Parsimonious modelling of winter season rainfall incorporating reanalysis climatological data

Andrew P. Garthwaite and N. I. Ramesh

ABSTRACT

Several Markov Modulated Poisson Process (MMPP) models are developed to describe winter season rainfall with parsimonious parameter use. We propose a methodology for determining the best form of seasonal model for fine-scale rainfall within a MMPP framework. Of those proposed here, a model with a fixed transition rate is shown to be superior over the other MMPP models considered. The model is expanded to include covariate data for sea-level air pressure, relative humidity, and temperature using reanalysis data over 14 years from the coordinates covering the Bracknell rainfall collection site in England. Results are compared using the likelihood ratio test and the second-order properties of aggregated rainfall.

Key words | bucket-tip, covariates, Cox model, fine-scale modelling, seasonal rainfall

Andrew P. Garthwaite (corresponding author)
N. I. Ramesh
Department of Mathematical Sciences,
University of Greenwich,
UK
E-mail: *a.p.garthwaite@greenwich.ac.uk*

INTRODUCTION

Beginning with a time series of precipitation arrivals where a 'tipping bucket' rain-gauge has accumulated and then discharged a small fixed volume of water, point process models can be used to describe rainfall at a fine time-scale and retain clustering properties relevant to this scale. The strength of this approach is that the inter-event durations between bucket tip times are observed from the time series along with the order of event arrival, allowing modelling of rainfall at scales as small as 5-minute intervals, in contrast to research where rainfall data is aggregated into hourly or daily volume and examined for first and second moment properties (Stern & Coe 1984; Hughes & Guttorp 1994; Kigobe *et al.* 2011).

Measuring these events as the accumulation of a very small volume of precipitation, the point process model posits that rainfall arrival is Poisson distributed, and better modelling can be performed by allowing the mean of this distribution to vary in different states of a dynamic system, where the transition between states is governed by a Markov chain. This yields a doubly stochastic Poisson process (Cox 1955). A useful form of this model is the Markov modulated Poisson process (MMPP), which assumes that the variation in the mean is controlled by a finite-state hidden Markov chain (Davison & Ramesh 1993; Ramesh 1995; Rydén 1996). An aptly titled comprehensive review of the model can be found in Fischer & Meier-Hellstern (1993). Unlike many Poisson cluster process models, the MMPP has a likelihood function that can be expressed in a tractable form, allowing for robust parameter estimation, albeit from a computationally demanding optimization process. Successful estimation and reproduction by simulation was demonstrated by Ramesh (1995), following a related contribution by Smith to the discussion of Stern & Coe (1984). As well as enabling parameters to be estimated through maximum likelihood optimisation, these results allowed comparison of nested sub-models through likelihood ratio tests.

doi: 10.2166/nh.2018.012

While aggregated hourly and daily rainfall have been effectively modelled through MMPP (see, for example, Ramesh 1998), the task of interest in this paper is the fine-scale modelling of winter rainfall, with comparison drawn by reference to aggregations between 5 and 60 minutes. For a great many hydrological applications, aggregations as coarse as hourly or daily measurements of precipitation are sufficient, but for certain catchment studies, including the modeling of urban drainage systems and storm water sewerage systems, and small catchment hydrology (Onof 2002), finer-scale modeling of rainfall is required. For modelling on coarser scales, Smith & Karr (1983) used a method similar to an MMPP to model inter-arrival times of summer rainfall, and extended their work to make statistical inferences about the model parameters (Smith & Karr 1985). Onof *et al.* (2002) considered a class of MMPP for the fine-scale modelling of the structure of the rainfall intensity distribution, using tipping-times of rainfall gauges. Cowpertwait *et al.* (2007) developed a Bartlett–Lewis pulse model that could also capture the fine time-scale properties of rainfall. More recently, Ramesh *et al.* (2013) derived second order properties of the aggregated rainfall from an MMPP and demonstrated that the model was capable of reproducing rainfall properties at sub-hourly resolutions. Thayakaran & Ramesh (2013) extended the model to analyse tipping-times recorded at multiple sites in a catchment area and Ramesh *et al.* (2013) further extended the model to also incorporate covariates.

The number of states to include in an optimal MMPP model has been the subject of some study. The MMPP has an unobserved underlying Markov process; the state of the system at any time can never be directly measured, but is instead inferred from the proximity of bucket-tips within the data. Ramesh (1998) used a model with two states; one state corresponding to high rainfall intensity and the other to low or no rainfall. Models with four states have also been fitted (Ramesh *et al.* 2012), but using three states is more common (Onof *et al.* 2002; Thayakaran & Ramesh 2013). Although a BIC analysis can be used to determine the best model, the three state models are chosen based on the improvement recorded in reproducing the statistical properties studied. The three-state model generally provides a strong foundation to obtain good description of the rainfall pattern, provided the effects of seasonality are avoided (Ramesh *et al.* 2013).

Onof & Wheater (1994) extended the Bartlett–Lewis method to include a random cell duration, and improved the reproduction of the proportion of dry periods of different duration. In earlier research (Onof & Wheater 1993), simulation studies demonstrated an improvement in the temporal characteristics where the optimization had included some value for the cell arrival rate, with the resultant arrival rates recognizably characteristic for each individual month.

Variation in rainfall patterns across the year and seasonal effects have been countered by modelling each calendar month separately (Ramesh *et al.* 2013; Thayakaran & Ramesh 2013), with each month modelled by a unique set of parameters. However, more parsimonious modelling is desirable if the rainfall model is to form part of a larger climate model involving many separate meteorological processes. In this paper, we formulate models with a reduced total number of parameters for the winter season rainfall. This is achieved by treating the four-month winter period from the start of November to the end of February as a single block of time, and taking advantage of similarities between the patterns of rainfall events to model this period with one fitting. Differing approaches to retain variation between these months are examined, and the results analysed by likelihood ratio tests as well as graphical summaries of simulation studies.

This investigation is by nature a comparison of various models and we describe four MMPP models, each having three states. The first two models are based on standard research: first, we take the MMPP model fitted to each month individually and, second, we consider the MMPP model fitted over the winter season as a whole, with no effort made to model variation between the winter months. We refer to the former as the *maximal model*, as it contains all the other models we consider as special cases, and the latter as the *Fixed Parameter (FP) model*. We go on to detail two original alternatives that we call the *Fixed Transition Rate (FTR) model* and the *Fixed Arrival Rate (FAR) model*. The purpose of introducing a winter seasonality is to model the season as a whole, and then to experiment with different approaches to introduce some variation from November through to February. This contrasts with other research, where the practice has mostly been to model each

calendar month separately. The imperative is to compare the new models with both the maximal model and null (FP) model.

The proposed models are used to analyse winter season rainfall data from England. Additional models that incorporate covariates to produce improved fits are considered. The capacity for our models to simulate extreme events is briefly explored before the conclusions are summarised.

DATA AND STUDY AREA

The rainfall bucket tip-time data used in this investigation came from a weather station in Bracknell, England over a fifteen-year period that included fourteen complete winters, made available by the Centre for Environmental Data Analysis. Over this period, times were recorded when a fixed volume of precipitation, 0.2 mm, had collected in the recording device, forcing the bucket to tip and discharge its cargo. The mean hourly rainfall across the four winter months is 0.086 mm of precipitation per hour, with a standard deviation of 0.392. About two-thirds of the days over the period are wet days with some rain during the winter season and the other third of them are dry days with no rain. The frequent occurrences of dry days in the data set suggests that a viable model would include a state of negligible or no rainfall. The distribution of the accumulated rainfall, both at hourly and daily scales, looks positively skewed. The maximum daily rainfall over the period is recorded as 31.2 mm. When considering meteorological covariates to include in our covariate models, we included temperature, relative humidity, and sea-level air-pressure. Values for the covariates are available from the data supplied by the United States National Oceanic and Atmospheric Administration. These data are reanalysis data, obtained by using a consistent modern analysis system wherein observational data from a historical period is reprocessed. Figure 1 shows the daily variation of the three covariates using box plots drawn separately for the four months. There appears to be little variation in their values across the winter months, except for slightly higher values for temperature in November when compared with other months.

