Journal of Hydrology (2006) **330**, 586–595

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jhydrol

A spatial–temporal point process model of rainfall for the Thames catchment, UK

Paul S.P. Cowpertwait

Institute of Information and Mathematical Sciences, Massey University at Albany, Private Bag 102-904, Auckland, New Zealand

Received 3 November 2005; received in revised form 5 April 2006; accepted 14 April 2006

KEYWORDS
Point processes;
Time series;
Spatial–temporal models;
Urban rainfall;
Multisite data

Summary A spatial–temporal point process model of rainfall is fitted to data taken from 23 sites across the Thames catchment, London. For the period January 1970 to December 1988 the sites have daily data, whilst for the period January 1989 to November 2003 the sites have hourly data. In addition, the records contain missing values. The fitted model is used to infill the missing values, to disaggregate the daily data to hourly data and to generate spatially consistent series at a further 25 sites across the catchment. The model is validated by considering statistical properties that are not used in the fitting procedure, which include extreme values, the proportion of dry intervals, and annual totals. Overall the performance of the model is good, with the exception of an over-estimation in cross-correlation between pairs of sites for the disaggregated series.

Objectives and background

In order to design and upgrade a sewer network, engineers require rainfall data as input to hydraulic models, and have sought models of rainfall because historical records of data are limited. In recent years, there has been considerable work on developing stochastic models of rainfall to simulate long records of rainfall time series data which can be used as input to a sewer flow network model to simulate system performance under a range of conditions. For example, in the UK the commercial software package 'Stormpac' was developed for this purpose (Water Research Centre, 1994). Therefore, when coupled with a hydraulic flow simulation model, a stochastic model of rainfall can be used to create a range of scenarios, which can give insight into probable system performance and therefore help decision making when upgrading or redesigning a drainage system.

The objective of the work described here is to fit a spatial–temporal stochastic model of rainfall to historical data taken from 23 sites in the Thames region and to use the fitted stochastic model to: (i) infill missing values; (ii) disaggregate daily data to hourly data; (iii) simulate data at a further 25 sites over the catchment which have no historical data; and (iv) extend the records by simulating additional hourly data to give a total of 100 years of

E-mail address: p.s.cowpertwait@massey.ac.nz

0022-1694/$ - see front matter
doi:10.1016/j.jhydrol.2006.04.043

data at each site. To achieve this objective, a methodology based on the Neyman–Scott rectangular pulses (NSRP) temporal stochastic rainfall model, originally developed by Rodriguez-Iturbe et al. (1987) and extended to a spatial–temporal process by Cowpertwait (1995), is used.

There are many alternative approaches and methodologies available to disaggregation, downscaling, infilling, or spatial modelling of rainfall. For example, there are models for downscaling the output from deterministic global circulation models (e.g. Skaugen, 2002; Venugopal et al., 1999; Charles et al., 1999; Mehrotra and Singh, 1998; Lebel et al., 1998; Gao and Sorooshian, 1994), and there are models aimed more specifically at simulating fine resolution data for urban catchment studies (e.g. Hingray et al., 2002; Cowpertwait, 2001; Durrans et al., 1999; Cowpertwait et al., 1996a,b; Koutsoyiannis, 1994; Ormsbee, 1989). In addition, models based on stochastic point processes have been used for spatial–temporal series (e.g. Cox and Isham, 1988; Northrop, 1998; Cowpertwait, 1995 and for disaggregation (e.g. Glasby et al., 1995; Onof et al., 1996; Bo et al., 1994; Koutsoyiannis et al., 2003). The spatial NSRP model is also based on a stochastic point process, and was shown to perform well when tested against multisite historical data taken from the Arno Catchment in Italy (Cowpertwait et al., 2002). In this paper, the model is applied to the problem of spatial–temporal disaggregation and infilling (or incomplete data).

The spatial NSRP model is simple relative to the underlying physical rainfall process and as such uses only a modest number of parameters. In addition, unlike some of the more complex space–time stochastic rainfall models (e.g. Mellor and Metcalfe, 1996; Northrop, 1998), the spatial NSRP model given here does not require radar data for calibration, which is an advantage from the practical viewpoint. Of course, model selection is highly dependent on the task in hand, and involves considering a range of questions, e.g. the required level of accuracy, which can go beyond simple practical constraints. Nevertheless, it is hoped that the results described herein demonstrate that the spatial NSRP model can be used to solve some general problems related to incomplete data, which arise in catchment studies and urban drainage design.

Spatial–temporal point process model

Model definition

Let the arrival times $\{T_i\}$ of storms ('storm origins') occur in a Poisson process with rate λ (per hour), so that the times between adjacent storm origins are independent exponential random variables with mean λ^{-1}. A 'storm' consists of a stochastic process of rain cells $\{(U_{ij}, V_{ij}), S_{ij}, L_{ij}, X_{ij}, R_{ij}\}$ where for the ith storm:

1. $\{(U_{ij}, V_{ij})\}$ forms a two-dimensional Poisson process with rate ζ (per km^2 per hour).
2. (U_{ij}, V_{ij}) and R_{ij} form discs in two-dimensional space, where (U_{ij}, V_{ij}) is the disc centre and R_{ij} is the disc radius which is taken to be an independent exponential random variable with parameter ϕ.
3. S_{ij} is the arrival time of the jth cell in the ith storm, where $S_{ij} - T_i$ are independent exponential random variables with parameter β, so that the cell arrival times $\{S_{ij}\}$ form a Neyman–Scott point process.
4. L_{ij} is the lifetime of the jth cell, which is taken to be an independent exponential random variable with parameter η, so that the jth cell in the ith storm terminates at a time $S_{ij} + L_{ij}$.
5. X_{ij} is a random variable representing the intensity of the jth cell in the ith storm, where X_{ij} remains constant throughout the cell lifetime and over the area of the disc $\{(U_{ij}, V_{ij}), R_{ij}\}$.
6. The total intensity at time t and location $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, $Y(\mathbf{x}, t)$ is the sum of the intensities of all cells alive at time t and overlapping $\mathbf{x}$.

For the purpose of model fitting and simulation, the cell intensity X will be taken to be a Weibull random variable with survivor function $P(X > x) = \mathrm{e}^{-(x/\theta)^{\alpha}}$ and moments given by $E[X^r] = \theta^r \Gamma(1 + r/\alpha)$ $(r = 0, 1, 2, \ldots)$. Furthermore, for each storm the number of cells C that overlap a point in two-dimensional space $\mathbb{R}^2$ is a Poisson random variable with mean $\mu_C = 2\pi\zeta/\phi^2$. Consequently, the model may be summarised by the following parameters:

λ^{-1}	– the mean time (h) between adjacent storm origins;
β^{-1}	– the mean waiting time for a cell origin after a storm origin;
μ_C	– the mean number of rain cells per storm;
η^{-1}	– the mean cell lifetime;
α	– shape parameter for cell intensity;
θ	– scale parameter (mm/h) for cell intensity;
ϕ^{-1}	– the mean cell radius (km).

