Spatial-temporal rainfall simulation using generalized linear models

C. Yang, R. E. Chandler, and V. S. Isham

Department of Statistical Science, University College London, London, UK

H. S. Wheater

Department of Civil and Environmental Engineering, Imperial College, London, UK

Received 15 October 2004; revised 26 July 2005; accepted 10 August 2005; published 16 November 2005.

[1] We consider the problem of simulating sequences of daily rainfall at a network of sites in such a way as to reproduce a variety of properties realistically over a range of spatial scales. The properties of interest will vary between applications but typically will include some measures of ''extreme'' rainfall in addition to means, variances, proportions of wet days, and autocorrelation structure. Our approach is to fit a generalized linear model (GLM) to rain gauge data and, with appropriate incorporation of intersite dependence structure, to use the GLM to generate simulated sequences. We illustrate the methodology using a data set from southern England and show that the GLM is able to reproduce many properties at spatial scales ranging from a single site to 2000 km^2 (the limit of the available data).

Citation: Yang, C., R. E. Chandler, V. S. Isham, and H. S. Wheater (2005), Spatial-temporal rainfall simulation using generalized linear models, *Water Resour. Res.*, *41*, W11415, doi:10.1029/2004WR003739.

1. Introduction

[2] Generalized linear models (GLMs) extend the classical linear regression model, and are well established in the statistical literature. Since the pioneering work of *Coe and Stern* [1982] and *Stern and Coe* [1984], their potential for use in hydrology and meteorology has also been recognized. These authors used a two-stage approach to model both rainfall occurrences and amounts at a single site. *Chandler and Wheater* [2002] extended their work, proposing a GLM-based framework for interpreting spatial-temporal structure and applying this framework to the analysis of daily rainfall sequences in the west of Ireland. *Yan et al.* [2002] used the same framework for the analysis of daily maximum wind speed in northwestern Europe. Both of these studies have demonstrated the power of the GLM methodology for analyzing and representing complex relationships among components of the climate system.

[3] The use of GLMs to study relationships is of interest in its own right. However, their potential goes beyond this, since they are effectively probability models and can therefore be used to simulate realistic sequences of climatological and meteorological variables, incorporating the complex structures that they represent. For example, *Yan et al.* [2005] have explored the use of GLMs to simulate daily wind speed sequences at a network of European locations. *Yang et al.* [2005] give an application of GLMs to the simulation of daily sequences of potential evaporation, for hydrological applications.

[4] In this paper, we explore the use of GLMs to simulate multisite sequences of daily rainfall. Multisite rainfall simulation is a well-studied problem, and a variety of methods have been proposed for tackling it. These include transforming the rainfall distribution to approximate normality, with zeroes corresponding to negative transformed values [e.g., *Stehlík and Bárdossy*, 2002]; techniques based on resampling [*Buishand and Brandsma*, 2001]; and those based on unobserved underlying weather states [*Hughes et al.*, 1999; *Charles et al.*, 1999]. Against this background, the need for yet another simulation approach is perhaps not immediately clear. However, compared with some other methods GLMs are easily interpretable and computationally inexpensive and, in our view, this justifies their addition to the simulation toolkit.

[5] When generating multisite rainfall sequences, it is necessary to allow for the fact that neighboring sites tend to experience similar rainfall amounts on the same day, due to common exposure to a single weather system. We refer to this as ''spatial dependence''; it should be distinguished from ''systematic regional variation,'' which is the tendency of neighboring sites to share a common climate because they are in a similar location. The GLMs of *Chandler and Wheater* [2002] specify models for the marginal time series at each site, into which systematic variation is easily incorporated via the use of appropriate covariates such as site altitude. To simulate at a network of sites however, it is necessary to specify a joint distribution for all the time series. The present paper provides some suggestions for achieving this, by constructing models for spatial dependence which respect the marginal distributions from a GLM.

[6] Section 2 gives a brief introduction to GLMs for daily rainfall sequences. In section 3 we develop a framework for the modeling of spatial dependence in both rainfall occurrence and amounts. Our treatment of dependence in amounts is correlation-based, and hence can be applied

0043-1397/05/2004WR003739

quite generally. However, our model for occurrence is designed for applications where intersite dependence is strong and does not vary much with distance: it is therefore most suitable for use in catchments that are small relative to the weather systems that affect them (this is the case for almost all UK catchments, for example). Section 4 provides an example of multisite daily rainfall simulation for a catchment in southern England; and the work is summarized in section 5.

2. Generalized Linear Models for Daily Rainfall

[7] The theory of GLMs is reviewed thoroughly by *McCullagh and Nelder* [1989]; for a more introductory account, see *Dobson* [2001]. Here we outline how GLMs may be applied in hydrology or meteorology. Further details are given by *Chandler and Wheater* [2002], *Yan et al.* [2002], and *Chandler* [2005].

[8] The fundamental idea is to predict a probability distribution for each day's rainfall at every site of interest, by relating the mean of that distribution to the values of various other related quantities which we call "covariates." Possible covariates include previous days' rainfalls (possibly at more than one site), the month of the year and variables representing topographic and other location effects. Our implementation broadly follows that of *Coe and Stern* [1982] and *Stern and Coe* [1984], who adopted a two-stage approach as follows.

[9] 1. Model the pattern of wet and dry days at a site using logistic regression. Let p_i denote the probability of rain for the ith case in the data set, conditional on a covariate vector $\mathbf{x}_i$; then the model is given by

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{1}$$

for some coefficient vector $\boldsymbol{\beta}$.

[10] 2. Fit gamma distributions to the amount of rain on wet days. The rainfall amount for the ith wet day in the database is taken, conditional on a covariate vector $\boldsymbol{\xi}_i$, to have a gamma distribution with mean μ_i where

$$\ln \mu_i = \boldsymbol{\xi}_i'\boldsymbol{\gamma} \tag{2}$$

for some coefficient vector $\boldsymbol{\gamma}$. All gamma distributions are assumed to have a common shape parameter, ν say (if $\nu = 1$ the distributions are exponential). This is equivalent to assuming that, conditional on the covariates, daily rainfall values have a constant coefficient of variation [*McCullagh and Nelder*, 1989, chapter 8].

[11] These two models are referred to as "occurrence" and "amounts" models respectively. The right-hand sides of (1) and (2) are called "linear predictors. " In the GLM framework, model fitting (estimation of the coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) and selection can be carried out using likelihood methods. Models can be checked using a variety of simple but informative residual plots, illustrated in section 4. Further features include the ability to model interactions between covariates (two covariates are said to interact if the effect of one of them depends upon the value of the other), and the estimation of nonlinear transformations of covariates. Interactions can yield useful information about the mechanisms driving the rainfall process. For example, *Chandler and Wheater* [2002] found in a model for Irish rainfall that there was a significant interaction between the North Atlantic Oscillation (NAO) and covariates representing the seasonal cycle; increases in the NAO are associated with increases in winter rainfall but have little effect in the summer months. This agrees with our understanding of the NAO as a phenomenon whose effects are mainly confined to the Northern Hemisphere winter [*Hurrell*, 1995]. One of the potential advantages of the GLM methodology is that it allows us to incorporate such structures into simulated rainfall sequences.

[12] It is useful to compare the GLM approach with other commonly used models for daily rainfall data, as reviewed by *Wilks and Wilby* [1999], for example. Many such models can in fact be regarded as special cases of GLMs. Markov Chain models for rainfall occurrence can be written in the form (1) by including binary covariates representing the occurrence or not of rain on previous days [see, e.g., *Chandler and Wheater*, 2002, section 3.1]; separate Markov Chain parameters can be defined for each month of the year via an interaction between "monthly" and "previous days" covariates. As far as rainfall amounts are concerned, it is common to use a mixture of two exponential distributions as by *Wilks* [1998], fitted on a month-by-month basis with no temporal dependence. Although our amounts model (2) is not exactly of this form, the overall distribution of rainfall amounts will be a mixture of gammas whenever some covariate values change on a daily basis. Such covariates might include daily circulation pattern indices, or functions of previous days' rainfalls. In our experience, such covariates are invariably found to be statistically significant. The implication is, in agreement with many previous investigations [see *Wilks and Wilby*, 1999, and references therein], that rainfall amounts are not well modeled using a single gamma distribution and that a mixture of distributions is more realistic.

3. Spatial Dependence Structure

[13] Models (1) and (2) specify probability distributions for daily rainfall at individual sites, conditioned on the values of various covariates such as previous rainfalls, time of year and "external" factors such as the NAO. A single-site sequence can then be simulated, given some initial conditions, by sampling a value from the first day's distribution, using this value to construct a distribution for the second day, sampling a new value for the second day and so on. Typically, the required initial conditions are provided by a few days' observed data.

[14] For the generation of simultaneous rainfall sequences at several sites, in general it will be necessary to account for dependence between the sites, unless they are widely separated in space. The single-site simulation procedure must now be modified: instead of specifying individual distributions for the next day's rainfall at each site, it is necessary to specify a joint distribution for the next day's rainfalls at all sites. We proceed in two stages, first defining a joint distribution for the wet-dry pattern of rainfall occurrence, and then the distribution of the rainfall amounts vector at the wet sites. The mean vectors of these joint

distributions are specified by the GLMs of the previous section.