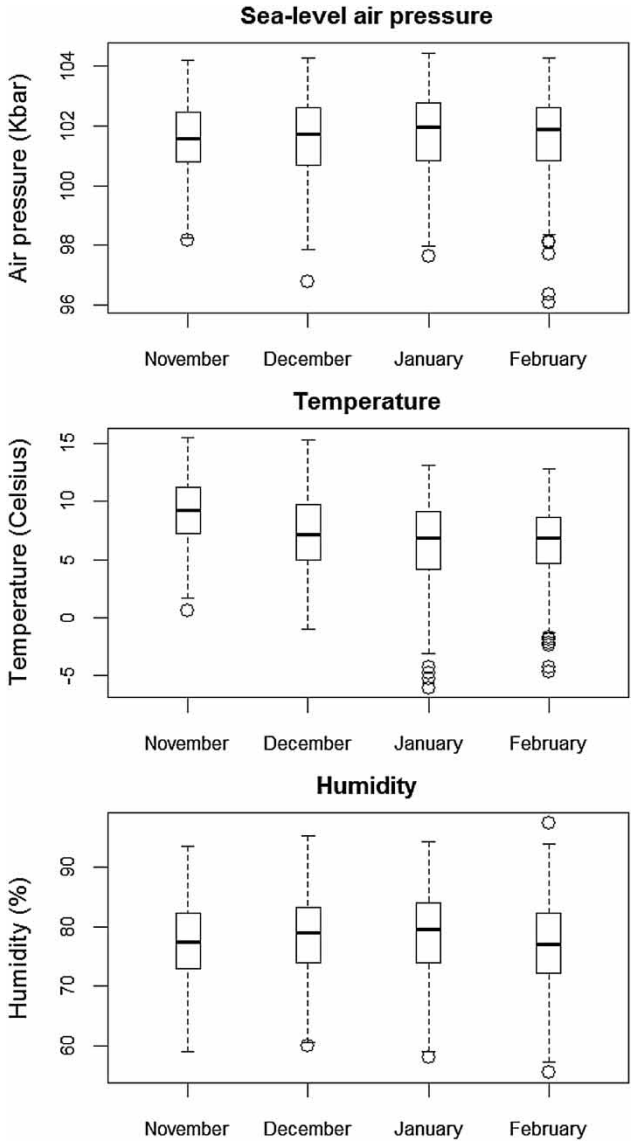

Figure 1 | Box plots of daily observations of temperature, sea-level air pressure, and humidity for the Bracknell site over fourteen years across the four winter months.

METHODS

The general MMPP model

We assume the process that controls the arrival of points is a stationary irreducible Markov chain, $\{X(t)\}$, with k states, labelled $1, 2, \ldots, k$, and Q is its infinitesimal generator. The rates of transition from one state to another are determined by the off-diagonal elements of the $k \times k$ matrix Q,

whose diagonal element for each row equals the negative of the sum of the remaining elements in that row. The mean sojourn time for state i (the average amount of time spent in that state) is $-1/q_{ii} = 1/\sum_{j\neq i} q_{ij}$, for $i = 1, \ldots, k$.

Let π_i denote the probability that the process is in state i when it is in equilibrium and take the steady-state probability distribution π equal to $(\pi_1, \pi_2, \ldots, \pi_k)$. We assume that the underlying Markov chain $\{X(t)\}$ is initially in equilibrium and let the point process $\{N(t)\}$ denote the number of bucket-tips. The rate that bucket-tips occur is dependent on the current state. Let ϕ_i be the mean rate of bucket-tips when $\{X(t)\}$ is in state i and assume that $\{N(t)\}$ is a Poisson process of rate $\phi_{X(t)}$. The arrival rate matrix, L, is the $k \times k$ diagonal matrix whose (i, i) element is ϕ_i for $i = 1, \ldots, k$.

We focus on the case $k = 3$, where:

$$L = \begin{bmatrix} \phi_1 & 0 & 0 \\ 0 & \phi_2 & 0 \\ 0 & 0 & \phi_3 \end{bmatrix} = \mathrm{diag}[\phi_1, \phi_2, \phi_3] \tag{1}$$

and

$$Q = \begin{bmatrix} -(q_{12}+q_{13}) & q_{12} & q_{13} \\ q_{21} & -(q_{21}+q_{23}) & q_{23} \\ q_{31} & q_{32} & -(q_{31}+q_{32}) \end{bmatrix}, \tag{2}$$

where q_{ij} is the transition rate from state i to state j.

Suppose that the process is observed in the interval $[0, T]$, wherein bucket tips occur at times $t_1 < t_2 < \cdots < t_n$, so $N(T) = n$. To obtain an expression for the likelihood function of the point process $\{N(t)\}$, following from Smith (1984), we define the following conditional probabilities:

$$\psi_{ij}(t) = P\{X(t) = j, N(t) = 0 | X(0) = i, N(0) = 0\},$$
$$i, j = 1, \ldots, k.$$

Here $\psi_{ij}(t)$ is the probability that the underlying Markov chain has transitioned to state j at time t, given that it began in state i at time 0, and that in the time period between 0 and t there were no bucket tips. The Chapman–Kolmogorov forward differential equations for the process are (Ross 1983):

$$\psi_{ij}(t+\delta t) = \psi_{ij}(t)(1 + q_{jj}\delta t)(1 - \lambda_j \delta t) + \sum_{k\neq j} \psi_{ik}(t) q_{kj} \delta t + o(\delta t).$$

These lead to:

$$\Psi(t) = \exp\{(Q - L)t\} = \sum_{n=0}^{\infty} \frac{t^n (Q - L)^n}{n!}, \tag{3}$$

where $\Psi(t)$ is the matrix function with entries $\psi_{ij}(t)$. See, for example, Ramesh *et al.* (2013) for a derivation of Equation (3).

Given $t_1, \ldots, t_n$, the likelihood function is obtained from:

- the steady-state probability, $\pi_l\, (l = 1, \ldots, k)$, that the process is in state l at time $t = 0$. For a three state model (where $l = 1, 2, 3$), the three elements of the steady state probability vector, π, are here described as:

$$\pi_1 = (1 - \pi_2 - \pi_3)$$
$$\pi_2 = \frac{(q_{12}+q_{13})(q_{32}-q_{12}) + q_{12}(q_{31}+q_{12}+q_{13})}{(q_{21}+q_{12}+q_{13})(q_{32}-q_{12}) + (q_{31}+q_{12}+q_{13})(q_{21}+q_{23}+q_{12})}$$
$$\pi_3 = \frac{(q_{21}+q_{23}+q_{12})(\pi_2) - q_{12}}{(q_{32}-q_{12})}; \tag{4}$$

- the probability of not observing any bucket tips of rainfall before the first bucket tip $[\Psi(t_1)]$;
- the (conditionally independent) probabilities of not observing any bucket tips of rainfall between those observed $[\Psi(t_i - t_{i-1})$ for the interval from t_{i-1} to t_i, for $i = 2, \ldots, n]$;
- the probability of observing bucket tips of rainfall at the times they occurred, using the rates ϕ_j that form the matrix L;
- the probability of not observing any bucket tips of rainfall between the last observation and the end of the recording period $[\Psi(T - t_n)]$.

As given in Smith (1984), Ramesh (1995, 1998) and Ramesh *et al.* (2013), the likelihood is:

$$f(t_1, \ldots, t_n | Q, L) = \pi \left[\prod_{i=1}^{n} \{\Psi(t_i - t_{i-1}) L\} \right] \Psi(T - t_n)\, l, \tag{5}$$

where multiplication by the vector l sums the products over all possible states l.

