Bayesian Image Analysis and the Disaggregation of Rainfall

R. E. CHANDLER

Department of Statistical Science, University College London, London, United Kingdom

N. G. MACKAY*, H. S. WHEATER, AND C. ONOF

Department of Civil Engineering, Imperial College, London, United Kingdom

(Manuscript received 5 April 1999, in final form 20 July 1999)

ABSTRACT

The applicability of meteorological general circulation models (GCMs) is limited by their spatial resolution. In this paper, a method is developed for improving the resolution of GCM-generated rainfall fields, using ideas from Bayesian image analysis to improve the resolution of the binary wet–dry image. This approach incorporates both the spatial and temporal memory of the rainfall field and can be adapted to utilize any available physical information. The method is illustrated using data from a network of weather radar stations in Arkansas, and some informal diagnostic procedures are developed for assessing the adequacy of the underlying model.

1. Introduction

A typical atmospheric general circulation model (GCM) outputs the values of many different variables, averaged over grid squares whose sides may be of the order of 100–200 km in length. The output from such models—rainfall, in particular—is of potential interest to the hydrologist. Rainfall, however, is notoriously variable in both space and time, and the use of a single averaged value over a large grid square is physically unrealistic. It is common, therefore, to apply a simple disaggregation scheme to introduce some finescale spatial variability into the rainfall. The resulting distribution of rainfall is used to update the surface hydrology. Typically, such techniques are used in short-term (up to 36 h) forecasting applications.

Existing methods

Current methodology for rainfall disaggregation in practice, as used by the U.K. Meteorological Office (Warrilow et al. 1986) and others, suffers from numerous disadvantages. In particular, there is no spatial location associated with rainfall occurrences, and spatial and temporal dependence are not accounted for. The procedure is described in Onof and Wheater (1996a,b), who address some of these deficiencies and introduce temporal dependence by modeling sequences of *coverages*—that is, proportions of grid squares that are wet—and total rainfall amounts as bivariate time series. A number of other authors have suggested improved schemes. Perhaps the approach that is most similar in spirit to our own is that of Perica and Foufoula-Georgiou (1996), who used a statistical model for finescale rainfall patterns conditioned on physical characteristics of a storm environment. This model relies on spatial scaling properties and assumes Gaussianity of standardized rainfalls: this latter assumption may not always be realistic. Moreover, the scheme does not incorporate temporal dependence.

Other recent work includes that of Lebel et al. (1998), who propose an event-based scheme using the turning-bands method to reproduce finescale spatial structure and using historical storm profiles to determine temporal structure. However, in its current form, Lebel et al.'s (1998) model is not suitable for interactive coupling with a GCM (Lebel et al. 1998, section 6). Bárdossy (1997) proposes a stochastic model conditioned upon atmospheric circulation patterns that is suitable for disaggregating to point rainfall sequences at a network of sites.

This paper introduces a new and general methodology for determining the spatial location of wet and dry regions within a GCM grid square, which is fully spatial–temporal and enables physical information to be incorporated. The emphasis is upon developing the methodology; some further results may be found in Onof et

* Current affiliation: D Cubed, Cambridge, United Kingdom.

Corresponding author address: Dr. R. E. Chandler, Department of Statistical Science, University College London, Gower Street London, WC1E 6BT, United Kingdom.
E-mail: richard@stats.ucl.ac.uk

al. (1998). Section 2 offers an overview of the relevant theory, presents the proposed algorithm, and discusses parameter estimation. Sections 3 and 4 offer an example of the performance of the method and present some diagnostic procedures for assessing the method's adequacy, and section 5 sets out some open problems and areas for future investigation.

2. Theoretical background

The proposed algorithm is based on Bayesian methods for the reconstruction of ''clean'' images from ''noisy'' data, as described by Besag (1986), Jennison and Jubb (1988), Jubb and Jennison (1991), for example. Most of these methods are iterative. Jennison (1986) suggested that in some cases a suitable starting value for the iteration could be obtained by aggregating the data up to a coarser spatial scale and thereby smoothing out the noise. The image is then reconstructed at successively finer spatial scales until the required resolution is obtained. This is called the ''cascade algorithm.''

This reconstruction at successively finer spatial scales offers a clear parallel with disaggregation. Moreover, in image reconstruction, the observed data are interpreted as ''evidence'' from which the ''true scene'' may be inferred. For disaggregation, evidence that is useful for reconstructing a finescale rainfall field at some time t may include the appearance of the field at time $t - 1$, the GCM-generated field of coarse-scale rainfall intensities, and any other GCM outputs that contain information regarding likely finescale structure [e.g., indices of convective activity, such as those used by Perica and Foufoula-Georgiou (1996)].

We now establish the relevant notation and define the model that will be used in the disaggregation algorithm.

a. Notation

We imagine each GCM square to be divided into smaller squares (which we call ''pixels'') at the scale of interest. Our aim is to determine which pixels are wet and which are dry at time t given some (fixed) evidence, denoted by $\mathcal{E}$. The field is to be reconstructed only within grid squares for which the rainfall is nonzero: m_t is the number of these squares, and n_t is the number of pixels contained within the squares. The field to be reconstructed is regarded as a vector of binary random variables (1 = wet; 0 = dry) $\mathbf{X}(t) = \{X_1(t), \ldots, X_{n_t}(t)\}$, with realizations $\{\mathbf{x}(t) = \{x_1(t), \ldots, x_{n_t}(t)\}$.

For our purposes, the evidence $\mathcal{E}$ will consist of the previous finescale wet–dry field, $\mathbf{x}(t - 1)$, and the vector of coarse-scale rainfall intensities generated by the GCM at time t interpreted as mean rainfalls over each grid square. This vector is denoted $\mathbf{r}(t) = \{r_1(t), \ldots, r_{m_t}(t)\}$ and is regarded as a realization of the corresponding random vector $\mathbf{R}(t)$.

