Res. Lett. Inf. Math. Sci., 2003, Vol.**5**, pp 61-70
Available online at http://iims.massey.ac.nz/research/letters/

Mixed Rectangular Pulses Models of Rainfall

P.S.P. Cowpertwait

Institute of Information & Mathematical Sciences
Massey University at Albany, Auckland, New Zealand.

p.s.cowpertwait@massey.ac.nz

In a recent paper ((3)) a fitting procedure for the Neyman-Scott rectangular pulses (NSRP) spatial-temporal model of rainfall was developed. In that paper, the NSRP third moment function was fitted to the equivalent sample value taken at one-hour time intervals. In this paper, the fitting and modelling procedure are extended to ensure a close fit is obtained to further sample properties over a range of time scales. The stochastic model is a 'mixed' model obtained as the superposition of two independent NSRP processes. The model is fitted to hourly data from Auckland, New Zealand, where a good fit to sample properties is obtained. It is found that a special case arises (the superposition of an NSRP process and a Poisson rectangular pulses process) for data over the summer period. A simulation study of extremes over a range of time scales supports the use of the model in hydrological applications.

1 Introduction

In a recent paper (3), a fitting procedure for the Neyman-Scott Rectangular Pulses (NSRP) spatial-temporal model was developed, and the model fitted to multisite data from the Arno Basin in Italy. In the previous paper, the focus was on developing an appropriate fitting procedure for multisite data by amalgamating the results from two previous pieces of work, which involved the development of a spatial-temporal NSRP model and and a third moment function of the temporal NSRP model (see the previous paper and references therein for details).

In the previous paper, the focus was on assessing goodness-of-fit at the $1h$ and $24h$ levels of aggregation, and no attempt was made to fit the model to sample properties (coefficient of variation ν, skewness κ, autocorrelation ρ) at time scales smaller than twenty-four hours and larger than one hour. Furthermore, the third moment function was only fitted to data sampled over one-hour time intervals. The range of properties was limited to enable accurate fit to $1h$ data to be obtained.

The objectives of the work described in this paper are to improve and extend the stochastic model and fitting procedure to enable good fits to be obtained to a wider range of sample properties (up to third order) for a range of time scales. The emphasis here is on fitting to temporal properties rather than the spatial cross-correlation, as the latter can always be fitted after obtaining a good fit to the temporal properties ((3), §3). In summary, the objectives are:

1. To expand and improve the fitting procedure developed in (3) by:
   (a) Including the NSRP third moment function in the fitting procedure at time scales greater than 1-hour;
   (b) Including sample properties taken at the $6h$ level of aggregation;
   (c) Using harmonic curves to provide 'smoothed' sample estimates for each calendar month (to reduce overall sampling error);

(d) Using bootstrap standard errors as weights in the minimization procedure ((3) equations 21, 23) to allow for the sampling error in the estimated coefficient of variation at different sampling intervals.

2. To propose the use of the superposition of more than one independent NSRP process (or a mixture of independent NSRP and Poisson rectangular pulses processes (PRP)) to improve the fit to sample properties over a range of time scales.

3. To provide empirical results which support the above methodology.

2 Mixed NSRP Process

A generalization of the NSRP model to include different cell types was developed in (1). However, the third moment function has not yet been derived for the generalized model (and the derivation is likely to be difficult, though not intractable). Hence, a mixed model using superposed independent NSRP processes is proposed which makes use of existing NSRP functions that have already been derived and cited ((3), equations 5–12).

Consider a temporal NSRP process. For a stationary period, the model is summarized by the following random variables and model parameters (see (3), 'Notation' on p13): (i) the time T between adjacent storm origins is an independent exponential random variable with parameter λ (so storm origins arrive in a Poisson process); (ii) the waiting time W for a cell origin after a storm origin is an independent exponential random variable with parameter β; (iii) the lifetime L of a cell is an independent exponential random variable with parameter η; (iv) the number of cells C per storm is taken to be an independent geometric random variable with mean μ_C; (v) the intensity X of a cell is an independent random variable that remains constant throughout the cell lifetime L, and is taken to be a Weibull random variable, so that $P(X > x) = e^{-(x/\theta)^\alpha}$.