Properties of the aggregated process

As rainfall data are usually sampled over discrete time intervals, it is necessary to consider the aggregated time series $\{Y_k^{(h)}(\mathbf{x})\}$, where

$$Y_k^{(h)}(\mathbf{x}) = \int_{(k-1)h}^{kh} Y(\mathbf{x}, t)\,\mathrm{d}t \tag{1}$$

Statistical properties of the aggregated time series have been derived by Cowpertwait (1995, 1998) and are cited below

$$\mu_h = E\{Y_k^{(h)}(\mathbf{x})\} = \lambda\mu_C\mu_X h/\eta \tag{2}$$

$$\begin{aligned}\gamma_{\mathbf{x},\mathbf{y},h,l} &= Cov\{Y_k^{(h)}(\mathbf{x}), Y_{k+l}^{(h)}(\mathbf{y})\} \\ &= \gamma_{\mathbf{x},\mathbf{x},h,l} - 2\lambda\{1 - P(\phi, d)\}\mu_C E(X^2)A(h, l)/\eta^3\end{aligned} \tag{3}$$

$$\begin{aligned}\gamma_{\mathbf{x},\mathbf{x},h,l} = \gamma_{\mathbf{y},\mathbf{y},h,l} &= \lambda\eta^{-3}A(h, l)\{2\mu_C E(X^2) \\ &\quad + \mu_X^2\beta^2 E(C^2 - C)/(\beta^2 - \eta^2)\} \\ &\quad - \lambda\mu_X^2 B(h, l)E(C^2 - C)/\{\beta(\beta^2 - \eta^2)\}\end{aligned} \tag{4}$$

where $A(h, 0) = (h\eta + \mathrm{e}^{-\eta h} - 1)$; $B(h, 0) = (h\beta + \mathrm{e}^{-\beta h} - 1)$; for l a positive integer, $A(h, l) = \frac{1}{2}(1 - \mathrm{e}^{-\eta h})^2\mathrm{e}^{-\eta h(l-1)}$, $B(h, l) = \frac{1}{2}(1 - \mathrm{e}^{-\beta h})^2\mathrm{e}^{-\beta h(l-1)}$; $P(\phi, d) = \frac{2}{\pi}\int_0^{\pi/2}\{1 + \phi d/(2\cos y)\}\exp\{-\phi d/(2\cos y)\}\,\mathrm{d}y$ and $d = \|\mathbf{x} - \mathbf{y}\|$ is the distance between points $\mathbf{x}$ and $\mathbf{y}$.

$$\xi_{\mathbf{x},h} = E[\{Y_k^{(h)}(\mathbf{x}) - \mu_h\}^3]$$
$$= 6\lambda\mu_C E(X^3)(\eta h - 2 + \eta h e^{-\eta h} + 2e^{-\eta h})/\eta^4$$
$$+ 3\lambda\mu_X E(X^2)E(C^2 - C)p(\eta,\beta,h)/\{2\eta^4\beta(\beta^2 - \eta^2)^2\}$$
$$+ \lambda\mu_X^3 E\{C(C-1)(C-2)\}q(\eta,\beta,h)/\{2\eta^4\beta(\eta^2 - \beta^2)$$
$$\times(\eta - \beta)(2\beta + \eta)(\beta + 2\eta)\} \quad (5)$$

where p and q are high-order functions in η and β and are given in Cowpertwait (1998). As C is a Poisson random variable, $E(C^2 - C) = \mu_C^2$ and $E[C(C-1)(C-2)] = \mu_C^3$ in Eqs. (4) and (5).

Following Cowpertwait et al. (2002), the following dimensionless properties are used when fitting the model (''Fitting procedure''):

1. the coefficient of variation: $v_h = \gamma_{\mathbf{x},\mathbf{x},h,0}/\mu_h$;
2. the coefficient of skewness: $\kappa_h = \xi_{\mathbf{x},h}/\gamma_{\mathbf{x},\mathbf{x},h,0}^{3/2}$;
3. the correlation: $\rho_{\mathbf{x},\mathbf{y},h,l} = \gamma_{\mathbf{x},\mathbf{y},h,l}/\sqrt{\gamma_{\mathbf{x},\mathbf{x},h,0}\gamma_{\mathbf{y},\mathbf{y},h,0}}$.

Fitted stochastic model

Moment estimators

Suppose we have $N(h)$ years of data sampled over discrete time intervals of width h hours at M sites across a catchment. Let $x_{ijkl}^{(h)}$ be the observed rainfall depth for site i, year j, month k, and interval l ($i = 1, \ldots, M$, $j = 1, \ldots, N(h)$, $k = 1, \ldots, 12$, $l = 1, \ldots, n(h,k)$, where $n(h,k)$ is the number of intervals of width h in month k).

To allow for seasonal variation, sample statistics are extracted for each calendar month. Initially, the sample hourly mean is found using data sampled over 24-h intervals, which would include daily data and aggregated hourly data. (For many applications, including the one described here, more data is available at the daily level of aggregation; the procedure below thus makes use of all data available at varying aggregation levels.)

$$\bar{x}_{ik}^{(1)} = \sum_{j=1}^{N(24)}\sum_{l=1}^{n(24,k)} x_{ijkl}^{(24)}/\{24N(24)n(24,k)\} \quad (6)$$

The following sample statistics are then found for each month k by pooling all available data across years and sites. The statistics are scaled by the sample mean (from Eq. (6)) so that they are dimensionless:

$$\hat{\sigma}_{h,k}^2 = \sum_{i=1}^{M}\sum_{j=1}^{N(h)}\sum_{l=1}^{n(h,k)}\left\{x_{ijkl}^{(h)}/\bar{x}_{ik}^{(1)} - h\right\}^2/\{MN(h)n(h,k)\} \quad (7)$$

$$\hat{\gamma}_{h,k} = \sum_{i=1}^{M}\sum_{j=1}^{N(h)}\sum_{l=1}^{n(h,k)-1}\{x_{ijkl}^{(h)}/\bar{x}_{ik}^{(1)} - h\}$$
$$\times\{x_{ijkl+1}^{(h)}/\bar{x}_{ik}^{(1)} - h\}/\{MN(h)(n(h,k)-1)\} \quad (8)$$

$$\hat{\kappa}_{h,k} = \sum_{i=1}^{M}\sum_{j=1}^{N(h)}\sum_{l=1}^{n(h,k)}\{x_{ijkl}^{(h)}/\bar{x}_{ik}^{(1)} - h\}^3/\{\hat{\sigma}_{h,k}^3 MN(h)n(h,k)\} \quad (9)$$

$$\hat{\rho}_{x,y,h,k} = \frac{\sum_{j=1}^{N(h)}\sum_{l=1}^{n(h,k)}\left\{x_{xjlk}^{(h)} - \bar{x}_{xk}^{(h)}\right\}\left\{x_{yjlk}^{(h)} - \bar{x}_{yk}^{(h)}\right\}}{\sqrt{\sum_{j=1}^{N(h)}\sum_{l=1}^{n(h,k)}\left\{x_{xjlk}^{(h)} - \bar{x}_{xk}^{(h)}\right\}^2\sum_{j=1}^{N(h)}\sum_{l=1}^{n(h,k)}\left\{x_{yjlk}^{(h)} - \bar{x}_{yk}^{(h)}\right\}^2}} \quad (10)$$

$$\hat{v}_{h,k} = \hat{\sigma}_{h,k}/h \quad (11)$$

$$\hat{\rho}_{h,k} = \hat{\gamma}_{h,k}/\hat{\sigma}_{h,k}^2 \quad (12)$$

Historical data

Rainfall data for the period January 1970 to November 2003 were available for 23 sites in the Thames catchment (see Table 1; Fig. 1). For each site, the data were available at the daily level of aggregation over the period 1970 (January) to 1988 (December) and at the hourly level of aggregation over the period 1989 (January) to 2003 (November).