Equations (3) and (5) form the likelihood function of the MMPP model. The parameters to estimate are the components of L and Q. A model with k states will have k^2 parameters. While Equations (3) and (5) express the likelihood in closed form, determining its value for a set of parameters is a sequential computation where the likelihood is updated at each event in the time series. Finding a maximum likelihood estimate is possible numerically, although time-consuming, and its duration increases with the number of parameters.

The FP and maximal models

The maximal model effectively fits a separate MMPP for each of the four winter months. With six transition rate parameters and three arrival rate parameters for each of the four months, this is the largest of our models with thirty-six parameters in total.

Rather than fit an MMPP model to each month individually, the FP model fits the MMPP model once for the whole winter data, with no difference between months. Thus, its arrival rate matrix and transition rate matrix are given by Equations (1) and (2), and its likelihood is given in Equations (3) and (4). The only distinction is in the choice of data it is applied to, being whole seasons concatenated into a single time series. For this model, there are nine total parameters for the whole four-month period, with six transition rate parameters and three arrival rate parameters, that do not change from month to month. Although we refer to the parameters as 'fixed', they are not fixed in the sense that the parameter values are fixed for the optimisation, simply that one set of parameter estimates is used for the whole season without variation between months.

The FTR model

In the FTR model the parameter values for the arrival rate matrix are allowed to vary as normal across months, but the transition rate matrix is held in common for all the months. This method of abridging the model for the winter season takes the structure of the FP model for the transition matrix Q and assumes that ϕ_1, the arrival rate parameter in state 1 (low to minimal rainfall), is the same in all four months. Thus, its only differences from the FP model is that the arrival rates in states 2 and 3 (ϕ_2 and ϕ_3) vary with month. Again, although in this method we refer to the transition rate matrix as 'fixed', it is not fixed in the sense that the parameter values are fixed for the optimisation.

To aid parameter interpretation, arrival rates are specified as the rates in November with additional adjustment parameters in subsequent months, like corner-point parameterization in general linear models. The following defines the notation, where L_j is the arrival rate matrix for month j, where $j = 1$ corresponds to the month November, and j increases correspondingly from the months of December to February:

$$\begin{aligned} L_1 &= \mathrm{diag}(\phi_1, \phi_2, \phi_3) \quad \text{for } j = 1, \\ L_j &= \mathrm{diag}(\phi_1, \phi_2 + \beta_{2j}, \phi_3 + \beta_{3j}) \quad \text{for } j = 2, 3, 4. \end{aligned} \tag{6}$$

The parameters ϕ_1, ϕ_2 and ϕ_3 may be viewed as the 'baseline' arrival rate parameters, while the six β_{ij} parameters are the monthly arrival rate adjustment parameters. The transition matrix Q does not vary from month to month.

In the FTR model, optimisation occurs over a single set of six transition rate parameters for the whole season, along with three baseline arrival rate parameters, and a further six arrival rate adjustment parameters (for states 2 and 3 across December, January and February) making a total of 15 parameters.

The changes in ϕ_i affect Ψ (Equation (3)), this becoming a function of the time interval and the month the points occur within, so at time t:

$$\Psi(t) = \exp\{(Q - L_{j_t})t\} = \sum_{n=0}^{\infty} \frac{t^n (Q - L_{j_t})^n}{n!}, \tag{7}$$

for $j_t = 1, 2, 3, 4$. This represents the use of the arrival rates related to the month at the time of the event. In addition, when we post-multiply $\Psi(t, j_{t_i})$ by the arrival matrix, again we must use the L matrix corresponding to the relevant month for the recorded events:

$$f(t_1, \ldots, t_n | Q, L) = \pi \left[\prod_{i=1}^{n} \{\Psi(t_i - t_{i-1}) L_{j_{t_i}}\} \right] \Psi(T - t_n) l, \tag{8}$$

for $j_t = 1, 2, 3, 4$.

The FAR model

The FAR approach to abridging the model across the winter months took the reverse approach to the FTR model. It assumes a single arrival rate matrix for the whole season, while the transition rate matrix is allowed to vary gradually between months. In this model, there are separate transition rate adjustment parameters $\alpha_{i,k,j}$ for the months December, January and February, acting on each of the six off-diagonal transition rate parameters. With 27 parameters in total, this model is larger than the FTR model, but it is still noticeably smaller than the maximal model that has 36 parameters.

Let Q_j denote the transition rate matrix for month $j (j = 1, \ldots, 4)$. With November as a baseline, Q_1 is equal to Q in Equation (2), and the other transition rate matrices can be written as:

$$Q_j = \begin{bmatrix} -q_{12} - \alpha_{12j} - q_{13} - \alpha_{13j} & q_{12} + \alpha_{12j} & q_{13} + \alpha_{13j} \\ q_{21} + \alpha_{21j} & -q_{21} - \alpha_{21j} - q_{23} - \alpha_{23j} & q_{23} + \alpha_{23j} \\ q_{31} + \alpha_{31j} & q_{32} + \alpha_{32j} & -q_{31} - \alpha_{31j} - q_{32} - \alpha_{32j} \end{bmatrix},$$

for $j = 2, 3, 4$. The arrival rate matrix, L, has the same form as in Equation (1). The role of the α_{ijk} parameters is to adjust the rate of transition from state i to state j, when in month k, adjusted from a baseline established during the first month in the series, so as to model the gradual variation in transition rates from month to month.

Adapting for covariate model

Often meteorological covariate information only gives a daily value for each covariate, and for this example data availability dictates that this is the case here. To include covariates in the MMPP model, the approach we adopt is to allow them to influence the rainfall arrival rates in the L matrix. Ramesh *et al.* (2013) described a model where time varying covariates influenced the transition and arrival matrix, and we make slight modifications to the expression of conditional probabilities previously suggested, as the rate matrix now varies with time. Again, we define the matrix $B(u) = (Q - L(u))$, where the arrival matrix L in B (u) is set to vary with time, while the transition matrix Q remains constant:

$$\Psi(t) = \exp\left\{\frac{1}{t}\int_0^t [(Q - L(u)]du\right\} = \exp\left\{\frac{1}{t}\int_0^t B(u)du\right\} = \exp \bar{B}t = e^{\bar{B}t} \quad (9)$$

The likelihood function is then written as,

$$f(t_1, \ldots, t_n, |Q, L) = \pi\left[\prod_{i=1}^{n}\{\exp\{\bar{B}(t_i - t_{i-1})\}L(t_i)\}\right] \bar{B}(T - t_n)\mathbf{1}. \quad (10)$$

with $\bar{B}(t_i - t_{i-1}) = 1/(t_i - t_{i-1}) \int_{t_i-1}^{t_i} [Q - L(u)]du$, and $t_0 = 0$.

The arrival rate matrix is allowed to depend on meteorological covariates, and we adjust the L matrix (c.f. Equation (6)) accordingly, where β_{2j} and β_{3j} vary with month, $\boldsymbol{x}$ is the 3×1 vector giving the daily values of the covariates, and $\boldsymbol{\gamma}$ is a vector of regression coefficients that does not vary with month. The monthly adjustments β_{21} and β_{31} applied to November are fixed parameters, both equalling zero, as November arrival rates are treated as baseline estimates for the FTR model. The lowest state of arrival has no adjustment parameter as the arrival rate in state one is treated as zero or of a negligible rate. The function $L(u)$, of the vector of daily meteorological covariate data $\boldsymbol{x}$ and month j, produces the daily arrival matrices, as follows:

$$L(t_i) = \begin{bmatrix} \phi_1 & 0 & 0 \\ 0 & \phi_2 + \beta_{2j} + \boldsymbol{\gamma}^T \boldsymbol{x_i} & 0 \\ 0 & 0 & \phi_3 + \beta_{3j} + \boldsymbol{\gamma}^T \boldsymbol{x_i} \end{bmatrix}, \quad (11)$$

where the tip times t_i fall in month j.