The parametrization of the temporal NSRP model is the same as that in the previous paper with the following two exceptions. First, for convenience, the Weibull power parameter is just α (rather than α^{-1} as in §2.1 of the previous paper). Second, C is taken to be a Geometric random variable instead of Poisson random variable, because the focus here is on the temporal process and we would like to ensure that each storm has at least one rain cell. This implies that $E(C^2 - C) = 2\mu_C(\mu_C - 1)$ in equations 7 and 10, and $E\{C(C-1)(C-2)\} = 6\mu_C(\mu_C - 1)^2$ in equation 10 (3) . Furthermore, using this parametrization implies the second-order properties of the NSRP model are the same as those for the Bartlett-Lewis Rectangular Pulses (BLRP) model (2).

Following closely the notation in the previous paper ((3), equations 5–13, but omitting the 'h' subscript), the following NSRP properties (which are functions of λ, μ_C, β, η, α, θ, and the aggregation level h) are available for fitting the NSRP model to data: the mean function μ; the variance σ^2; the lag 1 autocovariance γ; the third moment ξ. In the fitting procedure the dimensionless properties are used: the coefficient of variation ν; the lag 1 autocorrelation ρ; and the coefficient of skewness κ (see (3), equations 13).

Using the above notation, let $NSRP_i$ represent an NSRP process with parameter set $\mathbb{P}_i = \{\lambda_i, \mu_{C_i}, \beta_i, \eta_i, \alpha_i, \theta_i\}$, and statistical properties $\{\mu_i, \sigma_i^2, \gamma_i, \xi_i\}$ for $i = 1, \ldots, n$. Then the superposition of n independent NSRP processes gives the superposed NSRP process:

$$SNSRP(n) \equiv \sum_{i=1}^{n} NSRP_i \tag{2.14}$$

with parameter set $\mathbb{P}^n = \{\lambda_1, \ldots, \lambda_n, \mu_{C_1} \ldots, \mu_{C_n}, \beta_1, \ldots, \beta_n, \eta_1, \ldots, \eta_n, \alpha_1, \ldots, \alpha_n, \theta_1, \ldots, \theta_n\}$, where $SNSRP(1) \equiv NSRP_1$ (the original NSRP process). (Note that the summation in equation 2.14 is understood to be superposition. Refer to (4) for a general discussion on the superposition of point processes.) Furthermore, the use of superposed processes makes an allowance for different possible storm 'types', e.g. those with predominantly convective cells or stratiform cells.

The statistical properties, at aggregation level h, of the superposed process $SNSRP(n)$ (abbreviated below to S_n) are the sum of the equivalent properties for each NSRP process, i.e.

$$\mu_{s_n}(h) = \mu_1(h) + \cdots + \mu_n(h) \tag{2.15}$$

$$\sigma^2_{s_n}(h) = \sigma^2_1(h) + \cdots + \sigma^2_n(h) \tag{2.16}$$

$$\gamma_{s_n}(h) = \gamma_1(h) + \cdots + \gamma_n(h) \tag{2.17}$$

$$\xi_{s_n}(h) = \xi_1(h) + \cdots + \xi_n(h) \tag{2.18}$$

The dimensionless model functions (f) used to fit the model are then given by:

$$\text{Coefficient of variation, } \nu_{s_n}(h) = \sigma_{s_n}(h)/\mu_{s_n}(h) \tag{2.19}$$

$$\text{Autocorrelation (lag 1), } \rho_{s_n}(h) = \gamma_{s_n}(h)/\sigma^2_{s_n}(h) \tag{2.20}$$

$$\text{Coefficient of skewness, } \kappa_{s_n}(h) = \xi_{s_n}(h)/\sigma^3_{s_n}(h) \tag{2.21}$$

The set of dimensionless model functions for an $SNSRP(n)$ process is $\mathbb{S}_n = \{\nu_{s_n}(h), \rho_{s_n}(h), \kappa_{s_n}(h) : h = 1, \ldots\}$. (Note that an aggregation level h over *discrete* intervals is again used.) The set $\bigcup_i [\mathbb{S}_i \cup \{\mu_{s_i}\}]$ can be used to estimate the parameters in $\mathbb{P}^n$.

3 Fitting Procedure

The fitting procedure proceeds in stages as follows, and was applied to thirty-three years of hourly data from Auckland:

1. A subset $\widetilde{\mathbb{S}}_1 \subset \mathbb{S}_1$ of dimensionless model functions for the $NSRP_1$ process is selected to represent the rainfall process over a *range* of time scales.