A further 25 site locations (which had no historical data, but for which simulated data were required) were also given to provide a spread of sites over the Thames catchment (Fig. 1). For each month, a regression model of site mean on altitude was fitted and used to estimate the mean rainfall at these sites. Examples of the fitted regression models are shown in Fig. 2.

Fitting procedure

Sample properties $\hat{g}_i$, based on moments up to third order and autocorrelation, were calculated for each month by pooling all available data across the Thames region. For daily statistics, the daily data were included in the calculation; for aggregation levels of less than 24 h, the 15-year records of hourly data were used. (Following Cowpertwait et al. (2002), it is assumed that the ensemble properties are approximately stationary over the period of a calendar month.)

Table 1 Rainfall gauges from the Thames catchment: January 1970 to November 2003

Site reference	Easting (0.1 km)	Northing (0.1 km)	Altitude (m)
287864	5299	1661	35
246847	5208	1870	42
238097	5499	1863	16
287141	5247	1641	47
291467	5468	1754	50
289102	5377	1771	5
239258	5412	1981	115
239578	5447	1830	2
286392	5194	1682	12
246424	5246	1795	21
245176	5308	1894	33
288749	5375	1692	33
246213	5314	1828	25
287283	5234	1737	56
239320	5418	1923	17
239374	5415	1882	8
290007	5486	1805	8
247119	5141	1815	23
246627	5241	1920	78
238605	5476	2048	75
239315	5423	1926	15
289022	5433	1745	75
288020	5286	1712	40

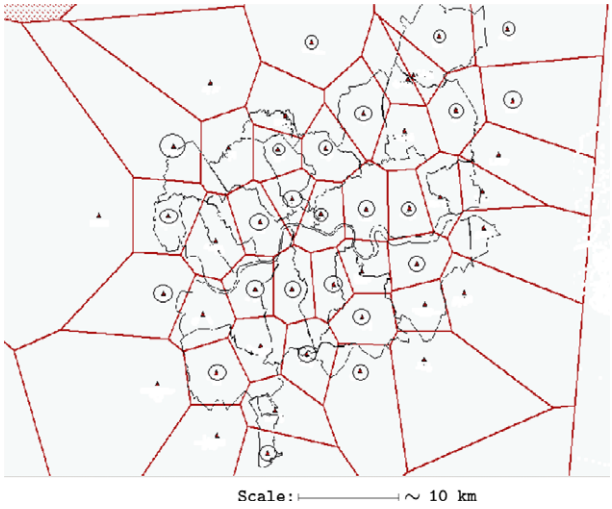

Figure 1 Sites for the Thames catchment. Gauges are shown as triangles within an associated sub-catchment; gauges circled are those that had no historical data available.

Harmonic curves were fitted through the sample estimates using least squares regression, i.e. if $\hat{g}(i)$ is the estimate for the ith calendar month ($i = 1, \ldots, 12$), and ϵ_i is random error, then the harmonic model $\hat{g}(i) = c_0 + \sum_{j=1}^{5}\{c_j \cos(2\pi ij/12) + s_j \sin(2\pi ij/12)\} + \epsilon_i$ was fitted using stepwise regression to ensure only those terms (c_j, s_j) of significance were included in the final model. (The S 'step' function discussed in Venables and Ripley (2002, p. 175), was used for this purpose.) This procedure assumes that the ensemble properties should have a seasonal variation that varies smoothly over the calendar months, and, therefore, reduces 'between month' sampling error. The fitted value for the ith month is denoted as $\hat{f}(i)$ to distinguish it from the pooled sample estimate $\hat{g}(i)$.

The model was fitted to the sample properties, by minimising the following sum of squares:

$$\sum_{i=1}^{n} w_i \cdot \left\{ \left(1 - \frac{\hat{f}_i}{f_i}\right)^2 + \left(1 - \frac{f_i}{\hat{f}_i}\right)^2 \right\} \qquad (13)$$

where f_i is a *dimensionless* Neyman–Scott model function and $\hat{f}_i$ is the equivalent sample estimate taken from the historical data. Dimensionless functions, which include the coefficient of variation, skewness and autocorrelation, are used because they do not depend on the scale parameter θ. Consequently, the scale parameter θ can be estimated directly from the mean rainfall when the other parameters are known or estimated. The scale parameter can therefore be modified to account for any non-homogeneity in the rainfall process over the catchment. In Eq. (13), w_i is a weight factor used to force a close fit to the coefficient of variation, because this tends to have a smaller sampling error when compared with skewness. The fitting procedure follows closely the procedure developed by Cowpertwait et al. (2002).

Parameter estimates

The temporal Neyman–Scott model parameters (λ, μ_c, β, η, α) were estimated following the procedure

