A BAYESIAN APPROACH TO MODELLING SPATIAL-TEMPORAL PRECIPITATION DATA [1]

Richard L. Smith [2] and Peter J. Robinson [3]

Most precipitation data comes in the form of daily rainfall totals collected across a network of rain gauges. Research over the past several years on the statistical modeling of rainfall data has led to the development of models in which rain events are formed according to some stochastic process, and deposit rain over an area before they die. Fitting such models to daily data is difficult, however, because of the absence of direct observation of the rain events. In this paper, we argue that such a fitting procedure is possible within a Bayesian framework. The methodology relies heavily on Markov chain simulation algorithms to produce a reconstruction of the unseen process of rain events. As applications, we discuss the potential of such methodology in demonstrating changes in precipitation patterns as a result of actual or hypothesized changes in the global climate.

1. Introduction

This paper represents the first part of research directed towards the question "What influence will future climate change have on the spatial and temporal patterns of precipitation?" General circulation models typically produce spatial averages of temperature and rainfall over large grid cells. However, the questions concerning rainfall that are of interest in hydrology, agriculture, etc., typically concern patterns of rainfall over a much smaller spatial scale. Thus there is a very general problem, sometimes referred to as "downscaling" or "disaggregation", concerned with turning predictions over large spatial scales into predictions over small spatial scales. As a first step to doing this, we need some understanding of current patterns of rainfall over the temporal and spatial scales of interest. This task motivates the present paper, which is concerned with contructing statistical models for the spatial- temporal distribution of rainfall.

It seems reasonable to try to construct models which reflect, even if only in a very crude and incomplete way, our physical understanding about precipitation, and in this connection there is a long history of stochastic models based on superpositions of point processes. An early attempt was that of Le Cam (1961), who proposed a three-level hierarchy of "storms", "fronts" and "rain cells". Subsequent modifications include the model of Waymire, Gupta and Rodriguez-Iturbe (1984). However, these papers did not directly attempt to fit the models to real data. Statistical methods were proposed by Rodriguez-Iturbe *et al.* (1987, 1988) and Cox and Isham (1988), using "method of moments" fits based on equating the theoretical and sample values of various quantities that can be calculated directly, such as moments and correlations. They pointed out the difficulties of a likelihood-based approach for these models. For recent discussion see Barnett and Turkman (1993).

In this paper we argue that a Monte Carlo simulation approach, making use of the Gibbs sampler and Hastings-Metropolis algorithms (see, e.g., Smith and Roberts 1993, for a review) provides a viable methodology for fitting these models and constructing Bayesian estimates (which, with a diffuse prior, should also approximate maximum likelihood estimates). Space permits only a brief outline of the method here; a fuller description is in preparation for publication elsewhere. The method is applied to daily rainfall data from North Carolina, and deliberately uses a somewhat simplified form of the model; however, once the general methodology becomes established, it should be relatively easily extended to other models.

2. The model and its Bayesian analysis

Our model is as follows. All random variables are mutually independent except where specified otherwise.

---

[1] Research supported by NSF grants DMS-9115750 and DMS-9205112, and by the EPSRC

[2] Address for correspondence: Statistical Laboratory, 16 Mill Lane, Cambridge CB2 1SB, U.K.
[3] Department of Geography, University of North Carolina, Chapel Hill, N.C. 27599-3220, U.S.A.

(1) Each day is either "wet" or "dry" over the region of study. The process of wet and dry days has a first-order Markov structure where

$$\Pr\{\text{Day } n \text{ is wet} \mid \text{Day } n-1 \text{ is wet }\} = p_1,$$
$$\Pr\{\text{Day } n \text{ is wet} \mid \text{Day } n-1 \text{ is dry }\} = p_2,$$

with $0 < p_1 < 1$, $0 < p_2 < 1$.

(2) Given that a day is wet, the number of rain events N has a geometric distribution with parameter q, $0 < q < 1$, i.e.

$$\Pr\{N = k \mid N > 0\} = (1-q)q^{k-1}, \; k = 1, 2, 3, \ldots$$

Dry days, of course, correspond to $N = 0$.

(3) Given $N > 0$ on a particular day, the kth rain event for $1 \le k \le N$ is specified by an origin (X_k, Y_k) (in cartesian coordinates), a direction Φ_k (in radians, measured clockwise from north), a duration D_k and a radius R_k. It is assumed that (X_k, Y_k) are uniformly distributed over a rectangle $(x_l, x_u) \times (y_l, y_u)$ containing all the measurement stations, D_k has a gamma distribution with parameters (a_D, b_D) and R_k has a gamma distribution with parameters (a_R, b_R). These random variables can reasonably be assumed independent from one rain event to the next, but there is a difficulty with the directions Φ_k because we would expect them to be highly correlated within a particular day. We therefore adopt the following hierarchical model for the event directions $\{\Phi_k, \; k = 1, \ldots, N\}$:

(a) For each day there is a dominant direction Φ_0 which has a von Mises distribution centred at some "prevailing wind direction" Φ^* and with concentration parameter κ_0,

(b) Conditionally on Φ_0, the individual directions $\Phi_1, \ldots, \Phi_N$ are independently drawn from the von Mises distribution centred at Φ_0 with concentration parameter κ_1.