Here, $\mathbf{P}(t) = \{P_1(t), \ldots, P_{m_t}(t)\}$ denotes the (unobserved and random) coverage vector whose ith entry is the proportion of wet pixels within the ith wet grid square, with realizations $\{\mathbf{p}(t)\}$.

We write $\Pr\{\mathbf{x}(t)\}$ instead of $\Pr\{\mathbf{X}(t) = \mathbf{x}(t)\}$, with a similar abbreviation for other random vectors. The notation $f_{\cdot}(\cdot)$ denotes a probability density, with the subscript indicating the random variable being considered. Conditional probabilities or densities are denoted by a vertical bar. For example, $\Pr\{\mathbf{x}(t) \mid \mathbf{x}(t - 1)\}$ means the probability that the wet–dry field at time t is $\mathbf{x}(t)$ if the wet–dry field at time $t - 1$ was $\mathbf{x}(t - 1)$.

b. The Bayesian framework

Given the evidence $\mathcal{E}$, which contains observations on continuous random variables, for the conditional distribution of $\mathbf{X}(t)$, by Bayes's Theorem, we have

$$\Pr\{\mathbf{x}(t) \mid \mathcal{E}\} = \frac{\Pr\{\mathbf{x}(t)\} f_{\mathcal{E}|\mathbf{x}(t)}(\mathcal{E} \mid \mathbf{x}(t))}{f_{\mathcal{E}}(\mathcal{E})}. \quad (1)$$

This is referred to as the ''posterior distribution.'' Because $\mathcal{E}$ is fixed, the denominator is constant for any configuration $\mathbf{x}(t)$. Hence, we consider only the numerator, and specify the model up to a constant of proportionality.

In image reconstruction, often the aim is to find a configuration that is in some sense optimal according to the posterior distribution (1). However, for disaggregation, it is not clear that an optimal reconstruction is useful at all, in view of the huge uncertainty regarding the finescale image. Moreover, rainfall patterns do not tend to exhibit the smooth boundaries that are found typically in more conventional images in a reconstruction problem. [See chapter 5 of Ripley (1988) for typical examples of such images.] We, therefore, propose to generate a sample of reconstructions from the posterior, using the Gibbs Sampler (Geman and Geman 1984; Ripley 1988). Then, this sample may be used in hydrological applications to quantify uncertainty regarding the response of surface processes. The sampler is initialized with some plausible configuration for the image. Then, each pixel, in turn, is visited and resampled according to its conditional distribution, given the current state of all of the other pixels in the image. A complete scan of the entire image in this way constitutes a single iteration of the sampler. Repeated iteration of the sampler ensures that the distribution of the image will converge to the required posterior distribution.

c. Model specification

We now specify the various components of the numerator in (1) that have been used as a first-guess model for the proposed disaggregation scheme. Assessment of the adequacy of the first-guess model is deferred to section 4. Two assumptions will be made. The assumptions are felt to be reasonable on physical grounds, at least

as a first-order approximation, and are made so that the model is mathematically tractable and so that every part of the model can be supported by data analysis. The assumptions are as follows.

Assumption 1: $\Pr\{\mathbf{x}(t-1)\,|\,\mathbf{r}(t), \mathbf{x}(t)\} = \Pr\{\mathbf{x}(t-1)\,|\,\mathbf{x}(t)\}$; that is, the mean rainfall over a GCM grid square at time t tells us nothing about the finescale wet–dry pattern at time $t-1$, once we know the finescale wet–dry pattern at time t.

Assumption 2: $\Pr\{\mathbf{x}(t)\,|\,\mathbf{p}_x(t), \mathbf{r}(t)\} = \Pr\{\mathbf{x}(t)\,|\,\mathbf{p}_x(t)\}$, where $\mathbf{p}_x(t)$ is the coarse-scale field of coverages corresponding to the finescale wet–dry field $\mathbf{x}(t)$. That is, the mean rainfall over a GCM square tells us nothing about the finescale wet–dry pattern at the same time, if we already know how much of the square is wet.

Under assumptions 1 and 2 it can be shown (Chandler et al. 1997) that the posterior distribution (1) can be written as

$$\Pr\{\mathbf{x}(t)\,|\,\mathcal{E}\} \propto \frac{\Pr\{\mathbf{x}(t)\}\Pr\{\mathbf{x}(t-1)\,|\,\mathbf{x}(t)\}\Pr\{\mathbf{p}_x(t)\,|\,\mathbf{r}(t)\}}{\Pr\{\mathbf{p}_x(t)\}}. \tag{2}$$

We need to specify some prior structure for $\Pr\{\mathbf{x}(t)\}$, a temporal dependence mechanism for $\Pr\{\mathbf{x}(t-1)\,|\,\mathbf{x}(t)\}$, and a plausible structure for $\Pr\{\mathbf{p}_x(t)\,|\,\mathbf{r}(t)\}$. In order to use the Gibbs Sampler, the model needs to be specified in such a way that we can calculate the probability of any pixel being wet at time t given the evidence and given the state of all of the other pixels at time t.

For the prior structure, we use a Markov Random Field (MRF) model (Besag 1974; Ripley 1988), which provides a convenient framework within which to specify local interactions. We have followed common image reconstruction practice and adopted an automodel (Besag 1974). The type-1 neighbors of a pixel are those pixels immediately horizontally and vertically adjacent, and its type-2 neighbors are those immediately diagonally adjacent. The model may be specified as

$$\frac{\Pr\{X_i = 1\,|\,\text{rest of scene}\}}{\Pr\{X_i = 0\,|\,\text{rest of scene}\}} = \exp\left[\alpha + 4\sum_{k=1}^{2}(\beta_k \times \text{proportion of wet type } k \text{ neighbors})\right], \tag{3}$$

where the use of the proportion rather than the number of wet neighbors circumvents problems at the edge of an image. Model (3) forces a symmetric dependence structure into the reconstructions. The adequacy of this will be investigated in section 4.