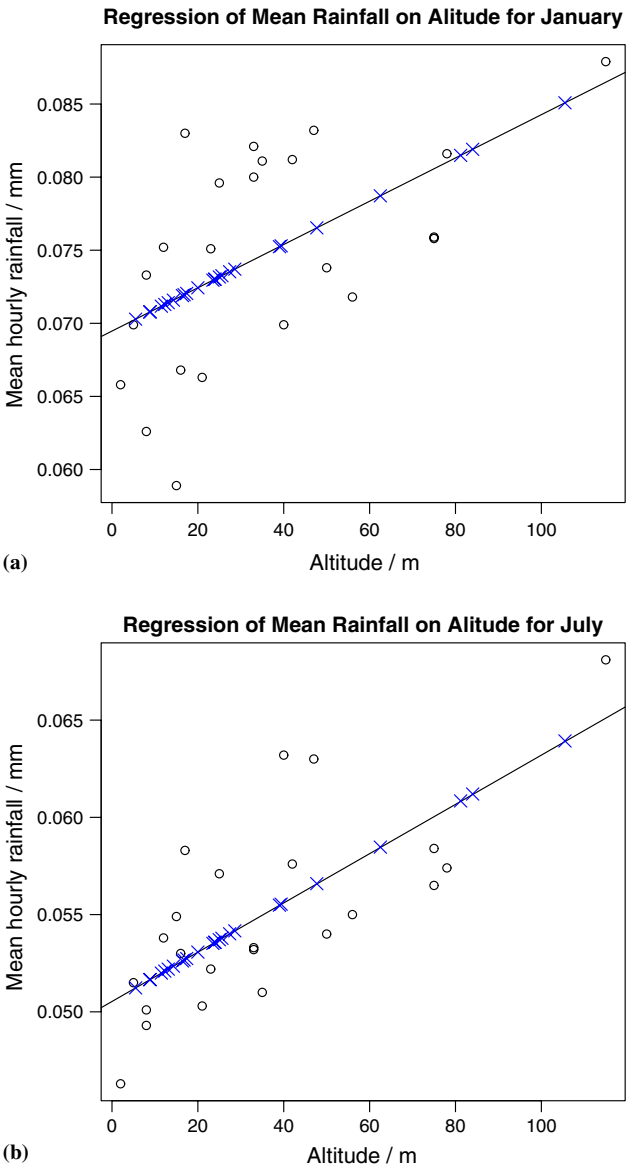

Figure 2 Regression of mean rainfall (○) on altitude (line) for (a) January and (b) July for each of the 23 sites. (Predicted values shown as ×.)

above using the model functions and sample values for: (i) the hourly, 6-hourly, and daily coefficient of variation; (ii) the hourly, 6-hourly, and daily (lag 1) autocorrelation; (iii) the hourly, 6-hourly, and daily skewness. The scale parameter θ was estimated directly from the mean rainfall for each site-month. The cell radius parameter (ϕ) was estimated by minimising the difference between the sample cross-correlation and the equivalent model function (at the 1 h level of aggregation). The sample cross-correlations were calculated for each month using all available hourly values. The fitted parameters are given in Table 2, where a seasonal variation in the estimates can be seen. For example, over the winter months $\hat{\lambda}$ increases and $\hat{\eta}$ and $\hat{\phi}$ both decrease, corresponding to an increase in low-intensity large-scale frontal weather.

Table 2 Parameter estimates

Month	$\hat{\lambda}$ (h^{-1})	$\hat{\mu_c}$ (per storm)	$\hat{\beta}$ (h^{-1})	$\hat{\eta}$ (h^{-1})	$\hat{\alpha}$	$\hat{\phi}$ (km^{-1})
1	0.0125	5.48	0.0752	1.11	1.1	0.0452
2	0.0129	5.51	0.0786	1.26	1.08	0.0432
3	0.0122	7.07	0.0897	1.28	0.893	0.0577
4	0.0106	10.0	0.0957	1.18	0.677	0.0628
5	0.00883	13.0	0.0989	1.43	0.583	0.0856
6	0.00774	11.9	0.0961	2.03	0.597	0.135
7	0.00734	9.64	0.0876	2.44	0.644	0.0978
8	0.00698	11.0	0.0872	2.31	0.623	0.116
9	0.00698	15.7	0.0934	1.77	0.574	0.126
10	0.00804	16.7	0.0962	1.24	0.596	0.0644
11	0.00992	11.1	0.0932	1.02	0.765	0.0571
12	0.0116	6.82	0.0831	0.993	1.01	0.0436

Assessment of fit

Overall the model had a close fit to the sample values used in the fitting procedure; Figs. 3–5 are given as typical examples. Whilst this might be highly desirable, it is also expected because these statistics appear in the minimisation algorithm (13). In the following section, more stringent tests are applied which look at properties that are not directly used in the fitting procedure. (In this section and those following, only the most essential validation plots are given; the remaining figures can be obtained via http://www.massey.ac.nz/~pscowper/pub.)

Model validation

Simulations

One hundred years of hourly rainfall data were simulated for each of 48 sites across the Thames catchment. Sample statistics taken at 1, 6, 12, and 24 h aggregation levels were extracted from the simulated series (48 sites) and compared with the equivalent statistics taken from the historical series (23 sites). The results of these comparisons showed that the implementation of the stochastic model reproduced the statistical properties of the historical series. The purpose of this exercise was to verify the simulation algorithm, and so the detailed plots are omitted. More important comparisons using the simulated data now follow.

Validation

When validating a model, it is necessary to assess the model against properties that are not used in the fitting procedure, because of possible bias in the estimation of sample statistics. Therefore, the following properties were selected because they had not been used to fit the model but were of practical importance: the extreme values (at the daily level of aggregation, because there were 34 years of historical daily data); the proportion of 'dry' intervals (less than 1 mm of rain) at the 1, 2, 6, 12, and 24 h levels of aggregation; the distribution of the annual totals (standardised to hourly means, to avoid bias due to missing values in the his-

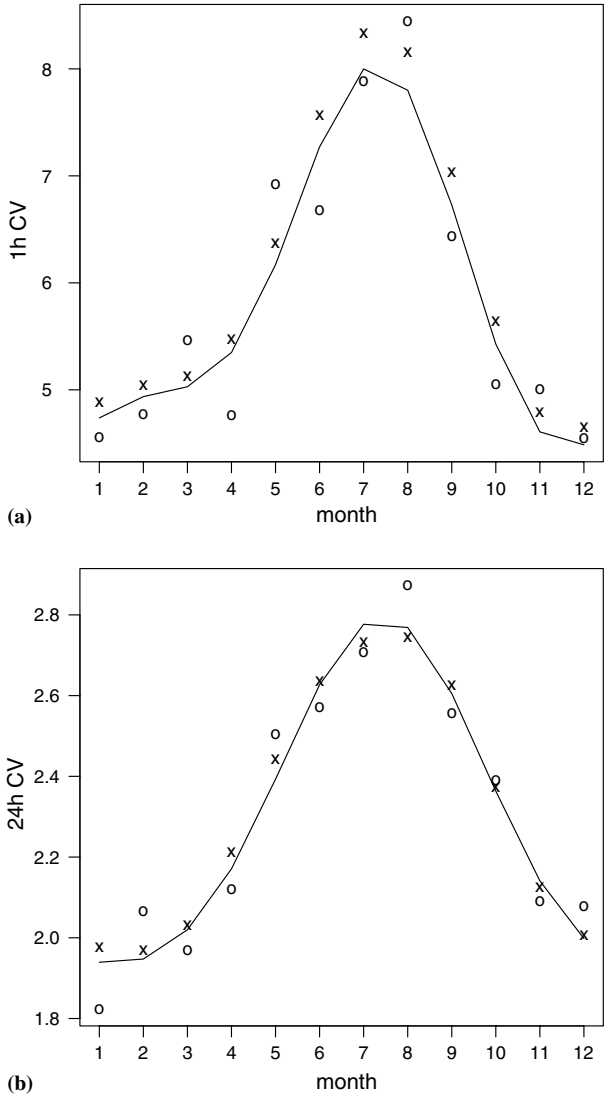

Figure 3 Comparison of coefficient of variation over a 24-h sampling interval for (a) January and (b) July: (——) harmonic estimate; (○) sample estimate (unsmoothed); (×) fitted value.