3.1. Rainfall Amounts

[15] We deal with the joint distribution of amounts first, since this is easier to specify than that for occurrence. It is convenient to proceed via a transformation to marginal normality at each site, since in this case the spatial dependence is conveniently summarized using the intersite correlation structure of the transformed values. If Y is a continuous random variable with distribution function $F(y) = P(Y \leq y)$, then an exact normalizing transformation is given by $Z = \Phi^{-1}[F(Y)]$, where $\Phi[.]$ is the distribution function of the standard normal distribution. In this case, a dependent vector $\mathbf{Y}$ could be generated by simulating a multivariate normal random vector $\mathbf{Z}$, and then setting $Y = F^{-1}[\Phi(Z)]$ for each element of $\mathbf{Z}$. However, when (as here) the Ys have gamma distributions, the evaluation of both F^{-1} and Φ is relatively expensive computationally. This can slow down simulations, which is a disadvantage if large quantities of data are to be generated. As an alternative, we note that if Y is gamma distributed, the distribution of $Y^{1/3}$ is normal, to a degree of approximation described as "absurdly accurate" by *Terrell* [2003]. In fact, if Y_i is the observed amount for the ith wet day in a database, and μ_i is the modelled mean of the gamma distribution for that case, then the quantities

$$r_i^{(A)} = (Y_i/\mu_i)^{1/3} \tag{3}$$

all share the same normal distribution, approximately. The mean and variance of this normal distribution can be calculated numerically, as described in section 3.3.2 of *Chandler and Wheater* [1998]. The calculation needs to be carried out only once in any simulation, since the mean and variance depend on ν, the common shape parameter of the gamma distributions, but not upon the means $\{\mu_i\}$.

[16] The quantities $\{r_i^{(A)}\}$ are called Anscombe residuals. Since they are approximately normal, it is natural to describe spatial dependence via a model for their intersite correlations (although this does not necessarily capture all of the dependence structure, because it is not guaranteed that the joint distribution of several Anscombe residuals is multivariate normal). One might choose to use a standard geostatistical model to represent these correlations as a function of intersite distance and direction [*Cressie*, 1991], or simply to calculate the empirical intersite correlation matrix from historical data. Note also that the normality of the Anscombe residuals can be used to check the assumption of gamma distributed rainfall amounts; if this assumption holds, a normal quantile-quantile plot of the Anscombe residuals should appear as a straight line except for the smallest rainfall amounts, as illustrated by *Chandler and Wheater* [2002].

[17] Simulation of rainfall amounts at "wet" sites for a particular day therefore proceeds: first, by sampling a vector of Anscombe residuals from a multivariate normal distribution with an appropriate mean and covariance structure (standard algorithms exist for this) and, second, by inverting the transformation (3) at each site. If any negative values are generated, they are discarded and a new vector is drawn. One potential drawback is that the scheme takes no account of "dry" sites, since the amounts model is only defined at sites where the rainfall amount is nonzero. Therefore it is not guaranteed to produce small amounts of rain near sites which are dry (a phenomenon labeled "spatial intermittence" by *Bárdossy and Plate* [1992]), although wet sites which are close to each other will tend to have similar rainfall amounts. There are, however, few currently available simulation methods that overcome this problem. One approach, adopted by *Stehlík and Bárdossy* [2002], for example, is to treat the combined occurrence-amounts process as a transformation of a single underlying Gaussian field, with nonzero rainfall occurring only when this field exceeds some threshold. The spatial correlation structure of the field ensures a tendency for small rainfall values to appear close to zeroes. One drawback of this approach is the implication that occurrence and amounts have the same underlying structure: this is questionable on both physical and empirical grounds. For example, in section 4 below, we find that systematic regional variation is different for the two processes. This is a typical finding [see, e.g., *Chandler and Wheater*, 2002, Figure 5].

[18] Alternative solutions to this problem have been suggested by *Wilks* [1998] and by *Charles et al.* [1999]. Wilks suggested a scheme that effectively amounts to the use of dependent sets of pseudorandom numbers to generate the two components (occurrence and amounts) of the rainfall process. Within his modeling framework, such an approach is relatively straightforward to implement and, although there is no particular theoretical justification for the precise form of dependence used, results are encouraging. The approach taken by Charles et al. was to regress (transformed) amounts at a site upon rainfall occurrence at neighboring sites. Of the various alternatives available, this is perhaps the most readily incorporated into our own framework. It would be of interest to explore these ideas within the context of the models we consider; however, such an exercise is well beyond the scope of the present paper. In any case, the importance of spatial intermittence in determining hydrological response is not clear: it will presumably be application-dependent.

3.2. Rainfall Occurrence

[19] For the binary rainfall occurrence model a transformation to marginal normality is not possible and an alternative approach to the modeling of spatial dependence is required. Possible approaches include specifying the dependence via the correlation structure of the wet/dry field, in the spirit of *Oman and Zucker* [2001] and *Lunn and Davies* [1998]; discretizing a continuous-valued process with appropriate dependence structure, as by *Emrich and Piedmonte* [1991]; specifying the dependence via odds ratios between pairs of sites [*Cox and Wermuth*, 1996, section 3.7]; conditioning on a hidden "weather state" variable as by *Hughes et al.* [1999]; and including simultaneous rainfalls at other sites as extra covariates in the logistic regression model. There are drawbacks to many of these approaches, however. For example, the correlation between binary variables is constrained by the marginal probabilities: since the probability of rain at each site changes each day, the correlation structure must therefore change as well, and it becomes difficult to specify a plausible correlation-based model for the dependence. Other approaches suffer from computational cost, either in esti-

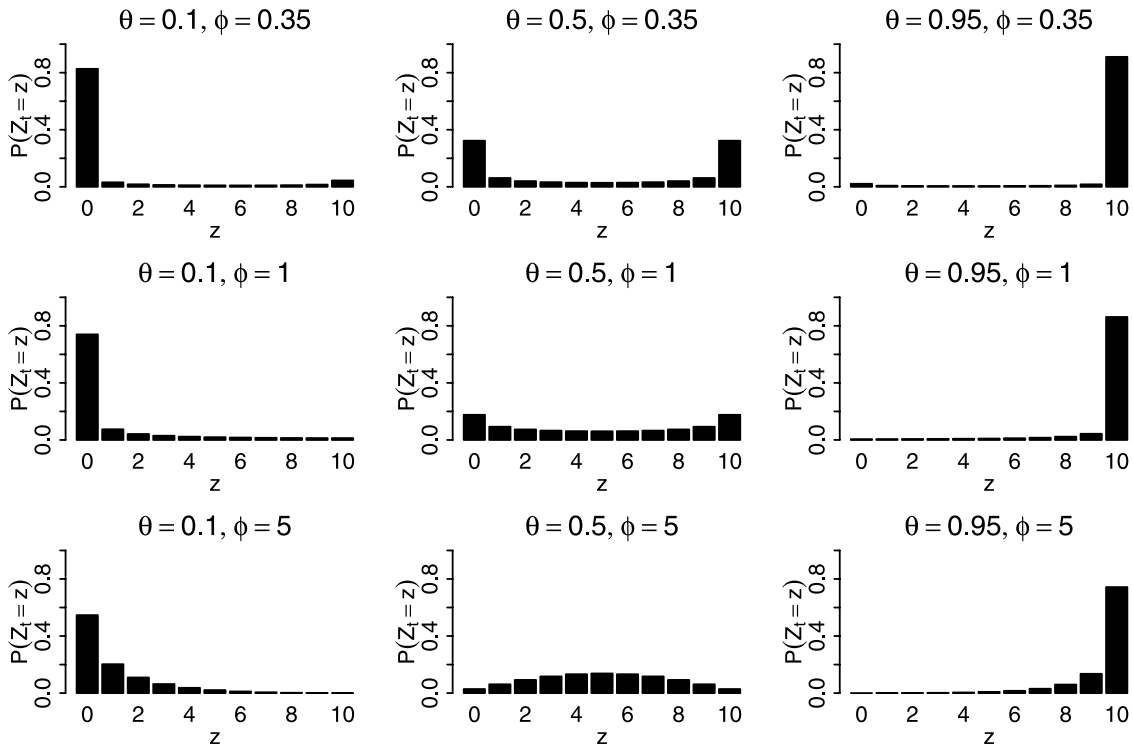

Figure 1. Examples of beta binomial distributions when $S = 10$. Each row corresponds to a fixed value of the dispersion parameter ϕ; each column corresponds to a fixed value of the mean parameter θ.

mating or in simulating the correlation structure. More details are given by *Wheater et al.* [2000, chapter 4].