The centre of the kth rain event begins at (X_k, Y_k), moves a distance D_k in the direction Φ_k, and then dies. Throughout this time the rain event is assumed to be a circle of radius R_k, whose centre is the centre of the rain event. From this it is possible to calculate the extent to which each rain event covers each measurement station. This is defined to be the distance that the rain event moves while the measurement station is inside the event. A rain event which does not cover a particular station at all is said to have coverage extent 0 for that station.

(4) Suppose the total extent of rain events covering station j in a particular day is denoted T_j. Then conditionally on T_j, we assume the actual amount of rainfall at station j has a gamma distribution with parameters $(T_j \alpha_j, \beta_j)$. The gamma parameters (α_j, β_j) are assumed fixed but different for each station.

Evidently this model has many arbitrary features, but our main purpose here is to demonstrate the feasibility of fitting such a model rather than trying too hard to justify the model itself.

We now turn to the details of a Bayesian analysis of this system. We can think of the model conceptually as consisting of five boxes.

Box A: the top-level parameters $p_1, \; p_2, \; q, \; \Phi^*, \; \kappa_0, \; \kappa_1, \; a_D, \; b_D, \; a_R, \; b_R$.

Box B: N, Φ_0 and the rain events $\{X_k, Y_k, \Phi_k, D_k, R_k, \; 1 \le k \le N\}$ for each day.

Box C: The coverage extents T_j for individual stations.

Box D: The gamma parameters for individual stations, i.e. the constants $\alpha_j, \; \beta_j, \; 1 \le j \le m$.

Box E: The observed rain gauge data.

The logical dependence between these boxes is given by the following diagram:

$$\begin{array}{ccccccc} & & & & & & \mathrm{D} \\ & & & & & & \downarrow \\ \mathrm{A} & \rightarrow & \mathrm{B} & \rightarrow & \mathrm{C} & \rightarrow & \mathrm{E} \end{array}$$

Here arrows represent direction of dependence. Given A we can generate the contents of B, given B we can calculate C, and given C and D together we can generate E. Note that the dependence from B to C is deterministic but the other dependencies define probability distributions rather than specifying exactly the contents of the box at the head of the arrow.

We assume proper but diffuse prior distributions for each of the parameters in boxes A and D. The Bayesian inference problem is then to compute the joint distribution of the boxes A, B, C and D, given E. We approach this via Gibbs sampling: set up a Monte Carlo simulation of the whole system, and then update the contents of each of the boxes A, B, C and D in turn, conditionally on the contents of all the other boxes. For A and D, this is an application of what are by now routine Bayesian calculations, while C is a direct calculation given B. For box B itself, we adopt a method based on randomly deciding whether to add a rain event, to delete a rain event, or to do both, with probabilities $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{2}$ respectively, and then applying the Hastings (1970) rule to decide whether to accept or reject the new configuration. The whole procedure is repeated a very large number of times until it appears to converge. In the studies which follow this is based on 20,000 iterations.

3. Results

The procedure of Section 2 has been applied to five years' daily rainfall data from 13 stations in the flat coastal plain region of North Carolina. Separate analyses were performed for "summer" (June, July, August) and "winter" (December, January, February) data. One of the main features of interest is the contrast between these, since winter rainfall in North Carolina tends to consist of cyclonic events of long temporal and spatial extent, whereas summer rainfall is much more dominated by small but intense convective storms.

Fig. 1 shows a time series plot for the summer data of the principal parameters across the 20,000 iterations. In the case of the distance and radius parameters, the mean distance a_D/b_D and mean radius a_R/b_R were plotted in place of the parameters b_D and b_R. Similar plots were obtained for the winter data. Although these plots do not provide any conclusive evidence that the Markov chain has reached stationarity, the visual impressions suggests that they have.

Fig. 2 shows estimated (smoothed) posterior densities for each of the ten principal parameters, computed separately for the summer data (solid curve) and the winter data (broken curve). Despite the uncertainties over such issues as whether the model is appropriate or whether enough iterations have been taken, some interesting comparisons can be made. The p_1 and p_2 values show that the overall probability of rain somewhere in the region is substantially higher in summer than in winter. On the other hand, the summer rain events have smaller mean radius, and also cover a smaller distance before they die. This is highly consistent with our expectations, given the contrast between winter cyclonic events and summer convective events mentioned above.