A simple model for $\Pr\{\mathbf{x}(t-1)\,|\,\mathbf{x}(t)\}$ assumes that the state of the ith pixel at time t depends on the previous image only through its own state at that time. In this case, we have

$$\Pr\{\mathbf{x}(t-1)\,|\,\mathbf{x}(t)\} = \prod_{i=1}^{n}\Pr\{X_i(t-1) = x_i(t-1)\,|\,X_i(t) = x_i(t)\}. \tag{4}$$

Strictly speaking, (3) and (4) may be compatible only when disaggregating at a single time instant, since the combination of the two equations implies that the overall distribution (i.e., prior) of $\mathbf{x}(t)$ will be different from that of $\mathbf{x}(t-1)$. If the model is used successively to disaggregate each image in a sequence, using the same prior at each time step, subtle theoretical questions arise regarding the consistency of the model's specification. The problem is that at time t we pretend effectively that the probability structure for time $t-1$ differs slightly from the structure that was actually used. For practical purposes, however, this is largely irrelevant, since the probability structure has no effect at all on our actual reconstruction at time $t-1$. Moreover, the prior (3) is intended only as a plausible approximation to the overall joint distribution. It is felt that the adequacy of the scheme should be assessed more in terms of its robustness in application than in terms of these theoretical subtleties.

For $\Pr\{\mathbf{p}_x(t)\,|\,\mathbf{r}(t)\}$, we will use an appropriate regression relationship between a grid square's coverage and its mean rainfall. If the residuals from this relationship are independent, then for the Gibbs Sampler we need to consider only the conditional probabilities of coverage in the *current* grid square instead of the full joint distribution $\Pr\{\mathbf{p}_x(t)\,|\,\mathbf{r}(t)\}$. The definitions of $\mathbf{r}(t)$ and $\mathbf{p}_x(t)$ suggest a linear relationship between their components, either on a log scale, which has been verified empirically by Onof and Wheater (1996b), or directly [Short et al. (1993) and Kedem and Pavlopoulos (1991), for example]. Kedem and Pavlopoulos (1991) also give a theoretical justification for such a relationship. However, both the log–log and linear–linear relationships are unsatisfactory for predicting $\mathbf{p}$ from $\mathbf{r}$, because predictions are not constrained in the range $(0, 1)$. We therefore seek some other form of dependence, empirically. Based on extensive analyses of rain radar data from both the United Kingdom and the United States, the most promising form of relationship found is between the empirical logit (Cox and Snell 1989) of the coverage and the log-mean rainfall. Residuals appear to have an approximately symmetrical distribution, which we approximate by a normal distribution with appropriate continuity corrections. Examples are shown in section 3.

The probability in the denominator of (2), in principle, can be obtained numerically from the MRF prior by exhaustive enumeration of all possible scenarios and by computation of the prior probability for each scenario. In practice, however, this is computationally expensive. The problem is reduced slightly when using the Gibbs Sampler, when we need only to consider local scenarios. However, a complete enumeration may still be infeasible.

The approach adopted here has been to estimate the probability distribution of coverage within a single grid square by simulation. In the work to be reported, in which each grid square contains 25 pixels, this is likely to provide a reasonable working approximation. However, the need to resort to simulation is slightly unsatisfactory. We return to this point in section 4.

The proposed model has eight parameters in total: three parameters for the MRF prior; two transition probabilities,

$$p_{11} = \Pr\{X_i(t-1) = 1 \,|\, X_i(t) = 1\} \quad \text{and}$$
$$p_{10} = \Pr\{X_i(t-1) = 1 \,|\, X_i(t) = 0\}; \qquad (5)$$

and three parameters (slope, intercept, and residual standard deviation) for the coverage-intensity regression relationship. All of the parameters can be estimated straightforwardly using data from a weather radar station. The parameters α, β_1, and β_2 in the MRF part of model (3) have been estimated using the coding methods of Besag (1972, 1974). Estimation of p_{11} and p_{10} is trivial, and the regression estimates are routine.

3. Example

We now present an example to illustrate the theory of the preceding sections. Data from a network of 16 weather radar stations covering the basin of the Arkansas Red River in the United States (an area of 1350 $\times$ 630 km^2) are used to estimate parameters required by the scheme and to provide aggregated images that will be used as dummy GCM images for testing the scheme. Use of radar data (which normally would not be available in forecasting applications, for example) gives us a known finescale image against which to assess performance.

In general, the quality of weather radar data is problematic (e.g., Collier 1989), but these data have been subject to an extensive amount of calibration, both in reconciling measurements from different radars and in comparing radar measurements with readings from rain gauges. (For details, see the Arkansas Red Basin River Forecast Center's Web site at http://www.abrfc.noaa.gov/pcpn_methods.html.) The spatial resolution of the data is 4 $\times$ 4 km^2, and the temporal resolution is hourly. Each radar image can be regarded as the mean rainfall during an hour. For this exercise, the dimension of the pixels is 8 $\times$ 8 km^2, and the dimension of the GCM squares is 40 $\times$ 40 km^2. The corresponding images are obtained from the original data by aggregating up to the required scale.

a. Parameters

The posterior density (2) describes the wet–dry rainfall field at the scale to which we are disaggregating. Therefore, it is appropriate to estimate parameters from data at this scale. Ideally, we would like to determine all of the parameters from the coarse-scale image. To this end, we have attempted—via analyses of estimated parameters for the Arkansas data—to link the estimated parameters obtained for 8 $\times$ 8 km^2 data to various properties of the 40 $\times$ 40 km^2 data. However, no clear relationships have been found. In this case, a reasonable strategy is to use a fixed set of parameter estimates for each month of the year. (The adequacy of this approach will be considered in section 4.)