   The choice made when fitting to the Auckland data was the coefficient of variation, lag 1 autocorrelation, and coefficient of skewness, each sampled over 1-, 6-, and 24-hour time intervals, i.e. $\widetilde{\mathbb{S}}_1 = \{\nu(h), \rho(h), \kappa(h) : h = 1, 6, 24\}$.

2. For each month i, sample properties ($\hat{g}(i)$) (the ensemble sample equivalents to the functions in $\widetilde{\mathbb{S}}_1$ (1 above)) are calculated by pooling all available data from a region that can be considered approximately homogeneous with respect to rainfall series ((3), equations 14–19). This assumes that the ensemble properties are approximately stationary over the period of a calendar month ((3), p4).

3. Harmonic curves are fitted through the sample estimates obtained above 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

   $\hat{g}(i) = c_0 + \sum_{j=1}^{5} \{c_j \cos(2\pi ij/12) + s_j \sin(2\pi ij/12)\} + \epsilon_i$

   is fitted using stepwise regression to ensure only those terms (c_j, s_j) of significance are included in the final model. (The S 'step' function discussed by (5) (p175), 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 $\hat{g}(i)$. (Note that $f \in \mathbb{S}_1$.)

4. For each calendar month ($i = 1, \ldots, 12$), an $SNSRP(1) \equiv NSRP_1$ process (i.e. the original NSRP process) is fitted by minimizing the following sum of squares ($SS(n)$ with $n = 1$):

$$SS(n) = \sum_{h=1,6,24} w_h \cdot \Big\{ \quad \Big(1 - \frac{\hat{\nu}_{s_n}(h)}{\nu_{s_n}(h)}\Big)^2 + \Big(1 - \frac{\nu_{s_n}(h)}{\hat{\nu}_{s_n}(h)}\Big)^2 + \Big(1 - \frac{\hat{\rho}_{s_n}(h)}{\rho_{s_n}(h)}\Big)^2 + \Big(1 - \frac{\rho_{s_n}(h)}{\hat{\rho}_{s_n}(h)}\Big)^2 + \Big(1 - \frac{\hat{\kappa}_{s_n}(h)}{\kappa_{s_n}(h)}\Big)^2 + \Big(1 - \frac{\kappa_{s_n}(h)}{\hat{\kappa}_{s_n}(h)}\Big)^2 \Big\} \tag{3.22}$$

giving the following set of parameter estimates $\{\hat{\lambda}_{1,i}, \hat{\mu}_{C_{1,i}}, \hat{\beta}_{1,i}, \hat{\eta}_{1,i}, \hat{\alpha}_{1,i}\}$ for $i = 1, \ldots, 12$. Bounded optimisation is used to reduce the parameter space in the search routine. In particular, $\mu_C \geq 1$ so a lower bound of 1 is needed on this parameter when minimizing SS.

The weights w_h are obtained by calculating a non-parametric bootstrap standard error for the sample estimates. For the Auckland data, the standard errors for the coefficient of variation (CV) sampled at 1-, 6-, and 24-hour time intervals were: 0.30, 0.20, and 0.13 respectively (these were the median values across the months). Thus, in equation 3.22, the weights w_1, w_6 and w_{24} for the Auckland series were taken as: 0.30^{-1}, 0.20^{-1}, and 0.13^{-1} respectively. Properties sampled at 24-hour intervals are thus more accurate than those sampled at 1-hour intervals and receive a correspondingly higher weight. (The same weights were also used for the autocorrelation and coefficient of skewness as these depend on the coefficient of variation.)

For the Auckland data the minimum was $SS(1) = 3.26$.

5. The scale parameter θ_i for the ith month ($i = 1, \ldots, 12$) is estimated directly from the sample mean using the equation: $\hat{\theta}_i = \hat{\mu}_i \eta_{1,i} / \{\mu_c \Gamma(1 + \alpha_{1,i}^{-1})\}$, where $\hat{\mu}_i$ is the estimated mean rainfall sampled over 1-hour time intervals. An exact fit is obtained to the sample mean for each month. The resulting estimates for the $NSRP_1$ process are: $\hat{\mathbb{P}}_{1,i} = \{\hat{\lambda}_{1,i}, \hat{\mu}_{C_{1,i}}, \hat{\beta}_{1,i}, \hat{\eta}_{1,i}, \hat{\alpha}_{1,i}, \hat{\theta}_{1,i}\}$ $(i = 1, \ldots, 12)$.