3.2.1. Modeling the Number of Wet Sites

[20] A particular feature of space-time rainfall is that, at typical scales of hydrological interest, intersite dependence is strong and sites tend to be either mostly wet or mostly dry. This reflects the fact that all sites tend to be influenced by the same weather systems on particular days. For hydrological purposes, it may be important to reproduce accurately the distribution of numbers of wet sites, since this is related to the proportion of an area which experiences rain. Our experience is that it can be difficult to reproduce the shape of this distribution well, using simple versions of the methods outlined above. We therefore explore here the idea of modeling the distribution directly. This approach has been used in other applications, notably in the analysis of teratology (developmental toxicity) data [see, e.g., *Ryan*, 1995, and references therein]. It has the advantage of being both conceptually straightforward and computationally feasible. At the present stage of development, a potential drawback is that there is no concept of intersite distance, except as implied by the regional variation of the probabilities from the rainfall occurrence model (1). This is probably acceptable at spatial scales where intersite dependence is uniformly high, especially since it is difficult to obtain very strong dependence using other methods. The model may not be suitable, however, at larger scales where there is substantial variation of intersite dependence with distance.

[21] We begin by establishing some notation. We wish to simulate, for day t, a vector of dependent binary random variables at S sites, $\mathbf{Y}_t = (Y_{1t} \ldots Y_{St})'$ say. The rainfall occurrence model (1) allows us to calculate $E(Y_{st}) = p_{st}$, say. A (nonunique) dependence structure can be specified for $\mathbf{Y}_t$ through the distribution of $Z_t = \sum_{s=1}^{S} Y_{st}$. Since the $\{p_{st}\}$ vary from day to day, so does the distribution of Z_t; in particular, we have $\mathrm{E}(Z_t) = \sum_{s=1}^{S} p_{st}$.

[22] A flexible family of distributions for discrete random variables taking values in $\{0, 1, \ldots, S\}$ is the beta-binomial:

$$P(Z_t = z) = \binom{S}{z} \frac{\Gamma(\alpha_t + z)\Gamma(S + \beta_t - z)\Gamma(\alpha_t + \beta_t)}{\Gamma(\alpha_t + \beta_t + S)\Gamma(\alpha_t)\Gamma(\beta_t)} \qquad (4)$$

for $z = 0, 1, \ldots, S$. The parameters of the distribution are $\alpha_t \in \mathbb{R}^+$ and $\beta_t \in \mathbb{R}^+$. The mean and variance are

$$\frac{S\alpha_t}{\alpha_t + \beta_t} \quad \text{and} \quad \frac{S\alpha_t\beta_t(\alpha_t + \beta_t + S)}{(\alpha_t + \beta_t)^2(\alpha_t + \beta_t + 1)}, \qquad (5)$$

respectively. It is convenient to reparameterize the distribution here: set

$$\theta_t = \frac{\alpha_t}{\alpha_t + \beta_t} \quad \text{and} \quad \phi_t = \alpha_t + \beta_t, \qquad (6)$$

so that

$$\mathrm{E}(Z_t) = S\theta_t \quad \text{and} \quad \mathrm{Var}(Z_t) = \frac{S\theta_t(1 - \theta_t)(\phi_t + S)}{\phi_t + 1}. \qquad (7)$$

We can think of $\theta_t = S^{-1}\,\mathrm{E}(Z_t) = S^{-1}\sum_{s=1}^{S} p_{st}$ as a mean parameter, which is determined by the ps from the rainfall occurrence model. As $\phi_t \to 0$, the distribution becomes increasingly concentrated around 0 and 1 (see Figure 1); as $\phi_t \to \infty$ the distribution tends to the binomial, with parameters S and θ_t. Since the binomial distribution arises if

all the Ys are independent and identically distributed, ϕ_t can be regarded as an overall summary of spatial dependence, with small values corresponding to strong dependence.

[23] As a first attempt at modeling in this way it is convenient, and not implausible, to assume that $\phi_t = \phi$ is constant for all t, so that θ_t is the only time-varying parameter of the distribution. As θ_t varies, the effect therefore is to move along one of the rows of Figure 1; in this way we hope to reproduce typical "summer" and "winter" distributions of numbers of wet sites, for example.

[24] In standard applications, the beta-binomial distribution arises as that of a binomial (S, η) random variable, where η is itself a random variable distributed according to a beta distribution with parameters α_t and β_t. Thus the following simple mechanism would give rise to a beta-binomial distribution for the number of wet sites in a fixed region sampled at S locations: a proportion η_t of the region is wet, where η_t is a beta-distributed random variable. Given η_t, individual locations are wet or dry independently of each other. Although this mechanism is clearly idealized, it provides a useful insight into the model. In particular, it suggests that the model may fail if sites are too close together, since in this case the assumption of conditional independence given η_t is unlikely to hold even approximately.

3.2.2. Estimation

[25] Given data $\{(S_t, Z_t, \theta_t = \sum_s p_{st}/S_t) : t = 1, \ldots, T\}$ (we write S_t here because in practice it is unlikely that the number of sites yielding data will be the same for all t), the parameter ϕ can be estimated using a method of moments. Let $R_t^2 = (Z_t - S_t\theta_t)^2/[S_t\theta_t(1 - \theta_t)]$; then, from (7), we have

$$\mathrm{E}\left(R_t^2\right) = \frac{\phi + S_t}{\phi + 1},$$

suggesting the estimator

$$\hat{\phi} = \frac{\sum_{t=1}^{T}(S_t - 1)}{\sum_{t=1}^{T}\left(R_t^2 - 1\right)} - 1. \tag{8}$$

The parameterization in terms of θ_t and ϕ, and proposal for a moment-based estimate of ϕ, are similar to those given by *Williams* [1982].

3.2.3. Simulation

[26] Having specified a plausible model for the distribution of Z_t, a natural strategy for simulation is to sample the number of wet sites from this distribution and then to allocate the positions of these wet sites. However, this needs to be done in such a way as to reproduce correctly the marginal probabilities of rain at each site, according to the rainfall occurrence model. Define $w_{s,z,t} = P(Y_{st} = 1 \text{ and } Z_t = z)$. Then we have

$$P(Y_{st} = 1) = \sum_{z=0}^{S} w_{s,z,t}. \tag{9}$$

The marginal probabilities will be reproduced correctly if (9) yields the value p_{st} for each s; therefore we seek an assignment of the probabilities $\{w_{s,z,t} : s = 1, \ldots, S; z = 0, \ldots, S\}$ that will ensure this. This assignment may not be unique, but this is not a problem since the aim is merely to reproduce accurately the distribution of the number of wet sites, while at the same time preserving the probabilities of rain at each site. Perhaps a more serious problem is that the postulated distribution of Z_t is not guaranteed to be compatible with the modelled marginal probabilities; obvious examples of incompatibility arise when $P(Z_t = S) > \min_s P(Y_{st} = 1)$ and when $P(Z_t = 0) > \min_s P(Y_{st} = 0)$, for example. In practice however, we have only ever encountered incompatibility in cases where the occurrence model (1) generated ps at one or two sites that differed substantially from the majority. This type of behavior is unrealistic, and is symptomatic of a poor occurrence model. In the work reported in section 4 below we have simulated almost four thousand years of daily rainfall at 10 sites, without once encountering this problem. Our simulations use an efficient algorithm for finding a valid set $\{w_{s,z,t}\}$ if it exists. Unfortunately, the details of this algorithm are far too lengthy to include here; a full description is given by *Chandler* [2002, appendix].

3.3. Imputation

[27] Daily rainfall records often contain missing values. If there are many missing values in a record, there will be considerable uncertainty regarding the historical values of various rainfall summary statistics. If we can determine the distribution of these missing values conditional upon the observed values at all sites, then we can simulate from this conditional distribution many times to construct uncertainty envelopes for historical rainfall statistics. We refer to this process as imputation. It can be an extremely helpful aid to the interpretation of historical records.

[28] The spatial dependence structures proposed here, for both amounts and occurrence models, are specified in such a way that imputation is straightforward. For the occurrence model, the conditional distribution emerges naturally as a by-product of the algorithm for allocating the locations of wet sites given Z_t, as described by *Chandler* [2002]. For the amounts model, dependence is specified via a multivariate distribution for the Anscombe residuals. If data from some sites are missing, but others are observed, then Anscombe residuals can be computed from the observed sites and the conditional distribution of the missing residuals, which remains multivariate normal, can be calculated [*Krzanowski*, 1988]. Missing residuals can then be simulated from this conditional distribution, and back transformed to yield imputed rainfalls.

4. Example

[29] In this section we illustrate the theory described above, by fitting GLMs to a daily rainfall data set and simulating the resulting models to evaluate their performance.

4.1. Data Overview

[30] The data used in this example are from a network of 34 gauges run by the UK Meteorological Office. The gauges are in a 50 km × 40 km region in southern England. The area is relatively flat; gauge altitudes range from 10 to 170 m above sea level. The earliest record starts in 1904, and data from some gauges are available until 2000.