This model is very flexible as there are no restrictions on the transition and arrival matrices, except that they must meet the conditions to be the transition rate and arrival rate matrices of an MMPP model.

RESULTS AND DISCUSSION

The models were employed to analyse a data set that gave the times of bucket-tips at a weather station in Bracknell, England over a fifteen-year period that included fourteen complete winters. The readings from the latter were accrued into a single vector of arrival times, together with a labelling vector to denote the month (November–February) wherein each rainfall event originated. For the maximal model, the data for each winter month were taken separately and an MMPP model fitted; summing their objective function values gave the log-likelihood for the maximal model.

All four models were fitted to the data and the parameter values for the maximal, FP, FTR and FAR models are given in Table 1 along with the fitted sojourn times for each state. In the table, q^*_{ik} is the (i, k) element of the transition rate matrix for the month specified in the first column; the ϕ^*_i are the diagonal elements of the arrival rate matrix for the specified month. The sojourn times are defined similarly. For the FP model, these quantities each have a single value for all the winter months. The parameter estimates for the maximal model confirm a similarity between months in terms of their Markov state transition rates.

The parameters of the transition rate matrix (the q^*_{ik}) cannot vary with month under the FTR model, and Table 1 shows that their values are similar to the FP model, as might be expected. Similarly, the parameters of the arrival rate matrix (the ϕ^*_i) cannot vary with month under the FAR model, and the parameter estimates of this are similar to those of the FP model. The parameter values that vary from month to month are the ϕ^*_i with the FTR model, and the q^*_{ik} with the FAR model. The values they take are quite close to the values given in corresponding months by the maximal model, as can be seen by comparing their values with the first four rows of Table 1.

Turning to hypothesis tests, parameters have been estimated by maximum likelihood so a natural criterion for model comparison is the likelihood ratio test. Table 2 gives values of likelihood ratio test statistics, parameter difference, and p-values for the other models (FP, FTR, and FAR) when compared with the maximal model. Each of these models is a simplified form of the maximal model and the first question of interest is whether there is evidence

Table 1 | Parameter estimates and sojourn times for the four models used

Month	q^*_{12}	q^*_{13}	q^*_{21}	q^*_{23}	q^*_{31}	q^*_{32}	ϕ^*_1	ϕ^*_2	ϕ^*_3	sj1*	sj2*	sj3*
Maximal model												
Nov	0.03	0.002	0.37	0.18	0.07	0.72	0.03	2.47	14.02	33.33	1.80	1.27
Dec	0.03	0.003	0.35	0.20	0.18	0.70	0.03	2.44	13.37	32.42	1.83	1.14
Jan	0.03	0.004	0.39	0.14	0.17	0.95	0.03	2.98	16.88	29.87	1.91	0.89
Feb	0.02	0.004	0.33	0.13	0.19	0.67	0.02	2.14	13.06	35.78	2.16	1.16
FP model												
Win.	0.03	0.003	0.36	0.16	0.14	0.83	0.03	2.62	15.02	33.20	1.89	1.03
FTR model												
Nov	0.03	0.003	0.36	0.16	0.15	0.82	0.03	2.66	15.94	33.13	1.89	1.03
Dec	0.03	0.003	0.36	0.16	0.15	0.82	0.03	2.67	14.35	33.13	1.89	1.03
Jan	0.03	0.003	0.36	0.16	0.15	0.82	0.03	2.86	16.34	33.13	1.89	1.03
Feb	0.03	0.003	0.36	0.16	0.15	0.82	0.03	2.29	13.45	33.13	1.89	1.03
FAR model												
Nov	0.03	0.003	0.38	0.17	0.07	0.80	0.03	2.61	14.97	33.58	1.82	1.15
Dec	0.03	0.002	0.34	0.20	0.20	0.78	0.03	2.61	14.97	35.06	1.84	1.02
Jan	0.03	0.005	0.39	0.16	0.13	0.97	0.03	2.61	14.97	29.34	1.83	0.91
Feb	0.02	0.004	0.34	0.13	0.19	0.76	0.03	2.61	14.97	35.65	2.13	1.05

Table 2 | Likelihood ratio tests comparing alternative models with the maximal model (left) and FP model (right)

vs. Maximal model	No. of para.	Para. diff.	D test stat.	*p*-value	vs. FP model	No. of para.	Para. diff.	D test stat.	*p*-value
Maximal	36	–	–	–	–	–	–	–	–
FP	9	27	32.565	0.2118	FP	9	–	–	–
FTR	15	21	4.551	0.9999	FTR	15	6	28.014	<0.0001
FAR	27	9	14.343	0.1106	FAR	27	18	18.222	0.4411

of model inadequacy for any of them. To this end, likelihood ratio tests were performed to test whether the additional parameters in the maximal model improved the fit relative to a simpler model nested within it. The sample size for the tests is 17,392 as this was the number of bucket-tips over the fourteen winters, so asymptotic theory should hold well.

Results of the tests are displayed in Table 2. The null hypothesis is that the additional parameters in the maximal model do not make it a better model than the simpler model. The test statistic, D, is twice the difference between the maximum log-likelihood of the maximal model and the maximum log-likelihood for the alternative model being tested. If the null hypothesis holds, then asymptotically D follows a chi-square distribution on v degrees of freedom, where v is the difference in model size. It can be seen that when the maximal model is compared with each of the new models, in every case the null hypothesis fails to be rejected. Hence there seems little justification for having the additional parameters.

The FP model does not distinguish between calendar months, and estimates a simple set of nine parameters that are not adjusted from month to month. The FTR and FAR models can each be obtained from this model by adding parameters to it, so the FP model is nested within each of the new models. Likelihood ratio tests were conducted to examine whether the additional parameters in the FTR and FAR models gave any improvement, and these results are also given in Table 2. The null hypothesis is that the additional parameters do not improve the model. This hypothesis was not rejected when the FAR model was compared with the FP model ($p = 0.44$), so there is no demonstrable improvement in the FAR model through adjusting transition rates. Hence there seems little justification in having monthly-varying transition rate parameters. However, there is a strongly significant result when the FTR model is compared with the FP model ($p < 0.0001$), so there is clear evidence that adjusting arrival rates is an improvement to the model. It is also worth noting that the performance of the FTR model was certainly better than that of the FAR model, as it has fewer parameters but still has the higher log-likelihood ratio test statistic. Hence, all the indications are that the best of the models is the FTR model.

To examine the goodness of fit of a model, simulations were run with the model's parameter estimates treated as population parameters making use of an algorithm called event-by-event simulation as described by Ramesh (1995). An event is defined as either an arrival from the point process $N(t)$ or a state transition of the underlying Markov chain $X(t)$. The initial state of $X(t)$ is simulated from its stationary distribution. Given that the Markov chain is in state i, the next event is taken as an arrival from $N(t)$ with probability $\phi_i/(\phi_i + q_i)$ or a transition of $X(t)$ to state j with probability $q_{ij}/(\phi_i + q_i)$. The time to the next event is then obtained from an exponential distribution with parameter $(\phi_i + q_i)$. This process is continued until the final point in the interval is simulated. If the population parameters are well-estimated and the model is appropriate, then the time series of rainfall measurements that this yields should resemble the original observed data. We compare the empirical and simulated values of various statistics for rainfall aggregated at different time-scales. As we are particularly interested in fine time-scales, we aggregate in 5, 10, 20, 30 and 60-minute intervals. For each model, we repeated the simulation 100 times and formed simulation bands using the minimum and maximum values in the 100 simulations.

