22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

Fitting the Bartlett-Lewis rainfall model using Approximate Bayesian Computation

Nanda R. Aryal [a], **Owen D. Jones** [b]

[a]*School of Mathematics and Statistics, University of Melbourne, Australia*
[b]*School of Mathematics, Cardiff University, UK*
Email: naryal@student.unimelb.edu.au

Abstract: The Barlett-Lewis (BL) rainfall model is a stochastic model for the rainfall at a single point in space, constructed using a cluster point process. The cluster process is constructed by taking a primary/parent process, called the storm arrival process in our context, and then attaching to each storm point a finite secondary/daughter point process, called a cell arrival process. To each cell arrival point we then attach a rain cell, with an associated rainfall duration and intensity. The total rainfall at time t is then the sum of the intensities from all active cells at that time.

Following Rodriguez-Iturbe et al. (1987), we suppose that the storm arrival process is a Poisson process, and that the cell arrival processes are independent Poisson processes, truncated after an exponentially distributed time (the storm duration). Rain cells are all i.i.d., with independent exponentially distributed duration and intensity.

Because it has an intractible likelihood function, in the past the BL model has been fitted using the Generalized Method of Moments (GMM). The puprose of this paper is to show that Approximate Bayesian Computation (ABC) can also be used to fit this model, and moreover that it gives a better fit than GMM. GMM fitting matches theoretical and observed moments of the process, and thus is restricted to moments for which you have an analytic expression. ABC fitting compares the observed process to simulations, and thus places no restrictions on the statistics used to compare them. The penalty we pay for this increased flexibility is an increase in computational time.

The ABC methodology supposes that we have an observation D from some model $f(\cdot|\boldsymbol{\theta})$, depending on parameters $\boldsymbol{\theta}$, and that we are able to simulate from f. Let π be the prior distribution for $\boldsymbol{\theta}$ and $S = S(D)$ a vector of summary statistics for D, then ABC generates samples from $f(\boldsymbol{\theta}|\rho(S(D^*), S(D)) < \epsilon)$, where $D^* \sim f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \sim \pi$, and ρ is some distance function. If S is a sufficient statistic, then as $\epsilon \to 0$ this will converge to the posterior $f(\boldsymbol{\theta}|D)$.

The choice of good summary statistics is important to the success of ABC fitting. To fit the BL model we used rainfall aggregated over six-minute and hourly intervals, and then compared the mean, standard deviation, auto-correlation at lags 1 and 2, probability of no rain, mean length of wet and dry periods, standard deviation of wet and dry periods, and the total number of wet and dry periods We note that for GMM fitting we can only use the first five of these statistics, because we do not have analytic expressions for the others. Using a simulation study we demonstrate that ABC fitting can give less biased and less variable estimates than GMM. We also give an application to rainfall data from Bass River, Victoria, July 2010. Again we see that the ABC fit is better than the GMM fit.

An important advantage of ABC fitting over GMM fitting is that we can use summaries of the data that capture useful information, whether or not we have an expression for their expectation. Moreover, this means that ABC can be used for models for which GMM fitting is not available. For example, if we used a gamma distribution for the duration of a rain cell, rather than an exponential distribution, then we would not be able to calculate the second order statistics of the model, making GMM fitting impossible. However ABC fitting would proceed as before, with the addition of a single parameter. This opens up the possibility of fitting much more realistic stochastic rainfall models.

Keywords: *Bartlett-Lewis process, rainfall, simulation, Generalized Method of Moments, Approximate Bayesian Computation, Markov Chain Monte Carlo*

1 INTRODUCTION

The Barlett-Lewis (BL) rainfall model is a stochastic model for the rainfall at a single point in space, constructed using a cluster point process. The cluster process is constructed by taking a primary/parent process, called the storm arrival process in our context, and then attaching to each storm point a finite secondary/daughter point process, called a cell arrival process. To each cell arrival point we then attach a rain cell, with an associated rainfall duration and intensity. The total rainfall at time t is then the sum of the intensities from all active cells at that time.

Because it has an intractible likelihood function, the BL model has been fitted using the Generalized Method of Moments (GMM). The puprose of this paper is to show that Approximate Bayesian Computation (ABC) can also be used to fit this model, and to show using a simulation study that it gives a better fit than GMM. GMM fitting matches theoretical and observed moments of the process, and thus is restricted to moments for which you have an analytic expression. ABC fitting compares the observed process to simulations, and thus places no restrictions on the statistics used to compare them. The penalty we pay for this increased flexibility is an increase in computational time.

Because our primary goal is to compare GMM and ABC fitting, we will restrict ourselves to a simple BL model, namely the rectangular pulse model introduced by Rodriguez-Iturbe et al. (1987). See Cowpertwait et al. (2007) for some more recent refinements.