[31] The data were checked carefully before use. Several of the "daily" values turned out to be monthly totals, and were discarded from the subsequent analyses. Moreover, the recording resolution changed from 0.3 to 0.1 mm during the early 1970s. This change may create illusory trends, particularly in rainfall occurrence, and should therefore be taken into consideration during modeling. Finally, there were spatial inconsistencies among the records, mainly relating to the recording of small rainfall amounts ("trace values"). A simple but effective solution to this problem is to threshold the data prior to modeling: we fit GLMs to the quantities

$$Y_{st}^{*} = \max(Y_{st} - \tau, 0) \tag{10}$$

for some threshold $\tau > 0$. The fitted models may then be used to simulate daily sequences of thresholded values, and the thresholding removed by computing

$$\tilde{Y}_{st} = \begin{cases} Y_{st}^{*} + \tau & Y_{st}^{*} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

In this procedure, the simulated output will contain no values between 0 and τ. However, for practical applications this will not cause problems providing the threshold is small enough. Our experience is that setting a small threshold can remove many apparent inconsistencies between gauges, without appreciably affecting the important properties of the sequences. The modeling task is simplified considerably if the inconsistencies are removed. In the work reported here, the threshold τ has been set to 0.5 mm.

4.2. Fitted Models

[32] To represent rainfall occurrence a logistic model of the form (1) has been fitted to the thresholded data from all sites. The model contains 19 terms, selected using a combination of formal tests and residual analyses in an approach similar to that described by *Chandler and Wheater* [2002]. Systematic regional effects are represented using site altitude as a covariate, as well as Legendre polynomial transformations of site eastings and northings (these are defined in such a way as to be approximately uncorrelated if the sites are uniformly distributed on the study region [see *Chandler*, 2005]). Seasonality is represented using sine and cosine terms; temporal dependence and persistence are modelled using indicators for three previous days' rainfall so that the structure is effectively a generalized Markov chain, as described in section 2. We refer to these indicators as "autoregressive terms."

[33] In addition to this basic structure, the model contains terms representing the effect of the NAO, which is generally regarded as the most important large-scale structure affecting climate in Europe [*Barnston and Livezey*, 1987]. The monthly NAO index used here is the extended version defined by *Jones et al.* [1997]. The fitted model involves significant interactions between the NAO and seasonal terms; this reflects the fact that the NAO is more strongly associated with UK rainfall in winter than in summer.

[34] Finally, to allow for the change in recording resolution in the early 1970s, the model includes an indicator variable defined as

$$I_{s,t} = \begin{cases} 1 & \text{for all observations before 1975} \\ 0 & \text{for all observations from 1975 onward.} \end{cases} \tag{12}$$

The change in resolution means that terms representing previous days' rainfall have a different meaning before and after 1975. This can be accommodated by considering interactions between the adjustment indicator (12) and any autoregressive terms. The occurrence model used here contains just the interaction involving the lag 1 autoregressive term; other interactions involving the adjustment indicator were not significant.

[35] For rainfall amounts, the gamma model (2) has been fitted to the same data set. The covariates are similar to those in the occurrence model, but fewer in number (only 13 are included). The reduction is mainly due to simpler site effects, implying that rainfall amounts are less affected by spatial variations within the region than rainfall occurrence. Additionally, the amounts model contains no terms involving adjustment indicators of the form (12), since these terms were not significant. The shape parameter is estimated as 0.6873.

[36] The fit of either model can be assessed by plotting mean Pearson residuals by month, site and year. The Pearson residual for an observation Y is proportional to

$$\frac{Y - \mathrm{E}(Y)}{\sigma} \tag{13}$$

where $\mathrm{E}(Y)$ and σ are the expected value and standard deviation of Y under the fitted model. If the model is correct, all Pearson residuals come from distributions with mean zero and constant variance. To illustrate their use, Figure 2 shows monthly and annual residual plots for the amounts model. There is some suggestion of a weak trend in the annual means (a block of predominantly negative values between 1940 and 1960) and some seasonality in the monthly standard deviations; however, neither effect is particularly large. There is no structure in the corresponding plots for the occurrence model (not shown). Overall, these analyses suggest that the models represent well the seasonal structure and trends in the rainfall sequences. The clearest residual structure is the seasonal variation in standard deviations for the amounts model. Although not substantial, this variation suggests that the model could be improved by relaxing the assumption of a constant shape parameter. This assumption may affect the ability of the model to reproduce aspects such as extremes; this is investigated below.

[37] Figure 3 shows mean Pearson residuals at each site for both occurrence and amounts models. If the model captures the spatial pattern correctly, the mean residuals should differ significantly from zero at about 5% of the sites (indicated by thick circles in the plots). It is clear that this is not the case. However, neither plot shows any systematic structure. For example, in the occurrence model the mean residuals at neighboring sites B19 and B20 are both significantly different from zero, but have opposite signs. This kind of spatial inconsistency cannot be modelled using a smooth surface. The inconsistencies are worse for the

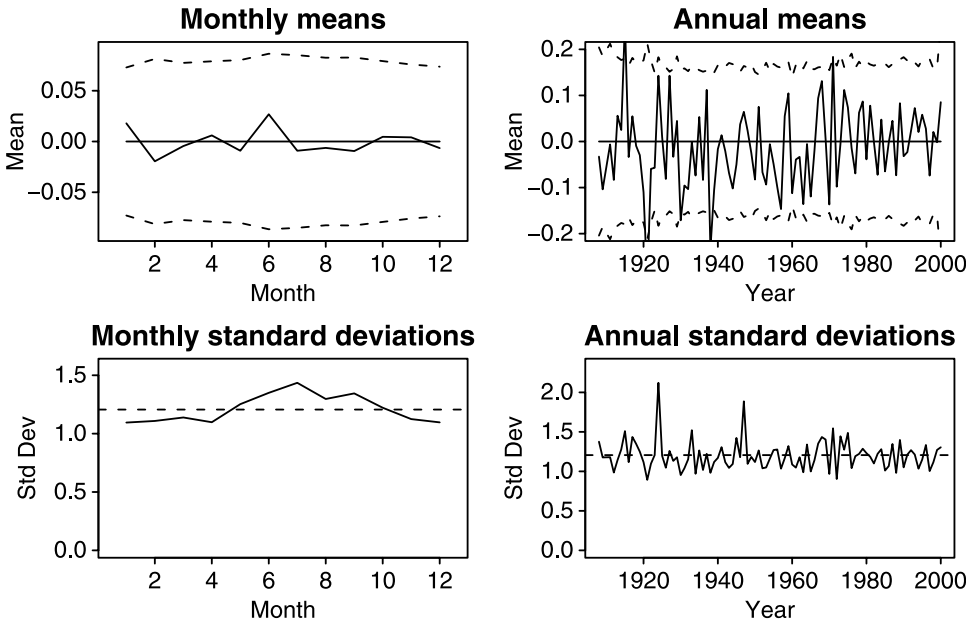

Figure 2. Mean and standard deviation (across all sites) of Pearson residuals for amounts model by month and year. In Figure 2 (top), dashed lines show 95% limits, adjusted for intersite dependence, under the assumption that the model is correct. In Figure 2 (bottom), dashed lines show the standard deviations expected under the model.

occurrence model than for the amounts; this suggests that problems with small values remain even after thresholding. However, the results here improve sufficiently upon those for unthresholded data (not shown) that the thresholding must be regarded as worthwhile.

[38] As described in section 3, a normal quantile plot of Anscombe residuals has been used to check the gamma distributional assumption for the amounts model. The plot is not shown here; it looks almost identical to that presented in Figure 4 of *Chandler and Wheater* [2002], and indicates that the gamma distribution fits extremely well.

4.3. Spatial Dependence Structure

[39] As described in section 3, when simulating sequences at a network of sites it is necessary to represent the dependence between them. Here, dependence in rainfall

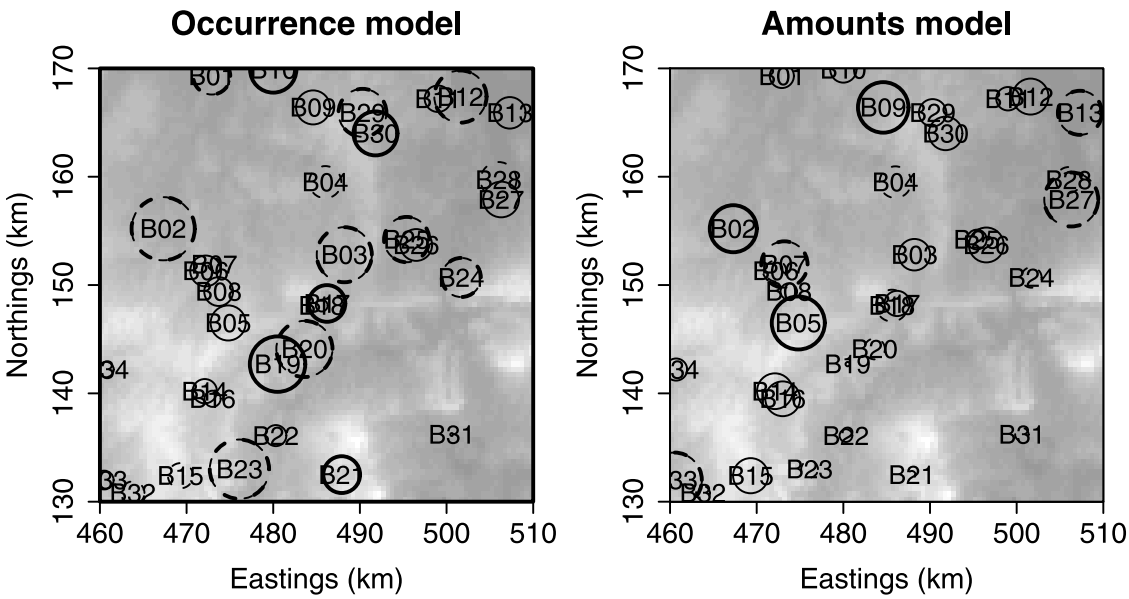

Figure 3. Mean Pearson residuals from GLMs fitted to thresholded data (threshold = 0.5 mm). Solid (dashed) circles represent positive (negative) residuals. The radii of the circles are proportional to the mean values of residuals. Thick lines indicate mean residuals that differ significantly from zero at the 5% level. The background images show the topography of the area; darker shading corresponds to lower ground. Gauge altitudes range from 10 to 170 m above sea level at sites B13 and B15, respectively.

