelation, (—) fitted autocorrelation along rows and columns; (---) fitted autocorrelation along diagonals.

2			-0.0481
1		0.0481	0.0254
0	0.3118	0.2343	-0.0364
	0	1	2

Time lag  $s = 0$

2			0.0312
1		0.0004	-0.0211
0	-0.0124	0.0401	-0.0258
	0	1	2

Time lag  $s = 1$

TABLE 2. Coefficients for the fitted GMRF. The value for  $(0, 0, 0)$  is the conditional standard deviation, the rest being conditional autocorrelations.

not estimated explicitly — it is fixed at 1 in the optimisation and then scaled to fit the variance. It should be noted that much more accuracy than that displayed is necessary to recover the quoted autocorrelation coefficients. Rue and Tjelmeland (2002) comment that these conditional autocorrelations are not easy to interpret, here being a mixture of positive and negative values, hence direct specification of a model by its conditional autocorrelation is not recommended.

### 3 Disaggregation

We use Gibbs sampling to update  $5 \times 5$  blocks, starting from a configuration in which rainfall is allocated uniformly within blocks (right-hand side of Figure 1). Let the vector  $y_A$  be the set of 25 locations to be updated and  $y_B$  be the vector of locations not conditionally independent of  $y_A$ , i.e. locations which have a non-zero entry in the precision matrix with at least

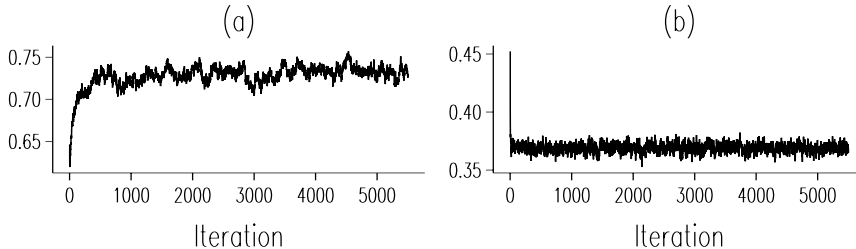


FIGURE 4. Trace plots of (a) lag 1 correlation in time, (b) proportion of wet locations (fine scale) within wet blocks (coarse scale).

one of the locations in  $y_A$ . Then we can write

$$\begin{pmatrix} y_A \\ y_B \end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q_{AA} & Q_{AB} \\ Q_{BA} & Q_{BB} \end{pmatrix}^{-1} \right).$$

Hence the full conditional distribution of  $y_A$  is multivariate normal and can be conveniently expressed in terms of the precision matrix  $Q$ :

$$y_A | y_B \sim \text{MVN} \left( -Q_{AA}^{-1} Q_{AB} y_B, Q_{AA}^{-1} \right).$$

For the given neighbourhood size,  $y_B$  is of length 218 and, as we are working on a torus, the neighbourhood size and shape is always the same and so the conditional variance needs to be calculated only once. Each conditional mean is calculated with a single matrix multiplication and so the whole simulation is computationally efficient. To weaken the wrap-round effect of the torus, we add onto the  $75 \times 150 \times 12$  lattice a spatio-temporal border of width  $75 \times 75 \times 12$ . This is an adequate size to allow the correlations to die away sufficiently.

At each update, the total rainfall in a block must be consistent with the observed total for that block. To ensure this, we repeatedly simulate values until the total of the back-transformed rainfall values are within 10% of the target value, and then make an adjustment to the wet values to match the totals exactly.

Statistics were collected from each realisation and plots and tests of convergence were applied using CODA (Best et al. 1995). It was decided that a burn-in of 500 iterations was required before convergence was achieved, and after this a further 5000 iterations were obtained. Figure 4 shows trace plots of two of the summary statistics considered, namely the autocorrelation at a time lag of one hour and the proportion of wet locations within wet blocks. Similar plots were examined for the mean and variance of the latent Gaussian variable, other autocorrelation estimates in space and time

	Observed	Simulated
Proportion wet locations	0.090	0.098
Proportion wet locations within wet blocks	0.334	0.369
Proportion correctly classified locations (wet/dry)	—	0.927
Proportion correctly classified locations in wet blocks	—	0.729
Lag 1 correlation in space	0.906	0.900
Lag 1 correlation in time	0.704	0.729

TABLE 3. Comparison of summary statistics of observed data with simulated realisations, averaged over 5000 realisations.