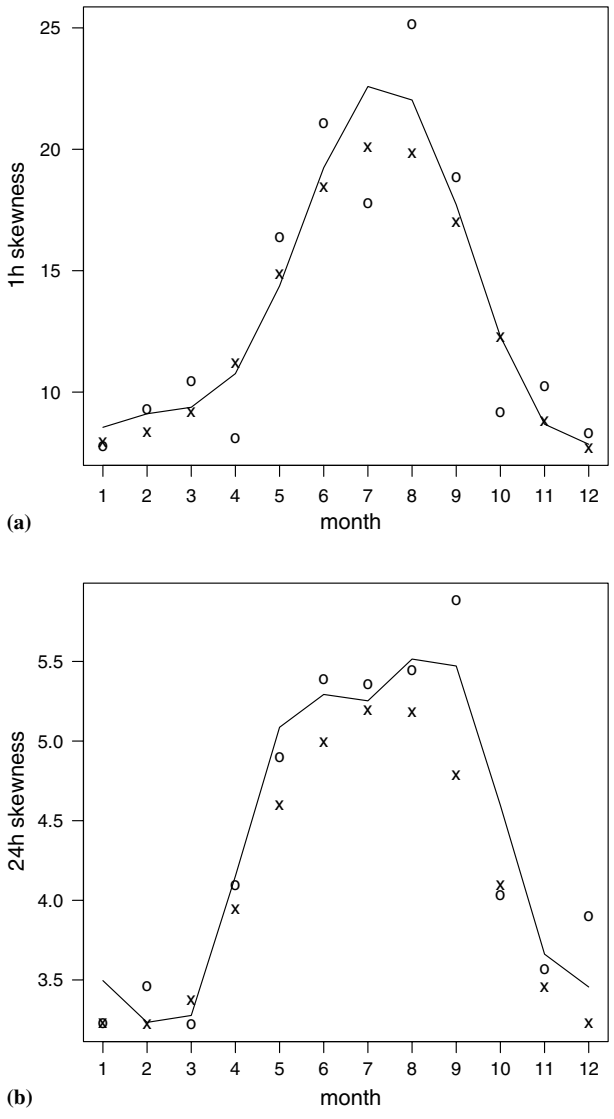

Figure 4 Comparison of coefficient of skewness over a 24-h sampling interval for (a) January and (b) July: (——) harmonic estimate; (○) sample estimate (unsmoothed); (×) fitted value.

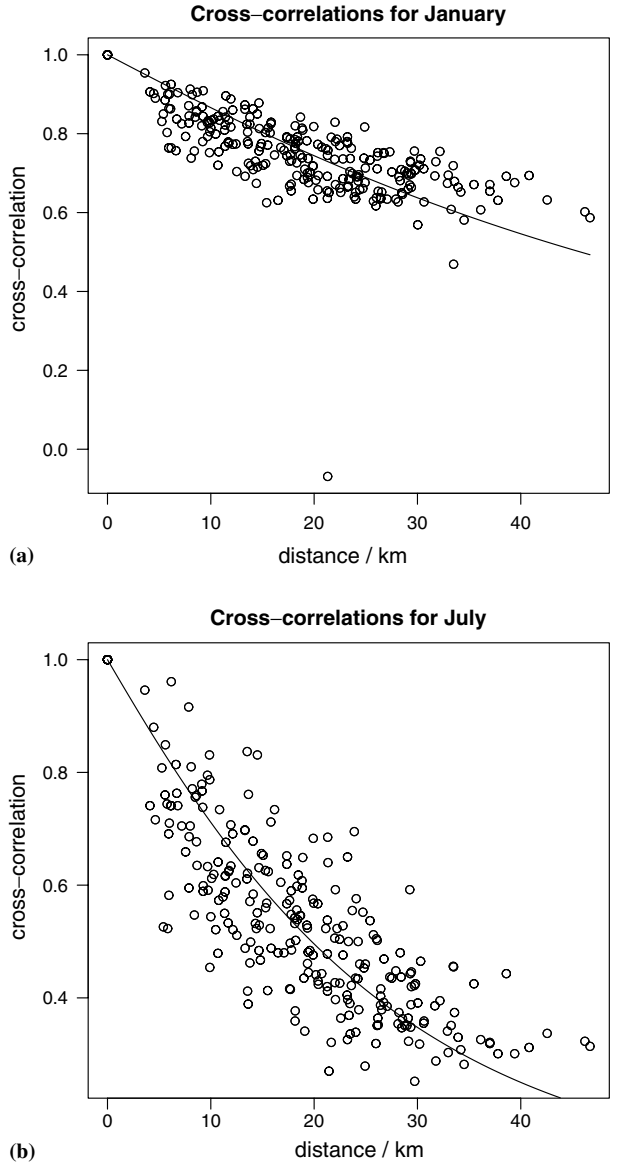

Figure 5 Comparison of historical and fitted 1 h cross-correlations against distance for (a) January and (b) July: (○) sample estimate; (——) fitted value.

torical series); the distribution of hourly depths exceeding a 5 mm threshold.

Simulated annual totals were extracted from the simulated data, and the distribution quantiles found. The distribution quantiles were also found for the historical series. The simulated quantiles were plotted against the historical quantiles and are shown in Fig. 6.

Fig. 6 indicates that the simulated data have annual totals that are representative of the historical data. In addition, a Kolmogorov–Smirnov test comparing the two distributions showed there was not sufficient statistical evidence to reject the null hypothesis that the two distributions were the same. Hence, we conclude that model simulations of annual totals are satisfactory.

For each month, the proportion of 1, 2, 6, 12 and 24 h intervals with rainfall less than 1 mm were extracted from both the historical and simulated series, and plotted pairwise (Fig. 7). The figure indicates that the model reproduces the dry proportions satisfactorily, with no obvious bias.

Two approaches were taken to assess the goodness of fit to extreme values. The first was a traditional Gumbel probability analysis applied to the historical data available at the daily level of aggregation. The second was based on comparing the historical and simulated distributions of hourly depths exceeding 5 mm. The details follow.

For the daily level of aggregation, there were 34 years of data, which were deemed a sufficient number for a standard Gumbel probability analysis. The median values across the sites were extracted for each of the 34 years and plotted with the equivalent simulated values (using the first 34 years of simulated data) on a Gumbel probability plot (Fig. 8). The Gumbel plot indicates that the model is performing well with respect to the daily extremes. (There is a slight overestimation in the tail of the distribution, which could be caused by missing values in the historical series.)