amounts is modelled via correlation between Anscombe residuals. As indicated previously, it is natural to consider standard geostatistical models for the intersite correlation structure. However, in this particular example all of the correlations are similar (90% of them are between 0.76 and 0.94), which is to be expected since the study area is small relative to most weather systems. In view of this, in our opinion the additional complexity of a distance-dependent correlation model is not warranted in this case. In the simulations reported below we have therefore used a common correlation, estimated as 0.786, between each pair of sites. To check the sensitivity of the results to this simplification, we have repeated all of the analyses for a further set of simulations using the exact residual correlations between each pair of sites. All of the results were, in practical terms, indistinguishable from those reported below. We conclude that at the spatial scales considered here, the precise choice of intersite correlation structure is relatively unimportant.

[40] To summarize the dependence in rainfall occurrence, 10 sites have been selected that have few missing values over the period 1961–1999, and for each month of the year the frequency distribution of the number of wet sites has been tabulated (excluding days when any site had missing data). The proportion of days on which 2–8 of the 10 sites experienced rain ranges from just 0.16 (in September) to 0.21 (in May). Hence, as far as occurrence is concerned, the variation of intersite dependence with distance is of little interest in this particular example. We therefore model the dependence using the beta binomial structure described previously. The parameter ϕ is estimated, using (8), to be 0.359. The distributions of numbers of wet sites will therefore look similar to those in Figure 1 (top).

4.4. Simulation

[41] The previous residual analyses indicated that overall, the fitted models capture well the systematic structure in rainfall sequences. This does not necessarily guarantee good simulation performance, however, since small errors in model specification may cumulate over a long period of time. Moreover, the occurrence and amounts models have been assessed individually rather than in combination, and the simultaneous performance at several sites has not been investigated. To address these issues, the fitted models have been used to generate simulated sequences, and properties of these sequences have been compared with those of the observations at a variety of spatial scales. Reasonably complete observational series are needed for this exercise. To meet this requirement, simulations were carried out over the 1961–1999 period for the same 10 sites used in section 4.3 to assess dependence in rainfall occurrence.

4.4.1. Summary Statistics

[42] Prior to simulation, 10 sets of imputations were carried out, replacing any missing observations by simulated values conditional on the observations as described in section 3.3. For each set of imputed data, summary statistics were calculated; the range of each set of statistics indicates the uncertainty due to missing data. The statistics considered were: mean, standard deviation, proportion of wet days (i.e., proportion of days with nonzero rainfall after thresholding), conditional mean and standard deviation (computed for wet days only), maximum and autocorrelations at lags 1 and 2. Statistics were calculated separately for each month of the year, for each individual site and for daily time series obtained by averaging over groups of sites. These average series can be regarded as estimates of areal mean rainfall at scales up to 2000 km^2, and can be used to assess the appropriateness of the spatial dependence structures used in the simulations.

[43] 100 sets of simulated data were then generated at the same 10 sites, to simulate the modelled dynamics of the rainfall processes during the 1961–1999 period. Each simulation was initialized using the historical data for December 1960, and was conditioned on the historical NAO series. Summary statistics were calculated for each simulation, to yield a simulated distribution for each statistic. If the simulations are realistic, the observed values of the statistics should look like samples from these simulated distributions.

[44] Figure 4 shows the results for the regional average series (i.e. mean of all 10 sites). Similar results are obtained for other groups of sites, and for individual sites. The worst individual statistic is the conditional mean, for which the simulated distributions are slightly too high. This probably reflects a slight misspecification of the nonlinear autoregressive structure in the amounts model. Elsewhere, there are some isolated discrepancies (e.g., the standard deviation on wet days in September, and lag 1 autocorrelation in August), and a tendency for the simulated maxima to peak later in the year than the observations. The latter result is consistent with the previous analysis of Pearson residuals, which suggested that the constant shape parameter in the amounts model may lead to an underestimation of variability in summer and overestimation in winter. However, overall the observed structure is well reflected in the simulated distributions. Note in particular that the simulations reproduce features such as the seasonal variation in the degree of autocorrelation; this is achieved via the interaction terms in the models.

4.4.2. Simulated Rainfall Distributions

[45] As well as examining rainfall summary statistics, it is of interest to compare the overall shape of the simulated and observed rainfall distributions on wet days. This can be achieved by plotting the quantiles of the observed data against those of the simulations. Results, for a single site (code B18; see Figure 3) and for the areal average, are given in Figure 5. Site B18 is located in the center of the study area, with a small mean Pearson residual for the amounts model. Therefore any discrepancies here are due to the way in which the individual daily gamma distributions are mixed in the simulations, rather than to any systematic bias in the individual distributions themselves.

[46] The plots in Figure 5 are typical and indicate good agreement overall between observed and simulated distributions (in the upper tails, some departures from the line of perfect agreement are to be expected due to sampling variability). However, when the distributions are split by month it becomes apparent that the simulations systematically overestimate the highest quantiles in January and underestimate in July. The latter problem seems particularly acute for the areal average, although in fact it affects only a small proportion of the distribution (in the areal average July plot, just 2.5% of observed wet day amounts exceed 21 mm, which is the point where the agreement breaks down;