The following statistics of the rainfall intensity are examined:

(i) the mean volume of rainfall in an interval;
(ii) the mean duration of 'dry' periods (a 'dry' period is defined as at least two consecutive intervals without rain);

(iii) the mean duration of 'wet' periods (a 'wet' period is defined as a period of at least two consecutive intervals with rain recorded);
(iv) the coefficient of variation of the rainfall in an interval;
(v) the standard deviation of the volume of rainfall in an interval;
(vi) the proportion of dry intervals in the time series, wherein no bucket-tips are observed;
(vii) auto-correlation with a range of lags from 1 to 5.

The same process of aggregation and calculation of summary statistics was conducted with the empirical observed data set. Here we restrict attention to the maximal model and the FTR model (these are the two best competing models). For each aggregation level and each summary statistic a graphical display was created to chart the summary statistic across the four winter months. These are given in Figures 2 and 3. The solid black line towards the centre of each graph is the value of the summary statistic for the observed data. Superimposed on the graphs are two sets of simulation bands: the dotted (blue) lines are the maximum and minimum values from the 100 simulated data sets for the FTR model, while the dashed (red) lines are the equivalent boundary lines for the maximal model.

Plots in the left-hand column of Figure 2 give the mean rainfall accumulation in 5, 10, 20, 30 and 60-minute intervals and those in the right-hand column give the mean duration of dry and wet periods for a selection of time intervals. The boundary lines for the FTR model are always close to those for the maximal model, but tend to be a little flatter. This is also true of the boundary lines in Figure 3, that give plots for the other summary statistics listed above. Sampling distributions of standard deviation appears to be skewed for both models at finer aggregations. In general, values for the observed data are comfortably within the boundary lines, with the exception of autocorrelation – we present here autocorrelation at lag 5, where the observed value is outside the boundary lines on only one occasion (autocorrelation of the 20-minute accumulation at lag 5 in January for the maximal model) but is always within the boundary lines for the FTR model. The simulations failed to deliver a good confidence band for autocorrelation with lags ranging from one to four – typically the autocorrelation was underestimated at such aggregations as 5 and 10 minutes, but improved for larger aggregations, and then was well estimated at a 60-minute aggregation for most months. Hence, the plots indicate that the FTR model gives an adequate fit to the data, but that some of the second order properties of the model are less well served with finer-scale aggregations.

MMPP models with covariates

To illustrate application of this covariate model with meteorological covariates we return to the data on bucket-tips at the weather station in Bracknell. The covariates we consider are temperature, relative humidity, and air-pressure, as earlier work by Ramesh *et al.* (2013) indicated that these had a significant effect on precipitation arrival within the MMPP framework. As stated earlier, these data are reanalysis data, obtained by using a consistent modern analysis system wherein observational data from a historical period is reprocessed. We focus on adding covariates to the FTR model, as results in the earlier section found the FTR model performed better than the alternative models we considered. The model assumes that the transition rate matrix (Q) does not vary with month during the season. In an attempt to obtain a parsimonious model, we assume that it also does not vary with the covariates.

The model incorporating covariates was fitted to the time series of bucket tips using maximum likelihood estimation. When compared with the FTR model without covariates with three fewer parameters this gives a test statistic for the likelihood ratio of 83.2, and hence there is overwhelming evidence that the covariates improve the model.

The covariates are clearly good predictors of rainfall, and it seems plausible that a simpler model than the FTR model could be used when the covariate information is available. Covariates were added to the FP model and fitted to the tip times data. This model has six fewer parameters than the FTR model with covariates, with a test statistic for the likelihood ratio of 6.60. This is only slightly poorer than the FTR model with covariates, but there is significant evidence ($p = 0.04$) that the FTR model with

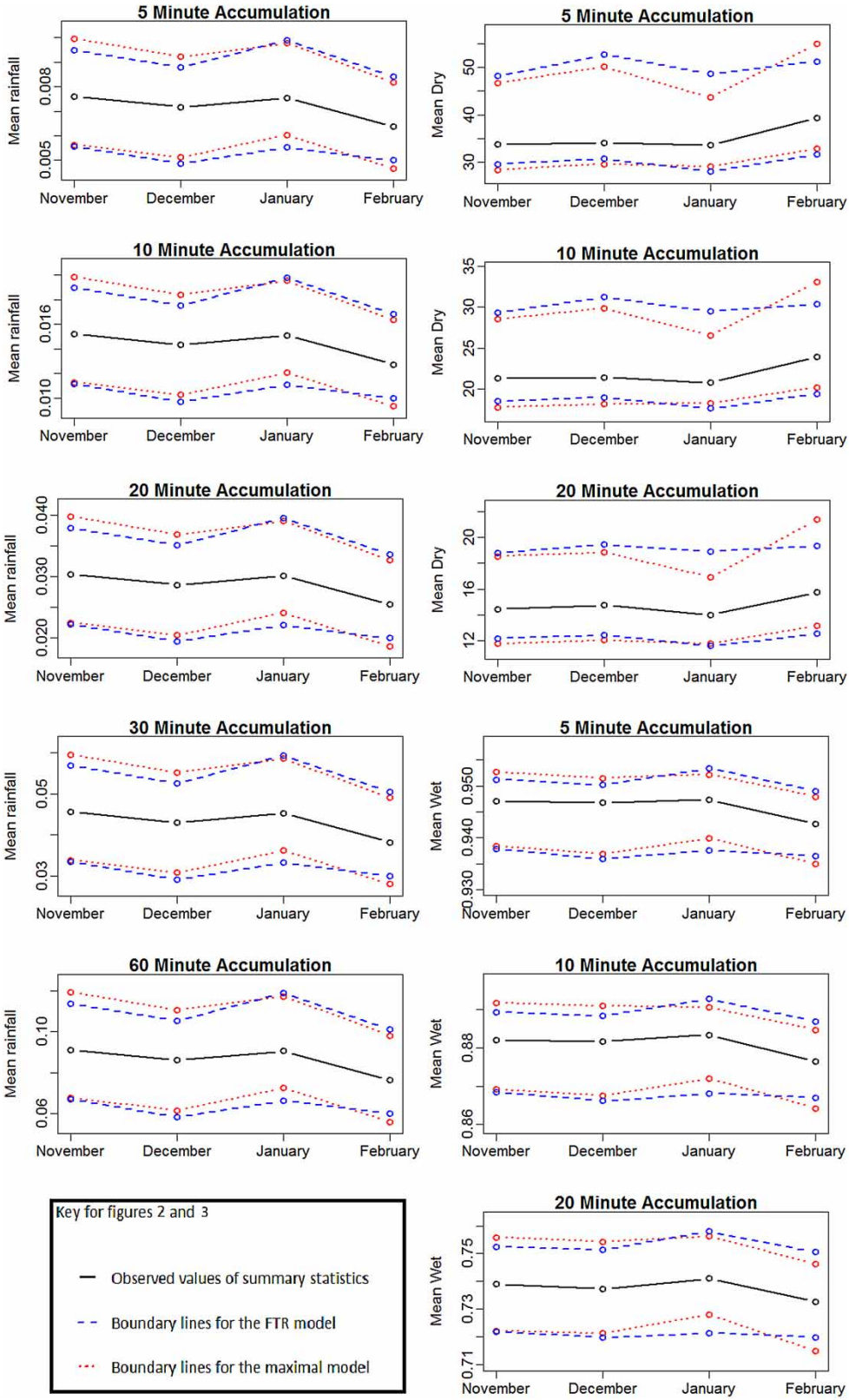

Figure 2 | Empirical mean rainfall, mean duration of wet and dry periods, with simulation bands from the FTR and maximal models.