6. The parameters for a superposed $NSRP_2$ process are estimated, by minimizing $SS(2)$, i.e. 3.22 above with $n = 2$. The parameter estimates obtained for the $NSRP_1$ process in 5 above are retained with the exception of λ_1 which is re-estimated. (λ_1 and λ_2 are both given small lower bounds to allow the possibility of the special cases $\lambda_1 \to 0$ or $\lambda_2 \to 0$ arising in the estimation procedure.)

   Some care is needed to avoid over-parametrization and over-fitting because the superposition of several NSRP processes could lead to large numbers of parameters. Hence, some of the estimates obtained for $NSRP_1$ in 5 above are also used in the superposed process $NSRP_2$. (i.e. a range of subsets of $\mathbb{P}_1$ need to be considered for inclusion in $\mathbb{P}_2$ to avoid over-parametrization.)

   For the Auckland data, the following parameter set was used in the fitting procedure for each calendar month: $\{\lambda_1, \mu_{C_1}, \beta_1, \eta_1, \alpha_1, \theta_1, \lambda_2, \mu_{C_2}, \eta_2\}$, which has $\beta_2 = \beta_1$, $\alpha_2 = \alpha_1$, $\theta_2 = \theta_1$ in the fitted $SNSRP(2)$ model. For the Auckland data, this resulted in an improved fit to the sample properties as the minimum $SS(2) = 1.24 < 3.26 = SS(1)$ obtained in 4 above. (This was the only 'objective' criteria used to compare different fits, and there is clearly scope for further improvements here which account for the number of parameters in the fitted models, for example.)

4 Fitted Model

The parameter estimates for the Auckland data are given in Table 4. Discussion is focused on January and July as representative of summer and winter respectively.

For January (summer), there are two storm types, one represented by an NSRP process and the other by a Poisson Rectangular Pulses (PRP) process which is the special case of $\mu_C \equiv 1$ in an

Table 2: Parameter Estimates for Auckland Data

month i	$\hat{\lambda}_{1,i}$	$\hat{\mu}_{C_{1,i}}$	$\hat{\beta}_{1,i}$	$\hat{\eta}_{1,i}$	$\hat{\alpha}_{1,i}$	$\hat{\lambda}_{2,i}$	$\hat{\mu}_{C_{2,i}}$	$\hat{\eta}_{2,i}$	$\hat{\theta}_{1,i}$
1	0.00409	14.4	0.0721	1.98	0.645	0.0173	1.0	2.43	1.86
2	0.00367	20.4	0.081	1.94	0.617	0.0185	1.0	2.01	1.65
3	0.0066	15.3	0.107	1.64	0.626	0.00608	1.0	1.24	1.24
4	0.00704	13.9	0.109	1.33	0.57	0.00457	13.2	2.42	0.954
5	0.00523	30.8	0.0947	1.13	0.433	0.0133	20.7	2.78	0.222
6	0.00485	32.2	0.104	0.755	0.391	0.0156	37.0	2.22	0.114
7	0.0125	15.2	0.163	0.837	0.457	0.0142	26.2	2.63	0.219
8	0.0229	6.24	0.145	1.09	0.65	0.00507	11.7	2.98	0.871
9	0.0197	7.56	0.0956	1.3	0.667	0.0	.	.	1.09
10	0.0158	10.1	0.0954	1.41	0.617	0.0	.	.	0.711
11	0.0151	8.14	0.109	1.54	0.641	0.0	.	.	1.12
12	0.00977	7.84	0.0889	1.75	0.659	0.000682	7.84	1.75	2.0

NSRP process. The PRP process is the more frequent of the two ($0.017 = \hat{\lambda}_{2,1} > \hat{\lambda}_{1,1} = 0.0041$), and represents isolated convective cells, whilst the NSRP process represents clusters of convective cells. When compared with the PRP process, the NSRP process has a slightly smaller value of η indicating that the clusters of cells tend to have longer lifetimes, which might correspond to statiform rain.

For July (winter), there are two storm types both with clustering, i.e. there are two superposed NSRP processes (Table 4). The more frequent (type 2) has shorter expected cell lifetimes ($\hat{\eta}_{2,7}^{-1} \approx 0.3h$) compared with the less frequent (type 1) storms ($\hat{\eta}_{1,7}^{-1} \approx 1h$) occuring in July.