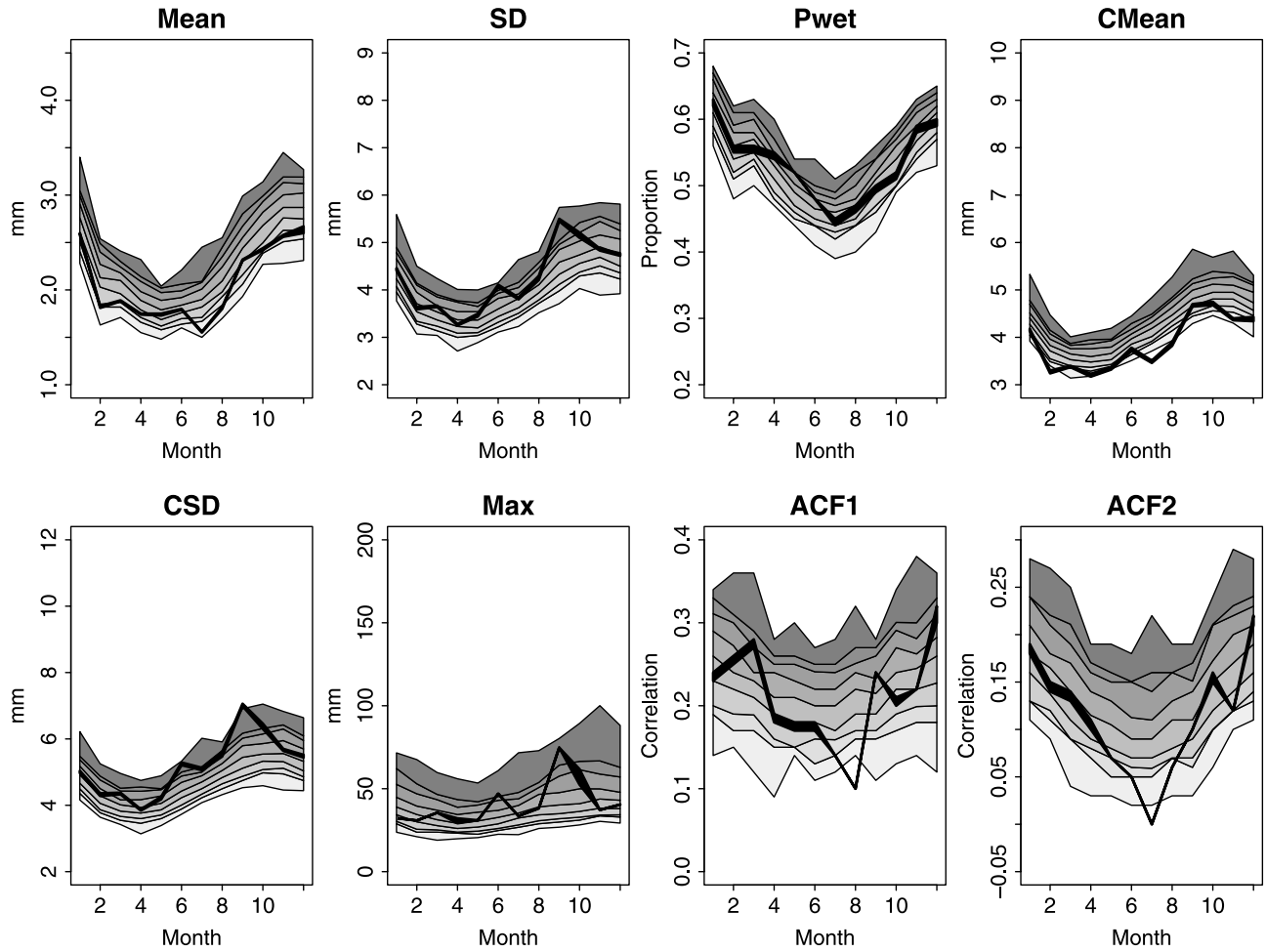

Figure 4. Observed and simulated monthly summary statistics for 10-gauge average daily series (models fitted to thresholded data). Thick lines show the envelope obtained from 10 sets of imputations of missing data; shading shows the range of the simulated distributions along with 5th, 25th, 50th, 75th, and 95th percentiles. Mean, standard deviation, proportion of wet days, and conditional mean (i.e., mean on wet days only) are shown from left to right in the top row, and conditional standard deviation, maximum, and autocorrelation at lags 1 and 2 are shown from left to right in the bottom row.

this corresponds to around 0.5% of all July days). Once again, the discrepancy is probably due to the assumption of a constant shape parameter in the amounts model. The good agreement between overall distributions suggests that the difficulty may be resolved if this assumption can be relaxed.

4.4.3. Seasonal Rainfall Totals

[47] Although the NAO contributes to both occurrence and amounts models, it explains only a fraction of a percent of the variance in the daily rainfall sequences. However, at a seasonal scale the effect is much more apparent. To illustrate this, Figure 6 shows the observed time series of summer (June, July, and August) and winter (December, January, and February) rainfall totals, averaged over the 10 sites used in the simulation. For each year, the distributions of simulated seasonal totals are also shown. Overall, the simulated distributions seem more or less consistent with the observations. This is not surprising, given that the simulations reproduce the mean structure of the time series as shown above. Of more interest is the fact that that the simulated summer distributions show little interannual variability, whereas the winter distributions are more erratic. The NAO provides the only possible source of interannual variability in the simulations; accordingly we conclude that it is responsible for the variation in winter distributions but has little effect upon summer rainfall. This is in agreement with previous studies [e.g., *Hurrell*, 1995] and could also have been deduced by examining its contribution to the linear predictors in the models, as by *Chandler and Wheater* [2002]. However, the simulations also indicate that, despite the very weak signal in the daily time series, the simulations generate a plausible level of interannual variation in winter rainfalls. Note in particular that the observations follow the simulated distributions closely in the latter half of the 1990s, suggesting that winter rainfalls during this period were strongly associated with variation in the NAO. There are places where the observations do not follow the simulated distributions particularly closely, for example, in 1990 and 1992, and this suggests the existence of other factors affecting winter rainfall in the area, that are not accounted for by our models. A possible candidate is the east Atlantic pattern (EA), defined by *Barnston and Livezey* [1987]. *Murphy and Washington* [2001] reported that the EA is

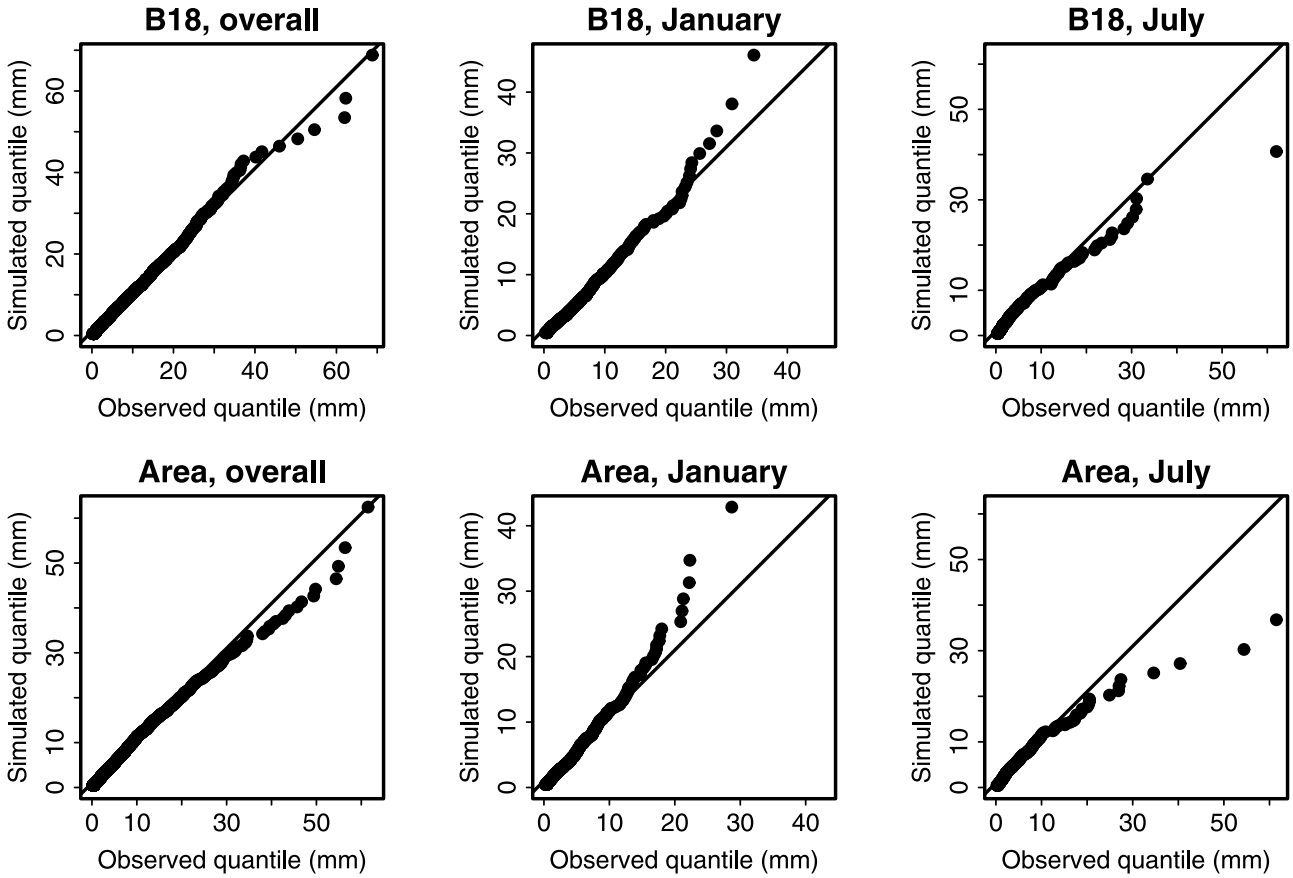

Figure 5. Quantile-quantile plots of observed and simulated daily wet day rainfall distributions for (top) a single site and (bottom) a 10-site average. (left) Overall rainfall distributions for each series and the distributions during (middle) January and (right) July.

more closely related than the NAO to rainfall variability in the British Isles, particularly in the southeast, from September to April. Our own calculations support this: for example, the correlations with December total rainfall in our data set are 0.375 and 0.146 for the EA and NAO respectively. Unfortunately, however, the EA is not defined between May and August, which makes it difficult to incorporate directly into a GLM. Further research is required to investigate this.

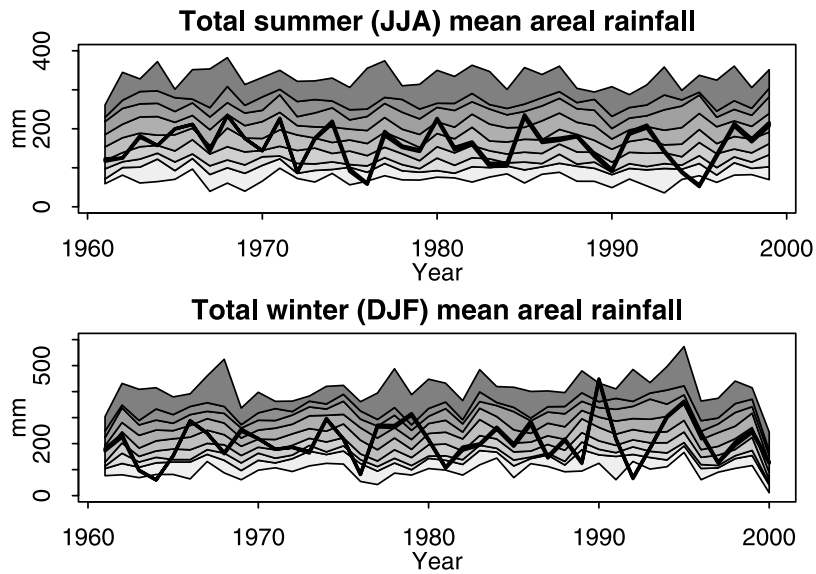

Figure 6. Observed time series (thick solid line) and annual simulated distributions of summer (June, July, and August) and winter (December, January, and February) rainfall, averaged over 10 sites.