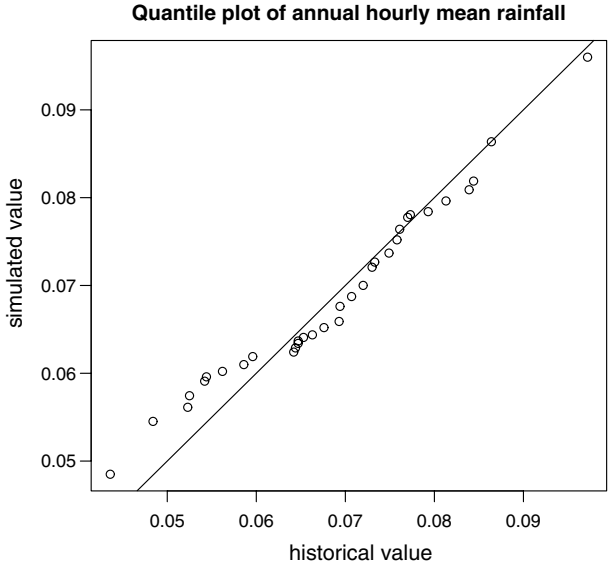

Figure 6 A comparison of the distribution of annual (mean hourly) rainfall. The purpose of the plot is to compare the distribution of annual totals, which are of a much greater aggregation level than the statistics used to fit the model. To avoid missing data causing an under-estimation of the historical value, the annual totals were 'standardised' to hourly means.

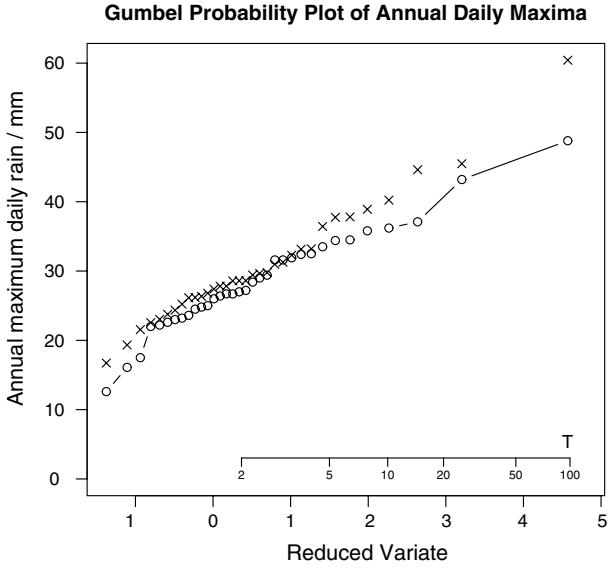

Figure 8 Median daily maxima (taken across all sites) for historical (—○—) and simulated (×) series. The maximum daily rainfall for each of the 34 years was calculated for each site. The medians across the sites were then used in the plots to give 34 values for both the simulated and historical series.

As there were fewer years of hourly data, and missing values would likely become influential in estimating quantiles of annual hourly maxima (≈23% of the hourly data were missing), rainfall depths exceeding a threshold of 5 mm were extracted from the hourly series at the 23 sites. A quantile plot was used to compare the distribution of these values with the equivalent distribution of the simulated values from the 48 sites (Fig. 9). The figure indicates that the two distributions compare very favourably.

Overall, the model fit to the historical data is good and the validation tests satisfactory. This indicates that the model performs well, and that the simulated series can be used with confidence in the disaggregation and infilling procedures which follow.

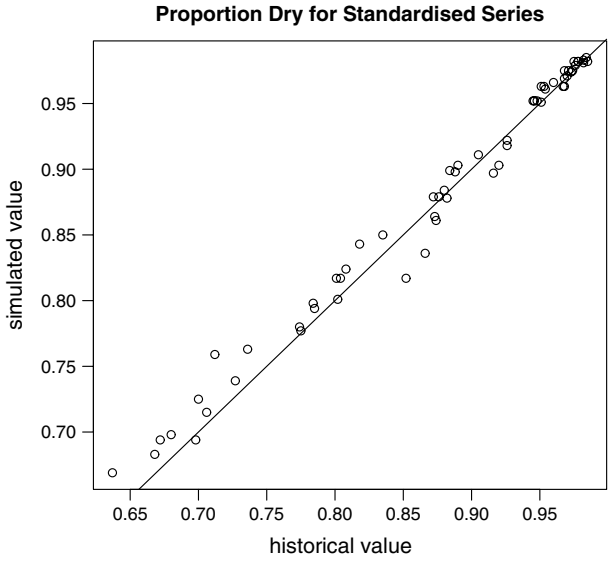

Figure 7 A comparison of historical and simulated proportion of dry intervals at the 1, 2, 6, 12, and 24 h levels of aggregation. The plotted pairs (x, y) correspond to an historical value (x) and a simulated value (y) for each month and aggregation level. The proportion of dry intervals were not used in the fitting procedure.

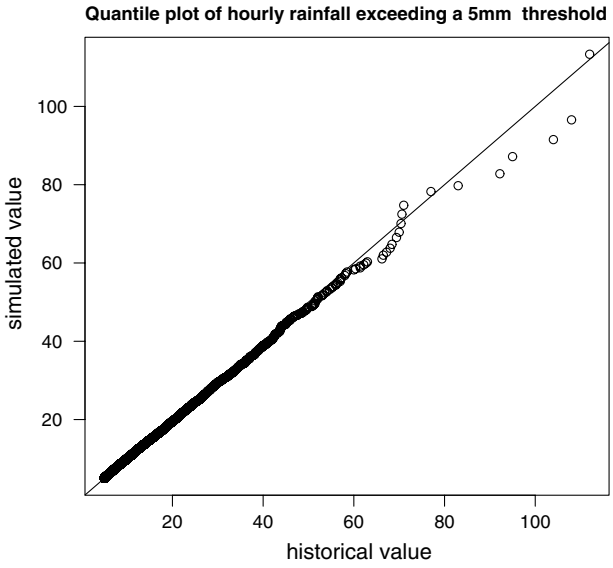

Figure 9 A comparison of the historical and simulated distributions of depths of hourly rainfall exceeding a 5 mm threshold. This plot provides a useful assessment of the distribution of the upper tail (extreme values) when the record is short or contains a large number of missing values.

Disaggregation and infilling of missing values

In the last section, the Neyman–Scott process was fitted to data from the Thames Catchment and used to simulate a 100-year record at 48 sites over the catchment. In this section, the model is used for disaggregating daily data to hourly data, and for infilling missing values in the hourly records. In addition, the model is used to extend the record of 23 sites to 48 sites over the catchment.

Extending the hourly records

Using the fitted model, a further 200-years of hourly data were generated at the 48 sites over the Thames Catchment (giving 300 years in total). For non-missing values, the rainfall in each hour in the historical record was compared with the rainfall in each hour of the simulated record. The date corresponding to the simulated values that were closest to the historical values was recorded. Simulated rainfall at this date were then used at sites with missing data (and for the additional 25 locations that had no data). To avoid the data being too 'patchy', the previous rainfall was also used in the criteria for selecting the best fitting simulated values. Thus, if $x_i(j)$ is the historical rainfall for the ith site (with data) in the jth hour, then the date (or time) k is sought, which minimises the following:

$$\sum_{i=1}^{n}[\{x_i(j)-y_i(k)\}^2+\{x_i(j-1)-y_i(k-1)\}^2] \quad (14)$$

where $y_i(k)$ is the simulated rainfall for the ith site in the kth hour, and n is the number of sites with data (i.e. $n = 23$ when the available historical records had no missing values). Thus, for the case $n = 23$, i.e. for when there are no missing values in the historical records, the 'predicted' values for the additional 25 locations across the catchment were given by: $y_i(k)$ for $i = 24, \ldots, 48$.