Comparing the estimates for January with those for July, we note that in January there are fewer but more intense cells, because $\hat{\lambda}_{j,1}\hat{\mu}_{c_{j,1}} < \hat{\lambda}_{j,7}\hat{\mu}_{c_{j,7}}$ and $\hat{\theta}_{1,1}\Gamma(1+\alpha_{1,1}^{-1}) < \hat{\theta}_{1,7}\Gamma(1+\alpha_{1,7}^{-1})$, which characterizes summer convective rainfall compared with winter stratiform rainfall.

The additional parametrization, i.e. the mixed model, was not needed from September to November. (Possibly a single NSRP model would also suffice for December, as this month had a small estimate of λ_2).

The fitted and sample properties are plotted in Figures 9 – 13. There is some slight underestimation of $1h$ skewness and overestimation of $6h$ skewness over the summer months (see Figure 12), which may have some effect on summer extreme values. However, overall the model fits the data well and the results provide support for the use of superposed processes.

5 Validation of Fitted Model

To validate the model it is appropriate to simulate data using the fitted model and to compare simulated and historical properties that were not used to fit the model but which are likely to be important in hydrological applications. Hence, five records of length equal to the historical record length (33-years) were simulated using the fitted model (Table 4). (Five records were used to give an indication of sampling variability in the simulated properties.)

Neither the proportion of dry intervals or the extreme values were used in fitting the model. As both of these may be important in hydrological applications, they were selected for model validation. In addition, properties taken over 12-hour sampling intervals were not used to fit the model, so this sampling interval was chosen for model validation.

Hence, to validate the model the proportion of dry intervals for 12- and 24-hour time scales were evaluated for each month and plotted (Figure 14). A bound of $1mm$ was used to define a 'dry' day (see also (2) for a further discussion on dry bounds in a similar context). In addition, the ordered annual maximum rainfalls for 1-, 6-, 12-, and 24-hour sampling intervals were calculated

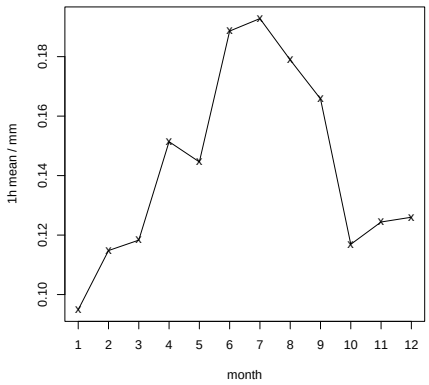

Figure 9: Mean rainfall for 1-hour data: $\times$ Sample estimate from historical data and fitted value (exact fit to sample estimate).

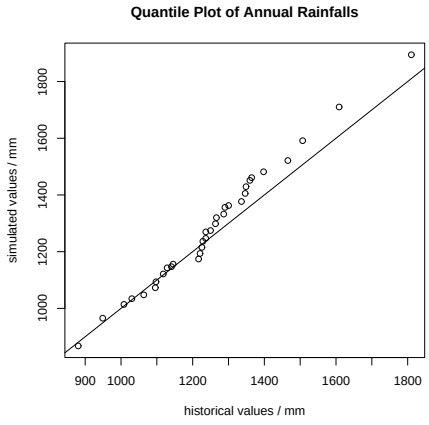

Figure 10: Comparison of distribution of historical and simulated annual totals (mm); Kolmogorov-Smirnov test has $D = 0.15$ and p-value 0.52 showing there is insufficient statistical evidence to reject the null hypothesis that the distributions are the same.

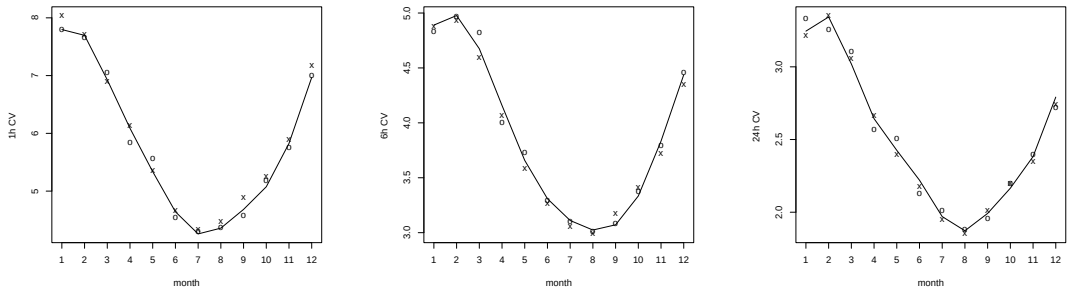

Figure 11: Coefficient of variation over a 1, 6, and 24-hour sampling intervals: — Harmonic Estimate; ∘ Sample Estimate (unsmoothed); × Fitted Value.