Monthly parameter sets were obtained using data from July 1994 to June 1995. Figure 1 shows typical examples of the relationship between logit coverage and log-mean rainfall. Each hourly image contributes a separate data point to these plots for each wet grid square: there are 24 836 points on the December plot, and 50 266 on the June plot. The scatterplots at the top of the figure should therefore be interpreted with caution, since it is impossible to show where the data are clustered. Interpretation is aided by the Q–Q plots at the bottom of the figure, which show that regression residuals are approximately normally distributed in both December and June.

Estimation of the MRF parameters α, β_1, and β_2 is performed using data at pixel resolution (discarding any pixels within completely dry coarse-scale grid squares). So each month, there are about 25 times as many data points to consider as for the regression relationships. To speed up the estimation, separate estimates were obtained for each image in the record. For each month, a weighted average of the individual estimates was then computed. Each image was weighted according to the mean rainfall over wet pixels, to avoid the estimates being dominated by contributions from insignificant rainfall events. (An alternative would have been to weight by coverage.)

The resulting parameter set showed distinct seasonal variation, although with much noise, which is to be expected from only one year of data. The monthly estimates were therefore smoothed using a 3-point moving average window with weights (0.25, 0.5, 0.25). The smoothed parameter estimates for each month are given in Table 1. The seasonal variation is much as expected. The MRF parameters indicate weaker neighborhood interactions in the summer than in the winter, and the strong seasonality of the intercept in the regression relationship between log rainfall and the empirical logit of coverage indicates that rainfall is much more intense in summer than in winter, a consequence of the highly convective nature of summer rainfall. It is perhaps worth commenting on the seasonality that is clearly present in α. Perhaps the most meaningful interpretation of this parameter is that it is the log odds for rain within a pixel that is surrounded completely by dry pixels [see (3)]. The higher values obtained in the summer months reflect the spottiness of summer rainfall.

b. Gibbs Sampler implementation

With the parameter set given in Table 1, the proposed scheme was applied to a 15-h event from July 1995.

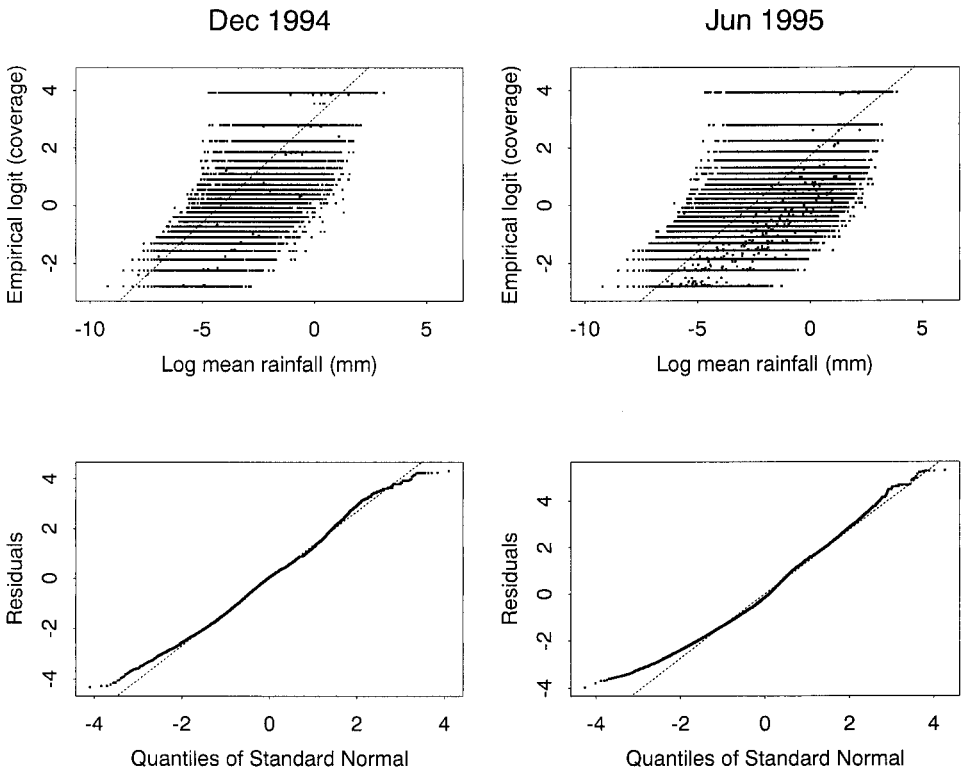

FIG. 1. Relationship between empirical logit of coverage and log-mean rainfall for Arkansas radar data, December 1994 and June 1995. (top) Scatterplots showing fitted regression lines. (bottom) The Q–Q plots of residuals from regressions. The dashed lines show the expected values for normally distributed residuals.

This event was chosen, first, because it was immediately preceded by an image in which the entire region was dry—whence the true scene is known—and, second, because the time period is outside that used to estimate parameters, so that we could obtain an out-of-sample assessment of the method's performance.

Two issues relating to the implementation of the Gibbs Sampler need to be addressed here. The first is the choice of an appropriate starting configuration from which to iterate. The second is the number of iterations required. Plausible starting configurations are

- a field with all pixels wet in a GCM square whose rainfall is nonzero,
- a copy of the finescale field at the previous time step, and
- a field with pixels set to wet independently, each with probability determined by the regression relationship on grid square mean rainfall.

To determine which of these should be used, a single disaggregation of the entire 15 h was carried out for each of the different initialization strategies. For each

TABLE 1. Parameters used in the operational version of the disaggregation scheme for the Arkansas data.