Disaggregation of daily data

Using a similar approach to the above, the 300-years of simulated data were used to disaggregate the daily data. For non-missing values, the rainfall in each day in the historical record was compared with the rainfall in each day of the simulated record. The date corresponding to the simulated values that were closest to the historical values was recorded. Hourly simulated rainfall at this date were then used to disaggregate the daily data, infill missing values, and extend the records to the further 25 locations. Again, the previous rainfall was also used in the criteria for selecting the best fitting simulated values. Thus, if $x_i(j)$ is the historical rainfall for the ith site (with data) in the jth day, then the date (or time) k is sought, which minimises the following:

$$\sum_{i=1}^{n}[\{x_i(j)-y_i(k)\}^2+\{x_i(j-1)-y_i(k-1)\}^2] \quad (15)$$

where $y_i(k)$ is the simulated daily rainfall for the ith site in the kth day, and n is the number of sites with data (i.e. $n = 23$ when the available historical records had no missing values). The simulated hourly values for the kth day are scaled to achieve an exact match to the daily total. These are then used to 'disaggregate' the rainfall depth in the kth day to hourly time steps.

Assessment of infilling procedure

To assess the performance of the infilling (and disaggregation) procedure, statistical properties were extracted from the infilled series and compared to the equivalent historical properties. As may be expected, the statistics from the infilled series were found to retain a good fit to the historical mean, variance, skewness, autocorrelation and proportion dry. However, some discrepancies were observed in the cross-correlations at the 1-h aggregation level which are shown in Fig. 10: there is evidence of over-estimation of the 1-h cross-correlation (although the actual magnitude of these differences was not very large). The cause of this

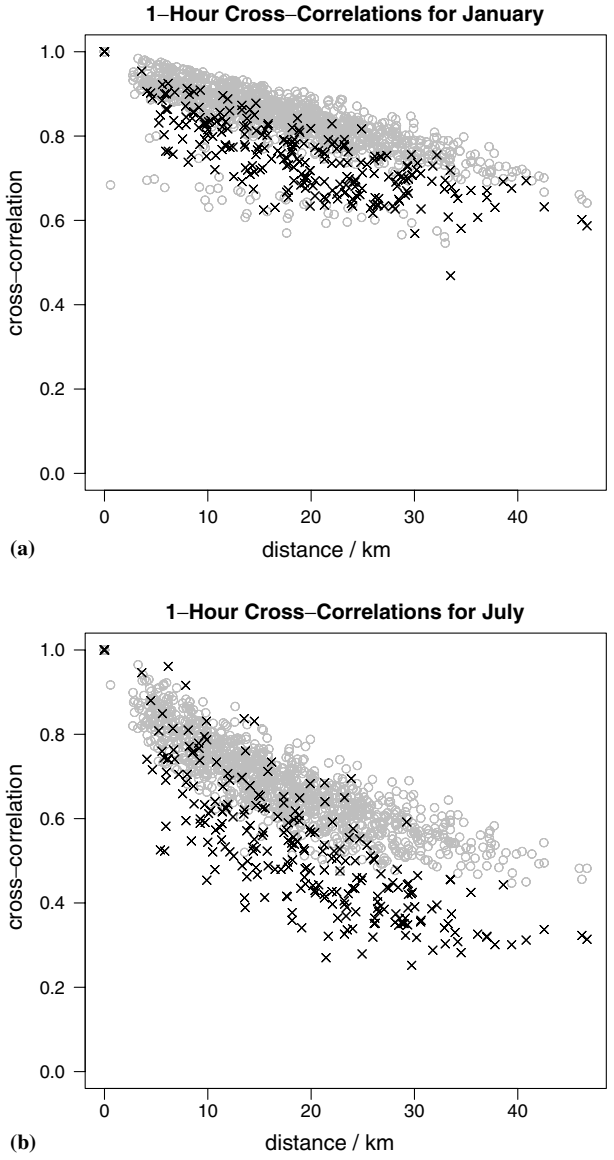

Figure 10 Comparison of historical (×) and disaggregated series (○): cross-correlations for (a) January and (b) July (evaluated using pooled data across all sites).

discrepancy is likely to be in the selection process applied to obtain the infilled series, i.e. minimising squared differences (Eqs. (14) and (15)), rather than in the simulated rainfall which have a good fit to the cross-correlations (Fig. 5).

To address the discrepancy in the cross-correlations, a range of other selection processes were tried, which included using absolute values instead of squared differences in Eqs. (14) and (15). However, these resulted in poor fits to the temporal statistical properties (e.g. using the absolute value gave a significant over-estimation in the mean rainfall). Thus, the use of squared differences was retained. It should be noted that the over-estimation in cross-correlation becomes notably less for higher aggregation levels (e.g. 24 h levels; Fig. 11), and is unlikely to be of practical significance.

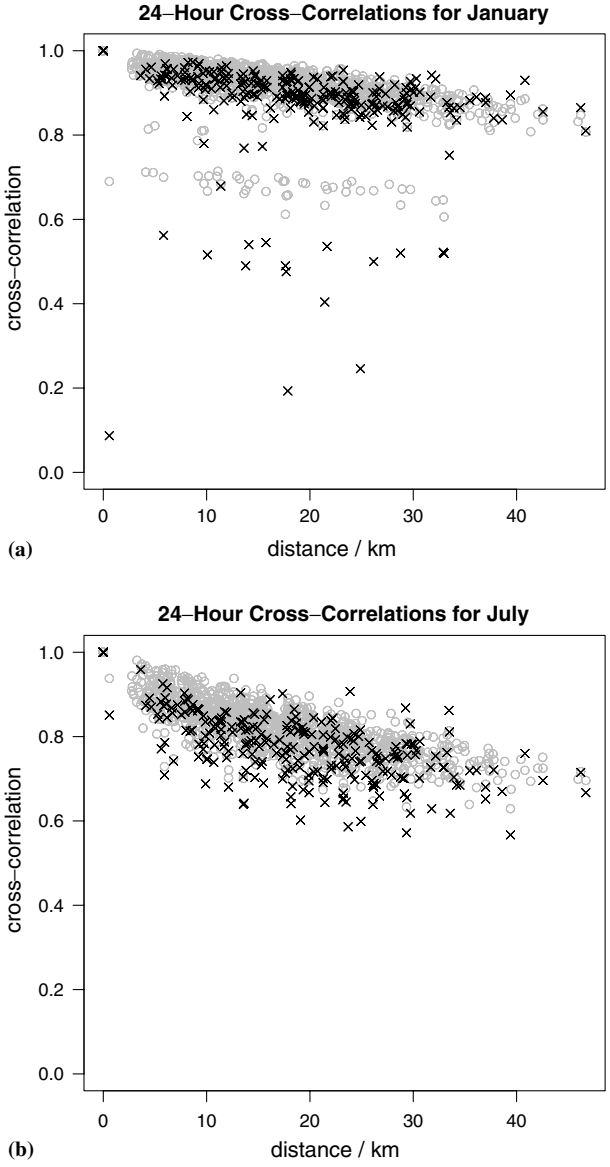

Figure 11 Comparison of historical (×) and disaggregated series (○): cross-correlations for (a) January and (b) July (evaluated using pooled data across all sites).