One the other hand, Fig. 3 shows that not everything is satisfactory about this model. In this plot, spatial correlations based on occurrences (i.e. code the data =1 if rainfall occurs on a given day, 0 otherwise) are computed using both the real data, and simulated data from the fitted model. The simulated correlations decay much faster with distance than the true correlations. Work is currently in progress to find more realistic models that will reflect the true correlations more accurately.

References

Barnett, V. and Turkman, F. (eds.) (1993), *Statistics for the Environment, Volume 2: Water Related Issues.* Chichester: John Wiley.

Cox, D.R. and Isham, V. (1988), A simple spatial-temporal model of rainfall. *Proc. Roy. Soc. Lond. A* **415**, 317-328.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Le Cam, L. (1961), A stochastic description of precipitation. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **3**, ed. J. Neyman, Berkeley, CA, pp. 165-186

Rodriguez-Iturbe, I., Cox, D.R. and Isham, V. (1987), Some models for rainfall based on stochastic point processes. *Proc. Roy. Soc. Lond. A* **410**, 269-288.

Rodriguez-Iturbe, I., Cox, D.R. and Isham, V. (1988), A point process model for rainfall: further developments. *Proc. Roy. Soc. Lond. A* **417**, 283-298.

Smith, A.F.M. and Roberts, G.O. (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J.R. Statist. Soc. B* **55**, 3-23.

Waymire, E., Gupta, V.K. and Rodriguez-Iturbe, I. (1984), Spectral theory of rainfall intensity at the meso-β scale. *Water Resources Research* **20**, 1453-1465.

Fig. 1: Time series plots of parameters

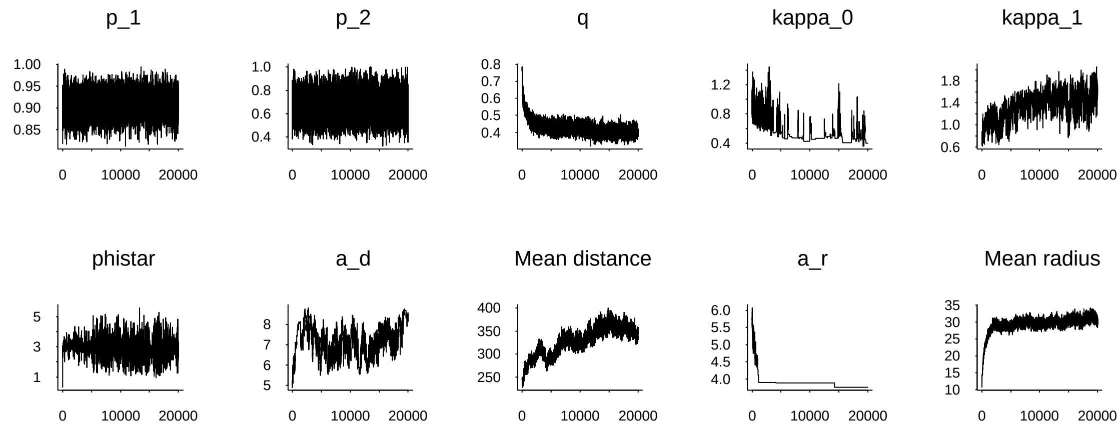

Fig. 2: Posterior densities (Solid curve summer, broken curve winter)

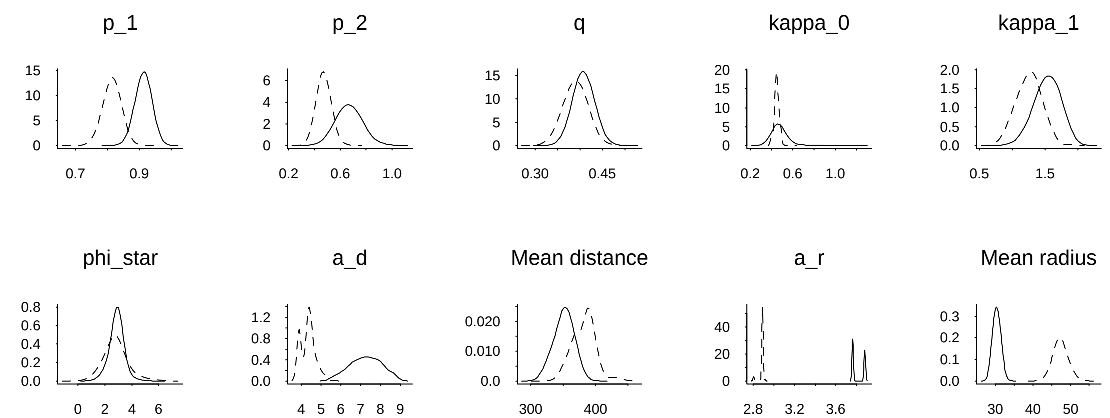

Fig. 3: Spatial correlations based on occurrences