	Markov random field parameters			Transition probabilities		Regression parameters		
Month	α	β_1	β_2	p_{11}	p_{10}	Intercept	Slope	RSD
Jan	−5.92	2.55	0.48	0.66	0.25	3.14	0.74	1.37
Feb	−5.79	2.57	0.44	0.60	0.23	2.98	0.73	1.38
Mar	−5.66	2.52	0.44	0.62	0.24	2.87	0.73	1.37
Apr	−5.50	2.46	0.41	0.63	0.25	2.63	0.73	1.33
May	−5.22	2.36	0.36	0.60	0.25	2.19	0.71	1.33
Jun	−4.96	2.25	0.35	0.56	0.22	1.80	0.67	1.37
Jul	−4.78	2.21	0.31	0.53	0.20	1.57	0.62	1.42
Aug	−4.71	2.19	0.29	0.51	0.19	1.68	0.61	1.47
Sep	−4.83	2.17	0.36	0.54	0.20	2.04	0.61	1.48
Oct	−5.17	2.22	0.46	0.60	0.24	2.36	0.65	1.40
Nov	−5.56	2.30	0.56	0.66	0.28	2.68	0.70	1.32
Dec	−5.86	2.43	0.56	0.69	0.28	3.05	0.73	1.33

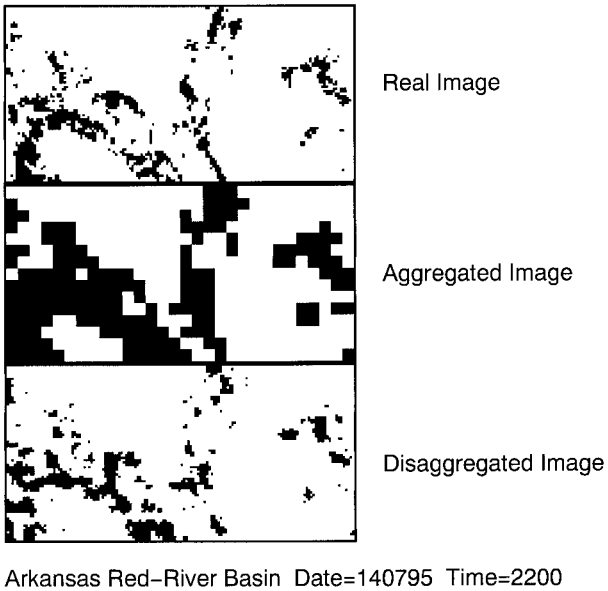

FIG. 2. Specimen reconstruction of an aggregated image, showing the original wet–dry configuration.

image, 100 000 iterations of the Gibbs Sampler were used in the disaggregation, and the sequence of log posterior probabilities, according to model (2), was plotted. After an initial ''burn-in,'' all of the plots settled down into a stable behavior, and this is when convergence was judged to have occurred.

The results indicated that the all-wet starting strategy produces much slower convergence than the other two strategies, which both reach similar equilibrium levels within about 200 iterations. For two of the images during the storm, the equilibrium level achieved by these two strategies appeared slightly different, however. This may be due to GCM squares that become wet between time steps, when the previous image is used to initialize the sampler. If this occurs, during the sampler's first iteration, the first pixel visited in such a square will always be set to wet, because the current coverage will be zero, which is impossible under the regression relationship between coverage and mean rainfall. This introduces a small element of determinism into the sampling, which may bias the results. For this reason, we prefer the random starting strategy. In the work reported herein, 400 iterations of the sampler have been used with the random starting strategy.

The random number generator used in this work is the subtract-with-borrow generator of Marsaglia and Zaman (1991), which is appropriate for extensive simulation studies.

c. Results

To give an idea of the appearance of reconstructions using this method, a typical scenario is shown in Fig. 2. Although this is only one possible reconstruction, in our experience all of the reconstructions of a particular image are broadly similar. For illustrative purposes, this reconstruction exhibits most of the main features. The top frame shows the original wet–dry pattern obtained from the radar data. The bottom is the result of applying the disaggregation algorithm to the coarse-scale field shown in the center frame. The figure illustrates the difficulty of the problem—it is perhaps optimistic to expect that we can obtain anything resembling the top frame from the middle one.

There is some qualitative similarity between the top and bottom frames in Fig. 2. However, there is a tendency toward horizontal and vertical, rather than diagonal, banding in the reconstruction, which also may have smoother outlines to the rain areas than the original image. We will return to this point subsequently.

4. Model verification

The ability to assess the adequacy of a model is an area that has received relatively little attention in the image reconstruction literature. Conventional model checking techniques, as used by statisticians and meteorologists alike, tend to be basically univariate and may not be suitable for checking models for entire images. In this section, we present some diagnostic procedures for checking the various components of the model, and apply them to the test storm described in the previous section.

a. Overall assessment

Perhaps the simplest (although not necessarily the most useful) measure of the success of the method is to examine the proportion of pixels correctly classified as wet or dry, within coarse-scale grid squares that are wet. In Fig. 2, for example, 64.2% of the pixels are classified correctly. However, this proportion is itself a random variable, so we consider its mean (over reconstructions), which can be estimated by simulation. We have performed 100 reconstructions of the event described in the previous section and, for each image, have calculated the mean proportion of correctly classified pixels, together with a 95% confidence interval. For the majority of the storm, the mean lies between 0.65 and 0.7, with little variation between reconstructions. At the start of the storm, the mean is higher, but this is due to the coverage being very low, so that both the original image and the reconstructions are predominantly dry, even within wet grid squares.

The dependence of this performance measure upon coverage is one of its drawbacks, in terms of assessing absolute model performance. Moreoever, from an applied perspective it is likely that the precise location of wet and dry areas is less important than their spatial structure in determining surface hydrological response. Hence, although the mean proportion of correctly classified pixels may be a useful simple indicator of per-

formance—for example, to compare with other possible approaches to the problem—it should be interpreted with some care, and more extensive analyses are needed.

b. Approaches to model checking

One of the reasons for the lack of diagnostic procedures in this type of situation is the complexity of the problem: formal methods, such as hypothesis testing, are unsatisfactory. Some discussion of this point may be found in Besag (1974). In the present context, the main arguments against the use of such techniques are the following.