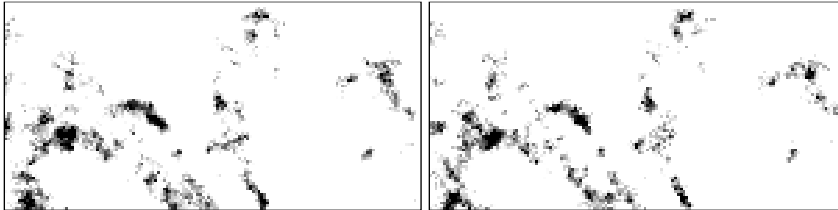


FIGURE 5. Two example disaggregated images, from two different runs.

and other proportions of wet and correctly classified (wet/dry) locations. Table 3 shows some summary statistics, illustrating the close agreement between the observed dataset and the simulated realisations.

Figure 5 shows two example disaggregations. Comparison with the corresponding observed data in Figure 1 (left) shows the disaggregations to be similar in most respects to the original data. An overall proportion of correctly classified locations within wet squares of 73% (Table 3) compares favourably with the 65% reported by Mackay et al. (2001).

## 4 Discussion

We have seen that rainfall can be successfully transformed to a thresholded Gaussian variable and hence access can be gained to existing methodologies for GMRFs. Using a model of this type, disaggregation is both straightforward and computationally efficient. For the GMRF, we empirically chose a neighbourhood of size  $5 \times 5 \times 3$ , the smallest size that gave an acceptable fit to the autocorrelation structure of the data. Further investigation would be useful to make formal comparisons of the efficiency of different neighbourhood sizes and ranges, extending the results in Rue and Tjelmeland (2002) to three dimensions. However it is a very attractive feature of these models that such a good fit to the autocorrelation structure can

be obtained with such a small neighbourhood and hence a relatively small number of parameters.

We have so far based all our analysis on the sequence of 12 hourly time points. A more rigorous approach, adopted by Chandler et al. (2000), would be to estimate model parameters from an independent, longer data sequence, e.g. for the same month in the previous year, and then use these parameters for the disaggregation. Nevertheless, we have demonstrated here that the modelling strategy works, the resulting realisations being convincingly close to the original data and visually much better than achieved previously.

**Acknowledgments:** Both authors acknowledge funding from the Scottish Executive Environment and Rural Affairs Department. We are also grateful to Richard Chandler and Havard Rue for helpful discussions.

### References

- Best, N. G., Cowles, M. K., and Vines, S. K. (1995). *CODA: Convergence Diagnostics and Output Analysis Software for Gibbs sampling output, Version 0.30*. MRC Biostatistics Unit, Cambridge, UK.
- Chandler, R. E., Mackay, N. G., Wheeler, H. S., and Onof, C. (2000). Bayesian image analysis and the disaggregation of rainfall. *Journal of Atmospheric and Oceanic Technology*, **17**, 641–650.
- Durban, M. and Glasbey, C. A. (2001). Weather modelling using a multivariate latent Gaussian model. *Agricultural and Forest Meteorology*, **109**, 187–201.
- Glasbey, C. A. and Nevison, I. M. (1997). Rainfall modelling using a latent Gaussian variable. In T. G. Gregoire et. al., editor, *Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, number 122 in Lecture Notes in Statistics, pages 233–242. Springer, New York.
- Mackay, N. G., Chandler, R. E., Onof, C., and Wheeler, H. S. (2001). Disaggregation of spatial rainfall fields for hydrological modelling. *Hydrology and Earth System Sciences*, **5**, 165–173.
- Mardia, K. V., Goodall, C., Redfern, E. J., and Alonso, F. J. (1998). The Krige Kalman filter (with discussion). *Test*, **7**, 217–285.
- Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, **29**, 31–49.