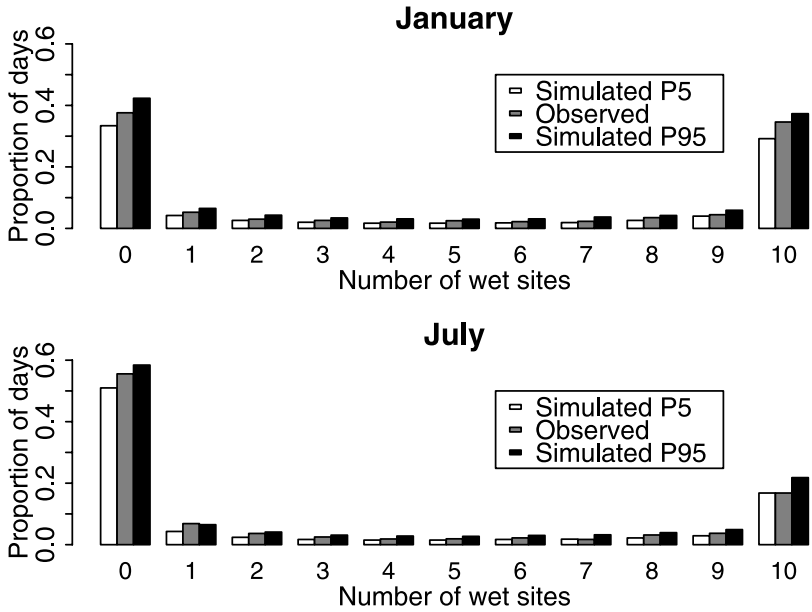

Figure 7. Distributions of numbers of wet sites for (top) January and (bottom) July. Middle bars indicate observed frequencies; left and right bars are 5th and 95th percentiles of the distribution of simulated frequencies, respectively.

4.4.4. Numbers of Wet Sites

[48] The results discussed so far indicate that the fitted models are able to reproduce various properties of rainfall, at spatial scales ranging from a single site to a 10-site average. This indicates that the models' representation of spatial dependence is adequate. As an additional check on the use of the beta-binomial dependence mechanism, Figure 7 shows the observed and simulated distributions of numbers of wet sites. The variability in the simulations is illustrated by calculating frequencies individually for each realization, and displaying the 5th and 95th percentiles of each frequency. The agreement between observations and simulations is excellent throughout. Of course, this does not represent an independent verification of the beta binomial model, whose parameters have been chosen to fit this particular set of data. However, it does indicate that the distribution is capable of reproducing observed histograms; moreover, the good agreement for both summer and winter suggests that the assumption of constant ϕ is reasonable.

4.4.5. Extremes

[49] As a final check on the simulations we examine their extremal behavior, since this is of primary importance if the models are to be used in applications such as flood risk assessment. In Figure 4 the comparison of observed and simulated maxima in each month of the year provides a preliminary confirmation that extremes are reproduced fairly well. A more sophisticated approach is to compare the simulations with the results of a classical extreme value analysis. For each of the observed daily time series studied here, a generalized extreme value (GEV) distribution has been fitted to the annual maxima using maximum likelihood, as described by *Coles* [2001]. The GEV distribution function (the probability that the annual maximum daily rainfall does not exceed m, say) is

$$G(m) = \Pr(M \le m) = \exp\left\{-\left[1 + \xi\left(\frac{m-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \qquad (14)$$

for all m with $1 + \xi(m - \mu)/\sigma > 0$. The parameters μ and σ control the location and scale of the distribution, and ξ is a shape parameter. If $\xi = 0$ a Gumbel distribution is obtained, if $\xi < 0$ the distribution has a finite upper bound, and if $\xi > 0$ it is heavy tailed. The value of ξ is often regarded as an important summary measure of extremal behavior in risk assessment, since it controls the magnitude of "rare but conceivable" events.

[50] As a first comparison between the fitted GEV distributions and the GLM-generated annual maxima, quantiles of the GEV distributions have been compared directly with those of the simulated maxima. To obtain reasonable estimates of the latter, we pool all of the 39-year simulations to give a combined sample of 3900 simulated maxima at each of the 10 sites. The results are summarized in Table 1 for all of the individual sites in the simulation as well as the time series of areal averages and an average of a group of three sites. For each time series, Table 1 shows estimates of the median, 90th and 99th percentiles of the distribution of annual maxima. Almost everywhere there is a striking agreement between the two sets of estimates; the main exception is at site B33, where the GLM simulations substantially overestimate relative to the GEV fit. However, standard diagnostics at this site (not shown) indicate that the GEV fits poorly; indeed, the largest of the 39 maxima here is 81.3 mm, which is far in excess of the 99th percentile of the fitted distribution. Moreover, the GEV fit at this site is very different from that at B15, which is nearby (see Figure 3) and at which the GEV and GLM agree much more closely. It appears therefore that the discrepancy at this site may be due to the GEV rather than to the simulations. The problem could be overcome by pooling data from several sites before fitting the GEV as in the Flood Estimation Handbook [*Institute of Hydrology*, 1999], for example. However, for current purposes it suffices to note that where the GEV fits reasonably, the two methods of estimating rare events are in close agreement.

Table 1. Estimated Percentiles of Distributions of Annual Maxima, Obtained Using Fitted GEV Distributions and From GLM Simulations[a]

Site/Group	P50, mm		P90, mm		P99, mm		ξ	
	GEV	GLM	GEV	GLM	GEV	GLM	Observed	Simulated
B05	34.8	35.3	50.2	52.7	83.9	80.1	0.263	0.094
B06	29.6	33.5	46.6	48.8	96.3	71.2	0.398	0.056
B13	32.0	30.9	47.8	45.1	66.2	66.9	−0.032	0.057
B15	37.9	42.4	59.1	61.8	96.8	89.8	0.167	0.053
B16	35.7	38.1	52.6	56.6	69.6	82.1	−0.104	0.066
B18	33.2	34.6	50.1	50.0	77.4	71.7	0.120	0.050
B19	33.9	36.9	50.3	53.3	76.5	76.8	0.118	0.048
B20	32.3	36.1	47.6	52.9	70.4	80.5	0.086	0.072
B27	33.6	31.9	51.4	46.6	73.8	68.3	0.004	0.072
B33	34.3	41.9	48.6	61.0	68.3	89.4	0.045	0.048
Average of B18, B19, and B20	33.3	31.2	49.4	44.8	76.6	64.0	0.144	0.052
Average of all 10 sites	31.5	29.7	44.3	42.4	64.9	63.2	0.120	0.053

[a]Also shown are the estimated shape parameters (ξ) for GEV distributions fitted to observed and simulated annual maxima. Site locations are shown in Figure 3.

[51] A second comparison is between the shape parameters (ξ) of GEV distributions fitted to observed and simulated annual maxima. As described above, ξ provides a convenient summary measure of tail behavior for risk assessment purposes; hence this analysis is a check on the upper tail of the simulated distributions. Again, Table 1 shows the results. The estimated standard errors for the observational estimates are all around 0.1 (except at site B06, which has an estimated standard error of 0.19); those for the simulations are around 0.01. The difference reflects the fact that there are 100 simulated sequences but only one set of observations. The majority of the observed estimates are positive, although few of them differ significantly from zero according to their standard errors. The simulation-based estimates are all consistent with the observational ones. More interesting, perhaps, is the fact that they all significantly exceed zero at the 5% level. In the literature, positive values of the shape parameter are widely reported for daily rainfall data [e.g., *Katz et al.*, 2002]. The ability of the GLMs to reproduce this may appear counterintuitive, since they are based on gamma distributions and it is known that, for a sequence of independent gamma random variables, the limiting distribution of the maximum has $\xi = 0$ [*Embrechts et al.*, 1997, Table 3.4.4]. We have checked (via simulation) that this limiting distribution is effectively achieved by the maximum of 365 independent observations. Our results therefore show that the marginal distributions of the GLMs have heavier tails than the gamma distributions used to construct them.

5. Summary

[52] GLMs provide a powerful and flexible environment within which to explore relationships among variables in the climate system. The main contribution of the current paper is to demonstrate their potential for simulating realistic sequences of daily rainfall at a network of sites. The models are cheap to simulate and have a simple structure. The biggest difficulty is the representation of spatial dependence in rainfall occurrence, particularly at spatial scales that are small relative to weather systems. In such situations the proposed beta-binomial scheme is conceptually simple and computationally tractable, and it works well in all the examples we have tried. At larger scales, however, intersite dependence will tend to be lower and to vary with distance: in such cases, other representations of dependence may be more appropriate.