- The model is necessarily an oversimplification of reality and therefore guaranteed to be wrong. The huge amounts of data involved ensure that virtually any hypothesis testing procedure of a null against an extended model will be guaranteed to find significant evidence in favor of the latter.
- There is no obvious hierarchical structure within which the model can be embedded. The result is that it is extremely difficult to specify a nested sequence of models in a meaningful way, as would be required by a formal approach.

For these reasons, the procedures to be adopted here are informal. We seek, first, to examine whether or not the model is nearly right, in some sense to be determined by the context, and then, if not, to determine which model components are underperforming and how their performance may be improved. For a more extensive overview, and reference list, see Chandler (1998).

c. Analyses for model checking

To determine whether or not the model is nearly right for informal checking, it is necessary to focus upon performance measures that are of practical interest. A quick method of assessment is to compare the method's ability to reproduce features of the data with its ability to reproduce feaures of its own output: the performances should be similar. This is straightforward to implement. First, choose some appropriate measure of (dis) agreement between two images. Then, for each image in the test storm, perform two disaggregations. One is regarded as a benchmark realization of the model (note that the coarse-scale images obtained from both data and model are the same), and the other as the official reconstruction. Compute the chosen (dis)agreement between the reconstruction and the benchmark, and between the reconstruction and the data. The difference between the two (dis)agreements should be close to zero.

This procedure has been applied to the test storm using the proportion of misclassified pixels as a performance measure. As before, 100 simulations have been carried out to obtain a mean and confidence interval for the difference in misclassification rates, for each image in the storm (for each simulation, a new benchmark realization is computed as well as a new reconstruction). The results are shown in Fig. 3. The mean difference is always significantly less than zero, indicating that model realizations resemble themselves more than they do data.

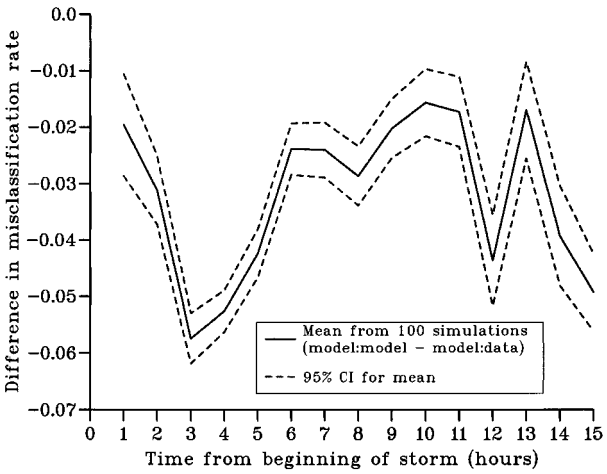

FIG. 3. Difference in misclassification rates, when two model realizations are compared and when the model and data are compared, for each image in the test storm.

Despite the drawbacks of the misclassification rate as a performance measure in its own right, therefore, it has proved useful in identifying some feature of the model that is inconsistent with the data. By way of illustrating other potential model checking techniques we now investigate this further, even though its practical significance is probably marginal. For the main part of the storm (hours 5–12), the discrepancy is mostly less than one pixel per coarse-scale grid square.

Before proceeding, we summarize the possible sources of error in the scheme as implemented here. Broadly speaking, as far as the model itself is concerned, there are three sources of error: misspecification of 1) the Markov Random Field component, 2) the temporal dependence component, and 3) the conditional distribution of coverage given mean rainfall. In addition, it is possible that error has been introduced by the use of monthly parameter values from outside the time period of the event under consideration. Finally, there is the approximation to $\mathbf{p}_x(t)$ in (2).

Many of these potential errors are readily checked. For example, the use of monthly parameter values has been investigated by rerunning the analysis using parameters obtained from the test storm itself. There is very little improvement in practical terms, so it appears that realizations are relatively insensitive to errors in the identification of parameters.

To assess the performance of the various components of the model that affect the coverage of the reconstructions—the conditional distribution of coverage given mean rainfall, the approximation to $\mathbf{p}_x(t)$, and the var-

ious assumptions regarding conditional independence that were made in section 2c—we examine the model's ability to reproduce the observed coverage. For each image, the error in the reconstructed coverage can be calculated, and confidence intervals for the mean error can be obtained by simulation.

For the test storm considered here, the mean bias in coverage reproduction (based on 100 reconstructions) is substantially less than one pixel per grid square for most of the storm, although for the first three images the model underestimates coverages by up to 2 pixels per grid square, which perhaps indicates that something is wrong (each grid square contains 25 pixels). This may contribute to the increased discrepancy between model and data for these images (Fig. 3), but is unlikely to be a problem in practice, as this part of the storm is hydrologically insignificant. In any case, the performance over the remainder of the storm is almost certainly adequate, particularly if the storm itself is regarded as a single realization of the underlying process.