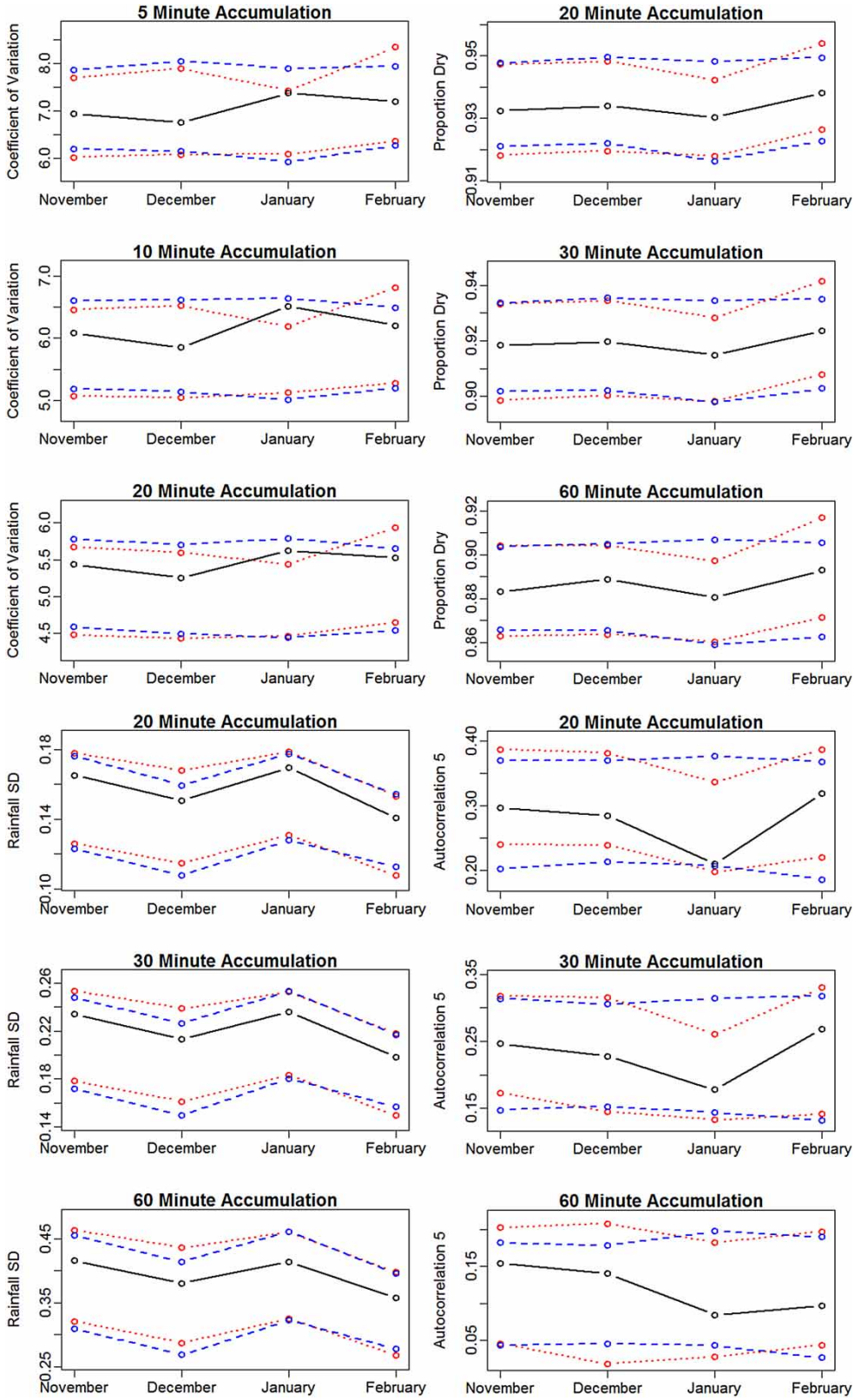

Figure 3 | Observed coefficient of variation, standard deviation, proportion of dry periods and autocorrelation at lag 5 of the aggregated rainfall with simulation bands from the FTR and maximal models.

covariates is the better model. Simulations were run to examine the goodness of fit of the models using the same procedure as before. Plots similar to Figures 2 and 3 are given for the FTR model with covariates and the FP model with covariates in Figures 4 and 5. These two models always show very similar bounds that almost always contain the observed data value, so the fits are adequate. We find that while the FTR model is superior to simpler models in cases without readily available daily meteorological covariates, where this data is available the simple covariate FP model will be quite capable of simulating rainfall series alike to the observed data. One explanation of the efficacy of this simple model, that does not see changes in the arrival rate modulated by calendar month, comes from the input of the meteorological data that varies daily. Given knowledge of the sea-level air-pressure, relative humidity and temperature, knowledge of calendar month becomes less necessary to describing appropriate rates of arrival.

Extreme events

We examined the extreme 30 minute, 60 minute, and daily rainfall volumes from simulations drawn from the FTR model, shown in Figure 6 plotted along with the empirical daily extreme rainfall against their Gumbel reduced variate (Gumbel 1954). This comparison showed that the model was capable of reproducing daily extremes, as the empirical evidence lay within the simulation boundaries provided by the boxplots for all fourteen years. At 60 minute intervals we found that the simulations are only successful for the lower stretch of the Gumbel reduced variate in producing simulations bands that include the empirical evidence. Clearly this is not as effective as with the daily extremes, and in plotting the 30 minute extremes the simulation is again less well reproduced with smaller time interval aggregations. When this was performed with the covariate FTR model, this same pattern was represented. As reported in previous studies, Verhoest *et al.* (1997) for example, the estimation of extreme values at fine time-scale is a common problem for most stochastic models for rainfall and our results reveal the same.

CONCLUSIONS

By identifying months with similar transition rates, we have been able to reduce the number of parameters used for fitting whole winter seasons and produce simulations and results comparable to that of the maximal model. This approach offers a benefit to larger environmental or hydrological models, where selecting methods of describing rainfall must be done with consideration to the overall size of a complex system incorporating a plethora of distinct parts.

For our purposes we have concentrated on the months November through to February, however there exists potential to further reduce any year-long model by collecting months with similar transition rates. While not employed here, a general guide for collecting together of months with common transition rate parameters would see as viable candidates any months where the standard deviation of the parameter estimates indicates a parameter region common to all, with transition rate estimates separated by no more than two standard deviations, when the months being examined run together consecutively.

From the reanalysis data set we employed covariate values that changed on a daily basis. The covariate information was incorporated into our overall model by the simple expedient of partitioning the time-series into days and computing the contribution to the likelihood from each day separately. This a flexible approach that can be applied in a number of different ways.

In the covariate case, we have shown statistically significant evidence for extending the covariate model from FP to FTR, but the strength of the evidence for additional parameters was of a lower level of significance than in the model without covariates. To account for this difference we consider sources of input into the arrival rates for each model. Conceptually, the difference between FTR and FP is in allowing the arrival rate matrix, in the FTR model, to contain variation between months. Given that the meteorological covariate data that accompanies both the FP and FTR covariate models contain information that acts as a predictor for rainfall arrival, and the meteorological data varies naturally across season, the covariate data itself contains much of the variation needed to effectively model rainfall arrival in each month distinctly. Equipped with knowledge