[53] At present, a potential problem with our models is that, conditional on the covariates, rainfall amounts and occurrence are generated independently. It is therefore not guaranteed, for example, that smaller rainfall amounts will be generated close to dry sites. The extent to which this is a problem in applications is not clear. The ultimate test is to use GLM simulations to drive, for example, a rainfall-runoff model, and to examine the resulting performance. Work in this area is currently under way; in the meantime, it has been demonstrated above that the models are able to reproduce a variety of hydrologically important properties of rainfall sequences, at different spatial scales. In particular, extremes were well reproduced, although there is some suggestion that the seasonal variation in extremes could be improved by relaxing the assumption of a constant shape parameter in the rainfall amounts distributions. Further refinement could result from the use of distance-dependent correlation structures for rainfall amounts, and by allowing for seasonal variation in the spatial dependence structures.

[54] The results reported in section 4 are typical of those obtained from several other data sets, from the UK and Ireland as well as other parts of the world. For example, *Yang* [2001] fitted GLMs separately to daily rainfall data from seven different regions representing different climate regimes in mainland China. Although he did not perform any simulations, the fitted models were remarkably similar to those reported above, albeit with lower shape parameters in the amounts models (interestingly, these shape parameters were almost identical for six of the seven regions, the exception being the Tibetan plateau). Moreover, in their original development of GLMs for daily rainfall, *Coe and Stern* [1982] worked with data from several African countries as well as Sri Lanka. This suggests that the methodology is potentially widely applicable. It would, however, be useful to verify this by carrying out further simulation exercises with contrasting data sets.

[55] As well as providing a means of generating "stand-alone" rainfall simulations, this work has potential application to the downscaling of future climate scenarios

generated by atmospheric general circulation models (GCMs) or regional climate models (RCMs). Despite recent increases in the resolution of RCMs, there are questions regarding the representation of rainfall in these models [*Wheater*, 2002], and there remains a need for statistical methods linking gridded climate model outputs to point rainfall sequences. Downscaling is commonly achieved using "weather generators" in which the parameters of a simple model for a rainfall sequence are linked to the output of a climate model [*Wilks and Wilby*, 1999]. Commonly, weather generators for rainfall are based on Markov chain models for rainfall occurrence, along with skewed distributions for rainfall amounts. As discussed in section 2, these models are all closely related to, and in many instances are special cases of, the GLMs presented here; the GLM framework allows a more realistic representation of the relationship between rainfall and large-scale climate. It is therefore of interest to see if GLM simulations, driven by the output of a climate model rather than by observed large-scale indices, are able to improve upon the performance of existing weather generators, at least as far as precipitation is concerned. Research in this area is ongoing.

[56] **Acknowledgments.** This work was funded by the Department of the Environment, Food and Rural Affairs (R & D project FD2105). The NAO index data were obtained from the Climatic Research Unit, University of East Anglia (http://www.cru.uea.ac.uk/cru/data/nao.htm). The authors are grateful to three anonymous reviewers and an Associate Editor for many helpful suggestions that have strengthened this work.

References

Bárdossy, A., and E. Plate (1992), Space-time model for daily rainfall using atmospheric circulation patterns, *Water Resour. Res.*, *28*, 1247–1259.

Barnston, A. G., and R. E. Livezey (1987), Classification, seasonality and persistence of low-frequency atmospheric circulation patterns, *Mon. Weather Rev.*, *115*, 1083–1126.

Buishand, T., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resour. Res.*, *37*, 2761–2776.

Chandler, R. (2002), GLIMCLIM: Generalized linear modelling for daily climate time series (software and user guide), *Tech. Rep. 227*, Dep. of Stat. Sci., Univ. College London, London.

Chandler, R. (2005), On the use of generalized linear models for interpreting climate variability, *Environmetrics*, *16*(7), 699–715.

Chandler, R., and H. S. Wheater (1998), Climate change detection using generalized linear models for rainfall—A case study from the west of Ireland. II. Modelling of rainfall amounts on wet days, *Tech. Rep. 195*, Dep. of Stat. Sci., Univ. College London, London.

Chandler, R., and H. S. Wheater (2002), Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland, *Water Resour. Res.*, *38*(10), 1192, doi:10.1029/2001WR000906.

Charles, S., B. Bates, and J. Hughes (1999), A spatiotemporal model for downscaling precipitation occurrence and amounts, *J. Geophys. Res.*, *104*(D24), 31,657–31,669.

Coe, R., and R. D. Stern (1982), Fitting models to daily rainfall, *J. Appl. Meteorol.*, *21*, 1024–1031.

Coles, S. (2001), *An Introduction to the Statistical Modelling of Extreme Values*, Springer, New York.

Cox, D., and N. Wermuth (1996), *Multivariate Dependencies: Models, Analysis and Interpretation*, CRC Press, Boca Raton, Fla.

Cressie, N. (1991), *Statistics for Spatial Data*, John Wiley, Hoboken, N. J.

Dobson, A. (2001), *An Introduction to Generalized Linear Models*, 2nd ed., CRC Press, Boca Raton, Fla.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997), *Modelling Extremal Events for Insurance and Finance*, Springer, New York.

Emrich, L., and M. Piedmonte (1991), A method for generating high-dimensional multivariate binary variates, *Am. Stat.*, *45*, 302–304.

Hughes, J. P., P. Guttorp, and S. Charles (1999), A nonhomogeneous hidden Markov model for precipitation, *Appl. Stat.*, *48*, 15–30.

Hurrell, J. W. (1995), Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation, *Science*, *269*, 676–679.

Institute of Hydrology (1999), *Flood Estimation Handbook*, 5 vols., Wallingford, U. K.

Jones, P. D., T. Jónsson, and D. Wheeler (1997), Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and south-west Iceland, *Int. J. Climatol.*, *17*, 1433–1450.

Katz, R., M. Parlange, and P. Naveau (2002), Statistics of extremes in hydrology, *Adv. Water Resour.*, *25*, 1287–1304.

Krzanowski, W. (1988), *Principles of Multivariate Analysis*, Oxford Univ. Press., New York.

Lunn, A., and S. Davies (1998), A note on generating correlated binary variables, *Biometrika*, *85*, 487–490.

McCullagh, P., and J. Nelder (1989), *Generalized Linear Models*, 2nd ed., CRC Press, Boca Raton, Fla.

Murphy, S. J., and R. Washington (2001), United Kingdom and Ireland precipitation variability and the North Atlantic sea-level pressure field, *Int. J. Climatol.*, *21*, 939–959.

Oman, S., and D. Zucker (2001), Modelling and generating correlated binary variables, *Biometrika*, *88*, 287–290.

Ryan, L. (1995), Comment on the article by Liang and Zeger, *Stat. Sci.*, *10*, 189–193.

Stehlík, J., and A. Bárdossy (2002), Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation, *J. Hydrol.*, *256*, 120–141.

Stern, R. D., and R. Coe (1984), A model fitting analysis of rainfall data (with discussion), *J. R. Stat. Soc., Ser. A*, *147*, 1–34.

Terrell, G. (2003), The Wilson-Hilferty transformation is locally saddle-point, *Biometrika*, *90*, 445–453.

Wheater, H. (2002), Progress in and prospects for fluvial flood modelling, *Proc. R. Soc. Lond., Ser. A*, *360*, 1409–1432.

Wheater, H. S., V. S. Isham, C. Onof, R. E. Chandler, P. J. Northrop, P. Guiblin, S. M. Bate, D. R. Cox, and D. Koutsoyiannis (2000), Generation of spatially consistent rainfall data, *Tech. Rep. 204*, Dep. of Stat. Sci., Univ. Coll. London, London.

Wilks, D. (1998), Multisite generalization of a daily stochastic precipitation generation model, *J. Hydrol.*, *210*, 178–191.

Wilks, D., and R. Wilby (1999), The weather generation game: A review of stochastic weather models, *Prog. Phys. Geogr.*, *23*, 329–357.

Williams, D. (1982), Extra-binomial variation in logistic linear models, *Appl. Stat.*, *31*, 144–148.

Yan, Z., S. Bate, R. Chandler, V. Isham, and H. Wheater (2002), An analysis of daily maximum windspeed in northwestern Europe using generalized linear models, *J. Clim.*, *15*, 2073–2088.

Yan, Z., S. Bate, R. Chandler, V. Isham, and H. Wheater (2005), Changes in extreme wind speeds in NW Europe simulated by generalized linear models, *Theor. Appl. Climatol.*, doi:10.1007/s00704-005-0156-x, in press.

Yang, C. (2001), Observed changes and simulative predictions of climate extremes in China (in Chinese), Ph.D. thesis, Inst. of Atmos. Phys., Beijing.

Yang, C., R. Chandler, V. Isham, C. Annoni, and H. Wheater (2005), Simulation and downscaling models for potential evaporation, *J. Hydrol.*, *302*, 239–254.

R. E. Chandler, V. S. Isham, and C. Yang, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK. (richard@stats.ucl.ac.uk; valerie@stats.ucl.ac.uk; chi@stats.ucl.ac.uk)

H. S. Wheater, Department of Civil and Environmental Engineering, Imperial College of Science, Technology and Medicine, London SW7 2AZ, UK. (h.wheater@imperial.ac.uk)