Conclusions

The NSRP model was fitted to a range of essential statistical properties extracted from historical data taken from the 23 sites across the Thames region, and a good fit was obtained to these properties. In addition, the model was able to closely match a range of important statistical properties that were not used in the fitting procedure, such as extreme values and annual totals. Consequently, the simulated data can be used with reasonable confidence in applications across the Thames catchment, subject to carrying out some additional testing before the data are finally used for design purposes (see below).

The infilled series also preserved most of the essential statistical properties of the historical series. The exception was the cross-correlations at the 1-h time steps, which were generally over estimated by the infilled series. However, this is unlikely to be of practical significance as a good fit was obtained to other statistical properties that are likely to be of greater influence in urban drainage design. In addition, the cross-correlation of the infilled series at larger aggregation levels (e.g. 6 and 24 h) closely matched the historical values.

Clearly in any modelling exercise some discrepancies will be observed. The question remains as to whether the observed discrepancies are of any practical significance for the intended application. Given that the fitted NSRP model performed well in stringent tests using properties not used in the fitting procedure, it is unlikely that the differences between the model and the data will be of any practical consequence. Nevertheless, it would be judicious to make some comparisons that relate more directly to the intended application. For example, a comparison of the distribution of annual spill volumes that result from using the simulated series to those that result from using the historical series would provide an appropriate test of the model for the intended application. A simple flow simulation model could be used for this purpose to gain further confidence in the model and methodology.

Acknowledgements

The author gratefully acknowledges Thames Water plc for giving permission to use the data and results. In addition, the author is grateful to the Water Research Centre (Swindon, UK) for coordinating the project. The referee and editor are also gratefully acknowledged for their helpful comments.

References

Bo, Z., Islam, S., Eltahir, E., 1994. Aggregation—disaggregation properties of a stochastic model. Water Resources Research 30 (12), 3423—3435.

Charles, S., Bates, B., Hughes, J., 1999. A spatiotemporal model for downscaling precipitation occurrences and amounts. Journal of Geophysical Research 104 (D24), 31657—31669.

Cowpertwait, P., 1995. A generalized spatial—temporal model of rainfall based on a clustered point process. Proceedings of the Royal Society of London, Series A 450, 163—175.

Cowpertwait, P., 1998. A Poisson-cluster model of rainfall: high order moments and extreme values. Proceedings of the Royal Society of London, Series A 454, 885—898.

Cowpertwait, P., 2001. A continuous stochastic disaggregation model of rainfall for peak flow simulation in urban hydrologic systems. Research Letters in the Information and Mathematical Sciences 2, 81—88.

Cowpertwait, P., Kilsby, C., O'Connell, P., 2002. A space—time Neyman—Scott model of rainfall: empirical analysis of extremes. Water Resources Research 38 (8), 1—14.

Cowpertwait, P., O'Connell, P., Metcalfe, A., Mawdsley, J., 1996a. Stochastic point process modelling of rainfall: I. Single-site fitting and validation. Journal of Hydrology 175, 17—46.

Cowpertwait, P., O'Connell, P., Metcalfe, A., Mawdsley, J., 1996b. Stochastic point process modelling of rainfall: II. Regionalisation and disaggregation. Journal of Hydrology 175, 47—65.

Cox, D., Isham, V., 1988. A simple spatial—temporal model of rainfall. Proceedings of the Royal Society of London, Series A 415, 317—328.

Durrans, S., Burian, S., Nix, S., Hajji, A., 1999. Polynomial-based disaggregation of hourly rainfall for continuous hydrologic simulation. Journal of the American Water Resources Association 35 (5), 1213—1221.

Gao, X., Sorooshian, S., 1994. A stochastic precipitation disaggregation scheme for GCM applications. Journal of Climate 7 (2), 238—247.

Glasby, C., Cooper, G., McGechan, M., 1995. Disaggregation of daily rainfall by conditional simulation from a point process model. Journal of Hydrology 165, 1—9.

Hingray, B., Monbaron, E., Jarrar, I., Favre, A., Musy, A., 2002. Stochastic generation and disaggregation of hourly rainfall series for continuous hydrological modelling and flood control reservoir design. Water Science and Technology 45 (2), 113—119.

Koutsoyiannis, D., 1994. A stochastic disaggregation method for design storm and flood synthesis. Journal of Hydrology 156, 193—225.

Koutsoyiannis, D., Onof, C., Wheater, H., 2003. Multivariate rainfall disaggregation at a fine time scale. Water Resources Research 39 (7), 1—18.

Lebel, T., Braud, I., Creutin, J., 1998. A space—time rainfall disaggregation model adapted to sahelian mesoscale convective complexes. Water Resources Research 34 (7), 1711—1726.

Mehrotra, R., Singh, R., 1998. Spatial disaggregation of rainfall data. Hydrological Sciences Journal 43 (1), 91—101.

Mellor, D., Metcalfe, A., 1996. The modified turning bands (MTB) model for space—time rainfall. III Estimation of the storm/rainband profile and a discussion of future model prospects. Journal of Hydrology 175, 161—180.

Northrop, P., 1998. A clustered spatial—temporal model of rainfall. Proceedings of the Royal Society of London, Series A 454, 1875—1888.

Onof, C., Faulkner, D., Wheater, H., 1996. Design rainfall modelling in the Thames catchment. Hydrological Sciences Journal 41 (5), 715—729.

Ormsbee, L., 1989. Rainfall disaggregation model for continuous hydrologic modelling. Journal of Hydraulic Engineering 115 (4), 507—525.

Rodriguez-Iturbe, I., Cox, D., Isham, V., 1987. Some models for rainfall based on stochastic point processes. Proceedings of the Royal Society of London, Series A 410, 269—288.

Skaugen, T., 2002. A spatial disaggregation procedure for precipitation. Hydrological Sciences Journal 47 (6), 943—956.

Venables, W., Ripley, B., 2002. Modern Applied Statistics with S, fourth ed. Springer, Berlin.

Venugopal, V., Foufoula-Georgiou, E., Sapozhnikov, V., 1999. A space—time downscaling model for rainfall. Journal of Geophysical Research 104 (D4), 19705—19721.

Water Research Centre, 1994. Stormpac User Guide — Version 1.1. WRc plc, Swindon, UK.