d. Checking the Markov Random Field

The final part of the model to check is the Markov Random Field component. There are two aspects of this that may be misspecified: 1) the use of a symmetric automodel [at (3)] and 2) the specification of the neighborhood relation to a distance of only one pixel. If this model structure is correct, then, in some sense, all of the information about the state of a pixel is contained in the counts (C_1 and C_2, say) of wet type-1 and type-2 neighbors. We can determine whether or not this is a realistic assumption for the test storm under consideration by estimating *perfect information misclassification probabilities* (PIMPs—our own terminology) for each image in the storm. A PIMP is the probability that a pixel is misclassified given perfect information, regarding both the configuration of the rest of the scene and its probability distribution given this configuration. If p is the true probability that a pixel is wet, then it will be misclassified with probability $2p(1 - p)$: the probability that it is sampled dry and is actually wet, plus the probability that it is sampled wet and is actually dry. Under the model structure used in this work, there will be a different value of p for each of the 25 different pairs (C_1, C_2). Each value can be estimated from radar data using an appropriate subset of pixels, and the results used to provide an overall estimate of the misclassification probability. It is important to use exactly, rather than asymptotically, unbiased estimators of the individual misclassification probabilities in order to ensure consistency of the overall estimate, as individual subsets of pixels may be small. We therefore use estimators of the form

$$\widehat{\text{PIMP}} = \frac{2n}{n - 1}\hat{p}(1 - \hat{p}), \tag{6}$$

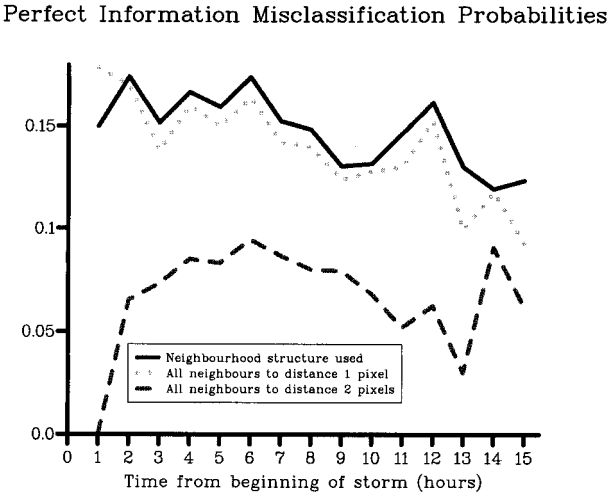

FIG. 4. Estimated PIMPS for the test storm of 14–15 July 1994.

where n is the number of pixels in the subset of interest, and $\hat{p}$ is the proportion of pixels that are wet.

PIMPs can be estimated for any given neighborhood structure and used to compare the amounts of information contained within different structures. We note in passing that the idea of examining the extra information to be gained from including wider neighborhoods is similar to that of partial correlations—but in a spatial setting and modified to deal specifically with sampled binary data. We have computed PIMPs for two neighborhood structures in addition to the one used: 1) a structure in which the neighbors of a pixel are again those pixels immediately horizontally, vertically, and diagonally adjacent but in which there is a different probability for each of the 2^8 possible neighborhood configurations [this will allow an assessment of the adequacy of the assumption of symmetry implicit in the automodel (3)]; and 2) a structure in which the neighborhood is extended to a distance of two pixels in each direction, with a different probability for each of the 2^{24} possible configurations. Clearly, there will, in general, be very few pixels in a subset corresponding to a particular configuration for the latter case, whence the use of unbiased estimators is of critical importance.

The results are shown in Fig. 4, in which separate estimates have been computed for each image of the event described in section 3. It is clear, comparing the solid and dotted lines, that there is very little extra information to be gained from discarding the assumption of symmetry. However, there is evidence that improved performance can be gained by extending the neighborhood to a distance of two pixels for, given perfect information, the misclassification probabilities are approximately halved for most images in the storm under this neighborhood structure. Although in practice we do not expect to obtain this level of performance (for this can be obtained only when we know the true configuration everywhere except at a single pixel), it gives a very clear indication that there is substantial information

in an extended neighborhood. It is therefore likely that this is the primary source of discrepancy between model and data. It is also probably responsible for the smoothness and lack of diagonal banding exhibited by the reconstruction in Fig. 2.

Further investigations of PIMPs have been carried out to try and determine a feasible neighborhood structure for practical purposes, by introducing various symmetries into the two-neighborhood structure. Unfortunately, the reduction seen in Fig. 4 is achieved only when each of the 2^{24} possible neighborhood configurations is treated separately. This is clearly hopelessly impracticable. So we conclude that the performance from the current neighborhood structure is probably as good as can feasibly be achieved.

e. Comparison with coverage-only model

Intuitively, it would appear that the scheme is essentially being driven by the relationship between coverage and mean rainfall, and it is therefore of interest to investigate the scheme's performance if the spatial and temporal dependence components are omitted. This amounts to setting $\alpha = \beta_1 = \beta_2 = 0$, and $p_{11} = p_{10} = 0.5$. The analyses above have been repeated for the simplified model. As expected, the appearance of the reconstructions is rather poor, but the performance is also much worse in terms of the statistical criteria. The percentage of correctly classified pixels is around five to seven percentage points lower in each image. This is to be expected. Spatial dependence will encourage wet pixels in a square to position themselves along the borders with other wet squares, whereas in a coverage-only model, the pixels will be allocated randomly.

A surprising result from this exercise is that the simplified model underestimates the coverage by approximately 0.15–0.2 throughout the storm. This means probably that the parameters used for the coverage–mean regression relationship are inappropriate for this particular storm. It is likely that the inclusion of spatial and temporal dependence has an effect of constraining the disaggregations in a realistic way, which renders the procedure insensitive to misspecification of the regression parameters. This suggests that the proposed scheme should be fairly robust and that the inclusion of spatial–temporal dependence into a disaggregation scheme is worth the additional complexity.

5. Conclusions

The disaggregation scheme presented here has a number of advantages over many existing schemes. The most important is the incorporation of spatial and temporal memory, which are likely to be important determinants of surface hydrological response. Many other authors—Onof and Wheater (1996b), Short et al. (1993), and Kedem and Pavlopoulos (1991), for example—have remarked upon the relationship between coverage and mean areal rain rate. The methodology proposed here uses this relationship in such a way as to ensure spatial and temporal consistency of the resulting scenarios. The basic idea can be made to incorporate further physical information where available. The starting point is Bayes' Theorem at equation (1). In the simplest case, further information can be incorporated by extending the regression relationship describing $\Pr\{\mathbf{P}_{\mathbf{x}}(t) \mid \mathbf{r}(t)\}$ as appropriate. An alternative would be to incorporate extra information via the αs and βs in (3). This would be particularly appropriate for incorporating physical information regarding convective activity, for example, since the βs effectively control the finescale spatial organization of rainfall. Topographic effects could also be handled in this way.