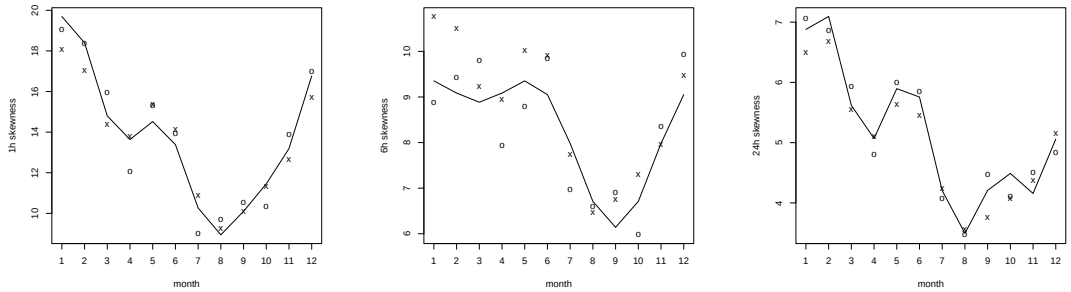

Figure 12: Coefficient of skewness over a 1, 6 and 24-hour sampling intervals: — Harmonic Estimate; ∘ Sample Estimate (unsmoothed); × Fitted Value.

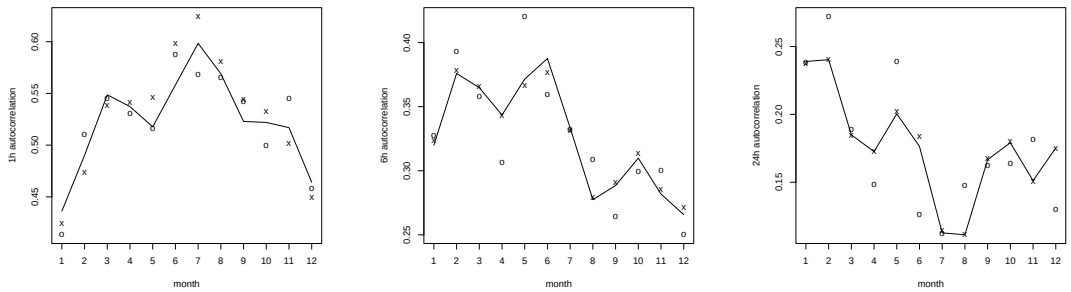

Figure 13: Autocorrelation (lag 1) over a 1-hour sampling intervals: — Harmonic Estimate; ∘ Sample Estimate (unsmoothed); × Fitted Value.

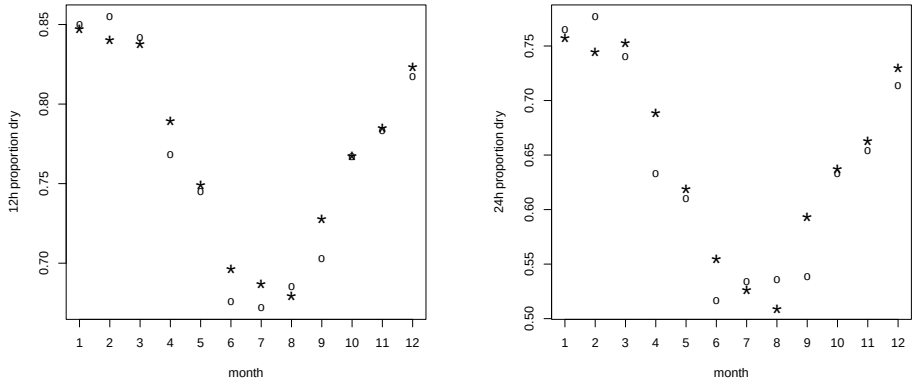

Figure 14: Proportion of dry rainfall ($< 1mm$) over 12 and 24-hour sampling intervals: $\circ$ Sample estimate from historical data (33-years); $\star$ Sample estimate from simulated data (5×33-years).

and plotted against the reduced Gumbel variate (the return period T was also plotted; Figures 15–16).