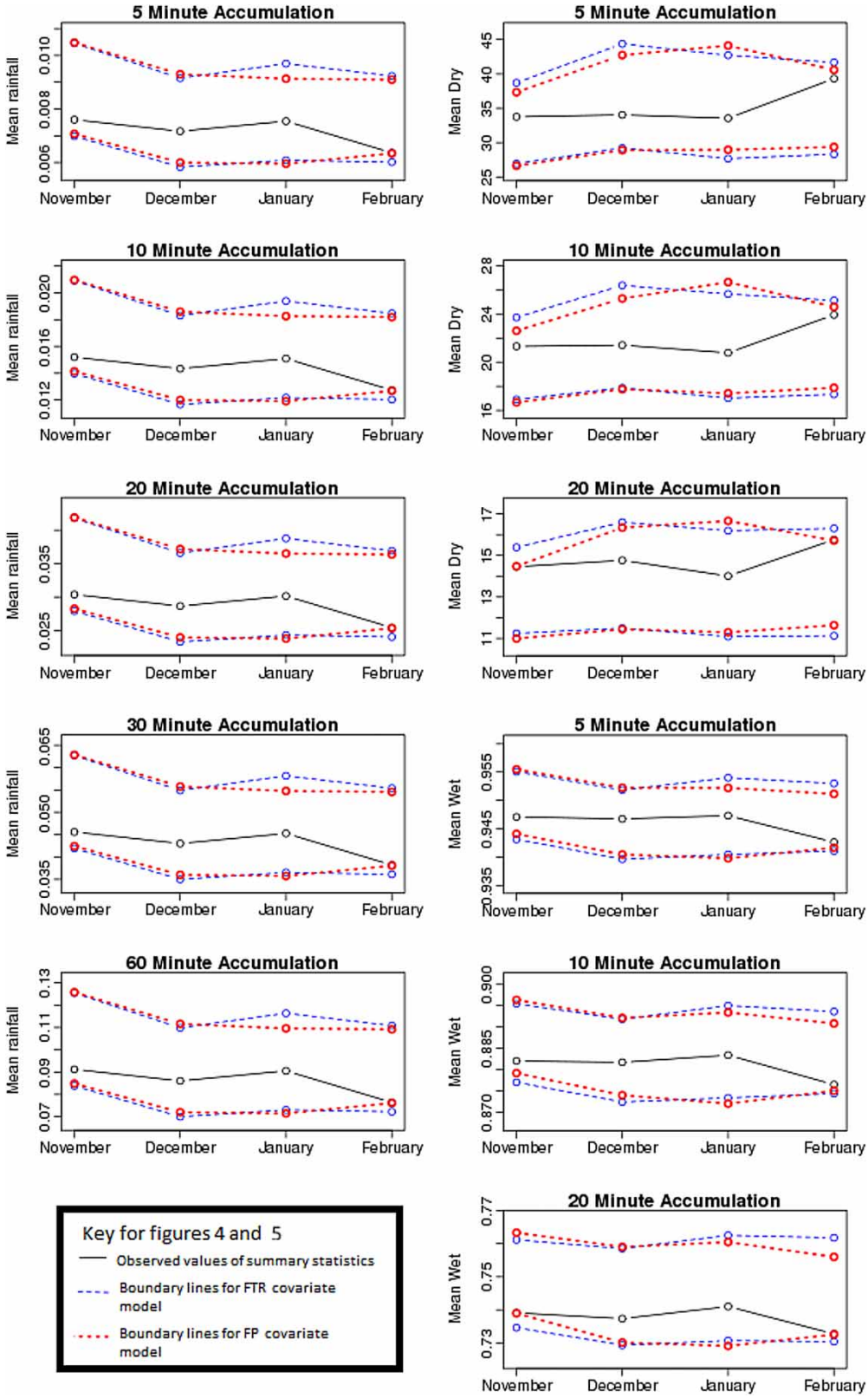

Figure 4 | Empirical mean rainfall and the mean duration of wet and dry periods, along with simulation bands from the FTR and FP covariate models.

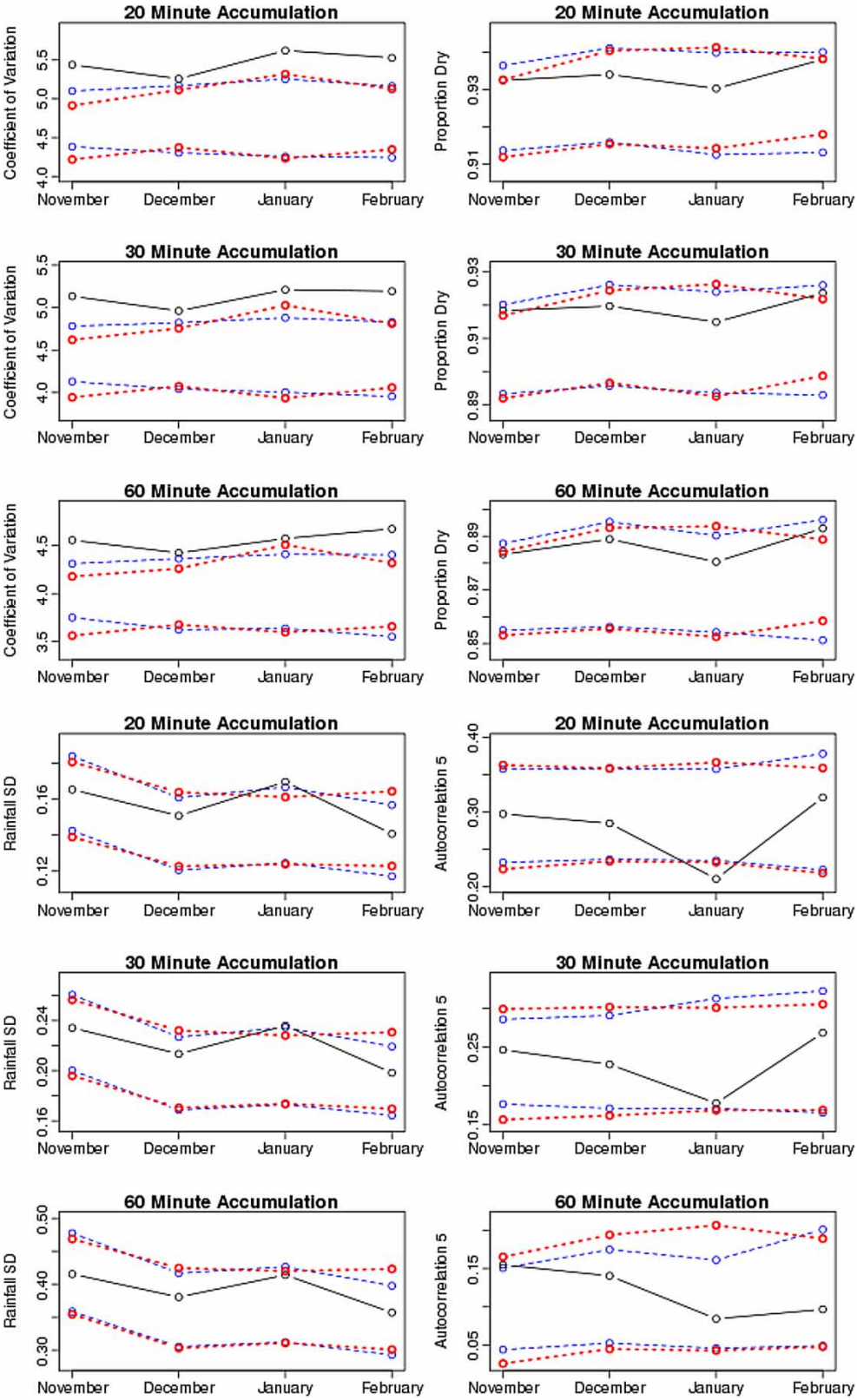

Figure 5 | Coefficient of variation, standard deviation, proportion of dry periods and autocorrelation at lag 5 of the observed rainfall with simulation bands from the FTR and FP covariate models.