The diagnostic procedures developed in section 4 are simple to implement and have been useful indicators of how the model could be improved. Many problems with an earlier version of the model, reported in Chandler et al. (1997), were identified using these and similar procedures, resulting in substantial improvements for the current paper.

The emphasis here has been to establish whether or not an image reconstruction approach to the disaggregation problem is viable. The methodology appears promising. A large program of data analysis is required to investigate the variation of model parameters with spatial and temporal scale and to apply the methodology to different areas. Space limitations preclude the presentation of more results in the current article. However, see Onof et al. (1998) for results obtained from the southwest of England. The methodology has also been applied successfully to data for all of northern Europe, as well as the area of the United States considered here. The robustness of the algorithm across a wide variety of weather types indicates that any theoretical problems with the model, as discussed in section 2c, have no practical significance. Nonetheless, this is an area that will be addressed in future work. In addition, the approach should be extended to allocate finescale rainfall intensities once the wet–dry rainfall pattern is known.

Acknowledgments. The authors would like to thank Chris Jennison (Department of Mathematics, University of Bath) for supplying preprints of several papers that were invaluable in the early stages of this work, and Tom Fearn for constructive comments regarding the use of the Gibbs Sampler. The data used in sections 3 and 4 were obtained via the Internet from the U.S. National Weather Service at http://www.abrfc.noaa.gov/

This work was funded by the Natural Environment Research Council under the auspices of the HYREX (REC) and TIGER (NM) projects.

REFERENCES

Bárdossy, A., 1997: Downscaling from GCMs to local climate through stochastic linkages. *J. Environ. Manage.,* **49,** 7–17.

Besag, J. E., 1972: On the statistical analysis of nearest-neighbour systems. *Proc. European Meeting of Statisticians,* Budapest, Hungary.

——, 1974: Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Stat. Soc., Ser. B,* **36,** 192–236.

——, 1986: On the statistical analysis of dirty pictures. *J. Roy. Stat. Soc., Ser. B,* **48,** 259–302.

Chandler, R. E., 1998: Model checking. *Encyclopedia of Biostatistics,* P. Armitage and T. Colton, Eds., Vol. 4, Wiley and Sons, 2666–2668.

——, N. Mackay, and C. Onof, 1997: Bayesian image analysis and the disaggregation of rainfall. *Proceedings in the Art and Science of Bayesian Image Analysis,* K. Mardia et al., Eds., Leeds University Press, 132–142.

Collier, C. G., 1989: *Applications of Weather Radar Systems: A Guide to Uses of Radar Data in Meteorology and Hydrology.* Ellis Horwood, 390 pp.

Cox, D. R., and E. J. Snell, 1989: *Analysis of Binary Data.* 2d ed. Chapman and Hall, 236 pp.

Geman, S., and D. Geman, 1984: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.,* **6,** 721–741.

Jennison, C., 1986: Contribution to the discussion of Professor Besag's paper. *J. Roy. Stat. Soc., Ser. B,* **148,** 288–289.

——, and M. D. Jubb, 1988: Statistical image resoration and refinement. *Information Processing in Medical Imaging,* C. de Graaf and M. Viergever, Eds., Plenum, 255–262.

Jubb, M. D., and C. Jennison, 1991: Aggregation and refinement in binary image restoration. *Spatial Statistics and Imaging,* A. Possolo, Ed., IMS Lecture Notes Series, Vol. 20, Institute of Mathematical Statistics, 150–162.

Kedem, B., and P. Pavlopoulos, 1991: An analysis of the threshold method for measuring area-average rainfall. *J. Amer. Stat. Assoc.,* **86,** 626–633.

Lebel, T., I. Braud, and J.-D. Creutin, 1998: A space–time rainfall disaggregation model adapted to Sahelian mesoscale convective complexes. *Water Resour. Res.,* **34,** 1711–1726.

Marsaglia, G., and A. Zaman, 1991: A new class of random number generators. *Ann. Appl. Probab.,* **1,** 462–480.

Onof, C., and H. S. Wheater, 1996a: Analysis of the spatial coverage of British rainfall fields. *J. Hydrol.,* **176,** 97–113.

——, and ——, 1996b: Modelling of the time series of spatial coverages of British rainfall fields. *J. Hydrol.,* **176,** 115–131.

——, N. Mackay, R. E. Chandler, and H. S. Wheater, 1998: A rainfall disaggregation scheme for forecasting. *Hydrology in a Changing Environment,* H. Wheater and C. Kirby, Eds., Wiley and Sons, 107–116.

Perica, S., and E. Foufoula-Georgiou, 1996: A model for multiscale disaggregation of spatial rainfall based on coupling meteorological and scaling descriptions. *J. Geophys. Res.,* **101** (D21), 26 347–26 361.

Ripley, B. D., 1988: *Statistical Inference for Spatial Processes.* Cambridge University Press, 148 pp.

Short, D. A., K. Shimizu, and B. Kedem, 1993: Optimal threshold for the estimation of area rain-rate moments by the theshold method. *J. Appl. Meteor.,* **32,** 182–192.

Warrilow, D. A., A. B. Sangster, and A. Slingo, 1986: Modelling of land surface processes and their influence on the European climate. UKMO Tech. Rep. DCTN38, 92 pp. [Available from The Meteorological Office, Bracknell, Berkshire RG12 2SZ, United Kingdom.]