The simulated data show a very good agreement to the proportion of dry intervals even though this is not used to fit the model (Figure 14). The fit to the distribution of annual totals is good, a Kolmogorov-Smirnov test showing no significant difference between the historical and simulated values (Figure 10). The fit to the extreme values is also satisfactory (Figures 15–16). However, the fit to the upper part of the distribution tail (i.e. the very extreme values) is very good, because the historical values all fall within the range of the 5 simulated values (Figures 15 and 16). However, some discrepancies are evident in the lower part of the distribution tail for the 1-hour sampling intervals (Figure 15), possibly due to the slight underestimation of 1-hour skewness over the summer months (Figure 12).

6 Conclusions

A stochastic modelling and fitting procedure based on the superposition of independent NSRP processes was developed and extends the method described in (3). The mixed model gives further flexibility in the parametrization thus providing a methodology for obtaining good fits to a wider range of data.

Overall, the fit to sample properties up to third order over a range of time scales was good. In addition, the simulated data generally showed good agreement to the historical proportion of dry intervals and extreme values, especially to the upper tail of the extreme value distribution. The results support the use of the extended model in hydrological applications, and, consequently, the method may be used to supplement the methodology described in the previous publication.

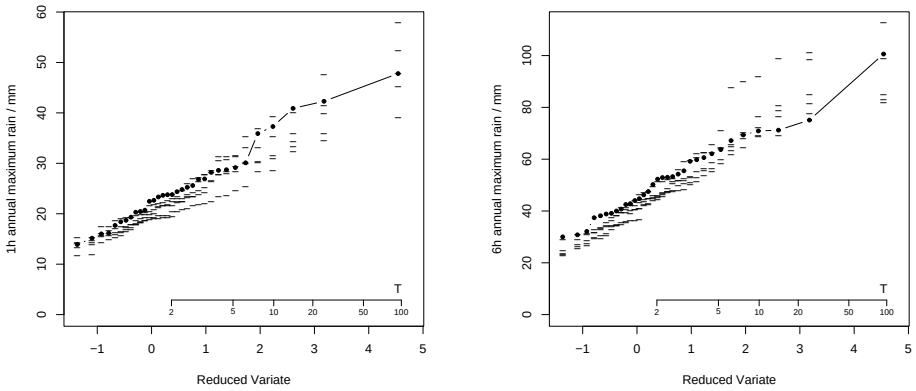

Figure 15: Gumbel probability plot for annual maximum rainfall over 1 and 6-hour sampling intervals: –●– Historical value; - Simulated values.

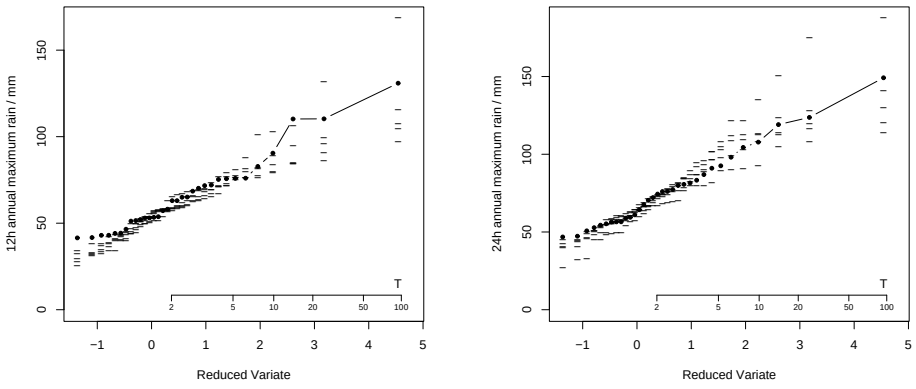

Figure 16: Gumbel probability plot for annual maximum rainfall over 12 and 24-hour sampling intervals: –●– Historical value; - Simulated values.

References

[1] Cowpertwait, P. (1994). A generalized point process model of rainfall. *Proc. R. Soc. Lond. A*, 447.

[2] Cowpertwait, P. (1998). A poisson-cluster model of rainfall: High order moments and extreme values. *Proc. R. Soc. Lond. A*, 454.

[3] Cowpertwait, P., Kilsby, C., and O'Connell, P. (2002). A space-time model of rainfall: Empirical analysis of extremes. *Water Resources Research*, 38(8):1–14.

[4] Cox, D. and Isham, V. (1980). *Point Processes*. Chapman & Hall.

[5] Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer, 4 edition.