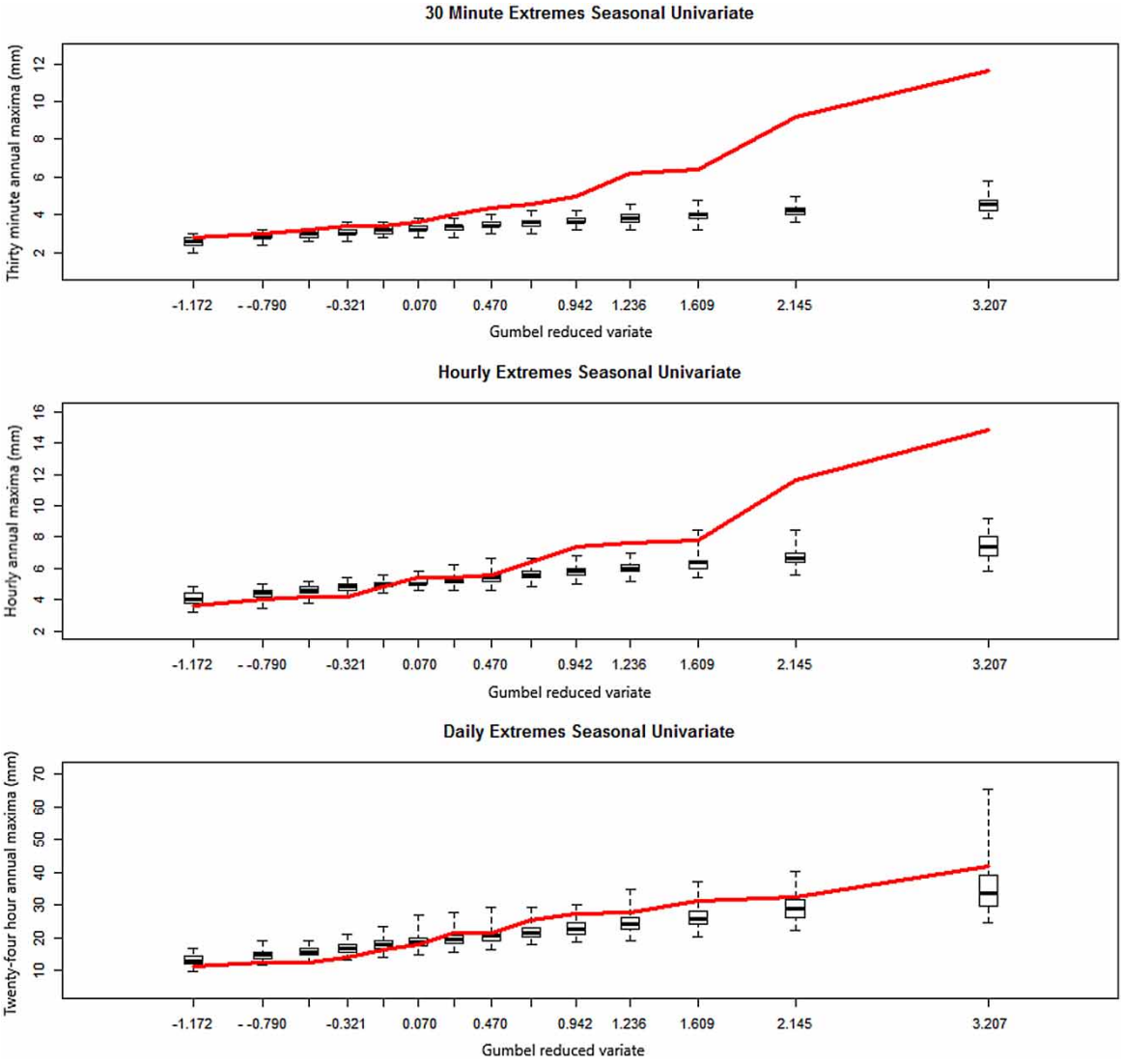

Figure 6 | Ordered annual extreme rainfall at different aggregations with boxplots using annual maxima of 100 simulations from the FTR model.

of relative humidity, sea-level air-pressure, and potential temperature, while the model does still gain some benefit from knowledge of calendar month, this benefit will be less pronounced than in the model without covariates.

ACKNOWLEDGEMENTS

This work was supported by a Vice-Chancellor's scholarship from the University of Greenwich (Ref. No: VCS-ACH-06-14).

REFERENCES

Cowpertwait, P., Isham, V. & Onof, C. 2007 Point process models of rainfall: developments for fine-scale structure. *Proc. R. Soc. A* **463** (2086), 2569–2588.

Cox, D. R. 1955 Some statistical methods connected with series of events (with Discussion). *J. R. Stat. Soc. B* **27**, 129–164.

Davison, A. C. & Ramesh, N. I. 1993 A stochastic model for times of exposures to air pollution from a point source. In: *Statistics for the Environment* (V. Barnett & K. F. Turkman, eds). Wiley, New York, pp. 123–138.

Fischer, W. & Meier-Hellstern, K. 1993 The Markov-modulated Poisson process (MMPP) cookbook. *Perf Evaf.* **18**, 149–171.

Gumbel, E. J. 1954 *Statistical Theory of Extreme Values and Some Practical Applications*. Applied Mathematics Series 33 (1st edn). U.S. Department of Commerce, National Bureau of Standards.

Hughes, J. P. & Guttorp, P. 1994 PA class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.* **30** (5), 1535–1546.

Kigobe, M., McIntyre, N., Wheater, H. & Chandler, R. 2011 Multi-site stochastic modelling of daily rainfall in Uganda. *Hydrol. Sci. J.* **56** (1), 17–33.

Onof, C. & Wheater, H. S. 1993 Modeling of British rainfall using a random parameter Bartlett–Lewis Rectangular ulse model. *J. Hydrol.* **149**, 67–95.

Onof, C. & Wheater, H. S. 1994 Improvements to the modeling of British rainfall using a random parameter Bartlett-Lewis Rectangular Pulse model. *J. Hydrol.* **157**, 177–195.

Onof, C., Yameundjeu, B., Paoli, J. P. & Ramesh, N. I. 2002 A Markov modulated Poisson process model for rainfall increments. *Water Sci. Technol.* **45**, 91–97.

Ramesh, N. I. 1995 Statistical analysis on Markov-modulated Poisson processes. *Environmetrics* **6**, 165–179.

Ramesh, N. I. 1998 Temporal modelling of short-term rainfall using Cox processes. *Environmetrics* **9**, 629–643.

Ramesh, N. I., Onof, C. & Xie, D. 2012 Doubly stochastic Poisson process models for precipitation at fine time-scales. *Adv. Water Resour.* **45**, 58–64.

Ramesh, N. I., Thayakaran, R. & Onof, C. 2013 Multi-site doubly stochastic Poisson process models for fine-scale rainfall. *Stoch. Environ. Res. Risk Assess.* **27**, 1383–1396.

Ross, S. M. 1983 *Stochastic Processes*. Series in Probability and Mathematical Statistics. Wiley, New York.

Rydén, T. 1996 An EM algorithm for estimation in Markov-modulated Poisson processes. *Comput. Stat. Data Anal.* **21**, 431–447.

Smith, R. L. 1984 Contribution to the discussion of Stern, R. D. and Coe, R. A model fitting analysis of daily rainfall data (with Discussion). *J. R. Stat. Soc. A* **147**, 24–25.

Smith, J. A. & Karr, A. F. 1983 A point process model for summer season rainfall occurrences. *Water Resour. Res.* **19**, 95–103.

Smith, R. L. & Karr, A. F. 1985 Statistical inference for point process models of rainfall. *Water Resour. Res.* **21** (1), 73–79.

Stern, R. D. & Coe, R. 1984 A model fitting analysis of daily rainfall data. *J. R. Stat. Soc. A* **147**, 1–34.

Thayakaran, R. & Ramesh, N. I. 2013 Multivariate models for rainfall based on Markov modulated Poisson processes. *Hydrol. Res.* **44**, 631–643.

Verhoest, N., Troch, P. A. & De Troch, F. P. 1997 On the applicability of Bartlett–Lewis rectangular pulses models in the modeling of design storms at a point. *J. Hydrol.* **202** (1), 108–120.

First received 16 January 2018; accepted in revised form 28 April 2018. Available online 11 June 2018