We use a (homogeneous) Poisson process with rate λ for the storm arrival process. The cell arrival processes are independent processes, each one a Poisson process of rate β, truncated after an exponential(γ) ammount of time, which we call the storm duration. Assuming that we are working in a finite time window, denote the storm arrival times $T_1, T_2, \ldots T_n$ and the storm durations $D_1, D_2, \ldots, D_n$. Let the arrival times for the i-th cell arrival process be $S_1^i, S_2^i, \ldots, S_{k(i)}^i \in [0, D_i]$, where $k(i)$ (possibly zero) is the number of cells in storm i. The cell arrival times are thus $\{T_i + S_j^i \,:\, i = 1, \ldots, n,\ j = 1, \ldots, k(i)\}$.

Rain cells are independent with duration and intensity having independent exponential(η) and exponential($1/\mu_x$) distributions. The intensity is constant during a cell's lifetime. Suppose the j-th cell in storm i has duration L_j^i and intensity X_j^i, then the overall intensity of rainfall at time t is

$$Y(t) = \sum_i \sum_j \mathbb{I}_{\{T_i + S_j^i < t \leq T_i + S_j^i + L_j^i\}}\, X_j^i.$$

An illustration of the components of a BL process is given in Figure 1.

Rain gauges record accumulated rather than instantaneous rainfall. Accordingly we define the rainfall for i-th time period of length h to be $Y_i^h = \int_{(i-1)h}^{ih} Y(t)\, dt$. Rodriguez-Iturbe et al. (1987) derive expressions for the mean, variance, and covariances of the Y_i^h, as well as $\mathbf{P}(Y_i^h = 0)$. GMM fitting requires moments with analytical expressions, and we used mean, standard deviation, auto-correlation at lags 1 and 2, and probability of no rain, for both six-minute and hourly aggregated data, giving a total of nine summary statistics. (Note that given the mean at six-minute intervals, the mean at hourly intervals contains no additional information, and so is not included.)

Suppose that $\mathbf{V} = (V_1, \ldots, V_k)'$ is a vector of statistics computed from our observed data, with expectations $\tau(\boldsymbol{\theta}) = (\tau_1(\boldsymbol{\theta}), \ldots, \tau_k(\boldsymbol{\theta}))'$, depending on some unkown parameter vector $\boldsymbol{\theta}$. The GMM estimate of $\boldsymbol{\theta}$ is then

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\, (\mathbf{V} - \tau(\boldsymbol{\theta}))'\, \mathbf{W}\, (\mathbf{V} - \tau(\boldsymbol{\theta})),$$

where $\mathbf{W}$ is a positive definite weighting matrix, usually assumed to be diagonal. It can be shown that the optimal weights are inversely proportional to $\text{Var}\,(V_i)$, and can be estimated iteratively. For specific details of GMM fitting for BL models, we refer the reader to Wheater et al. (2005) and Kaczmarska (2011), for example.

2 APPROXIMATE BAYESIAN COMPUTATION

ABC was introduced by Pritchard etl. (1999), and was later extended to incorporate Markov Chain Monte Carlo (MCMC) Marjoram et al. (2003), or alternatively Sequential Monte Carlo (SMC) Sisson et al. (2007, 2009); Beaumont et al. (2009). We will use the ABC-MCMC methodology of Marjoram et al. (2003).

We suppose that we have an observation D from some model $f(\cdot|\boldsymbol{\theta})$, depending on parameters $\boldsymbol{\theta}$, and that we are able to simulate from f. Let π be the prior distribution for $\boldsymbol{\theta}$ and $S = S(D)$ a vector of summary statistics for D, then ABC generates samples from $f(\boldsymbol{\theta}|\rho(S(D^*), S(D)) < \epsilon)$, where $D^* \sim f(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \sim \pi$,

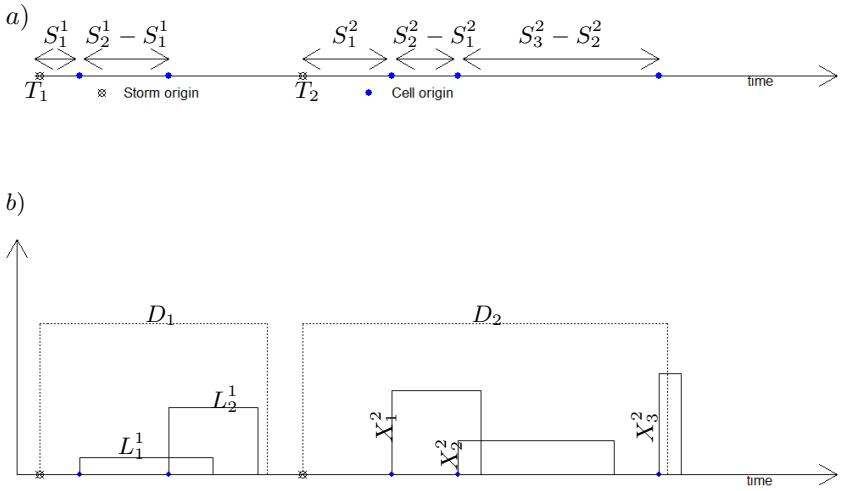

Figure 1. Constituent parts of a Bartlett-Lewis process: a) Storm process and cell processes b) Storm durations, cell durations, and cell intensities

and ρ is some distance function. If S is a sufficient statistic, then as $\epsilon \to 0$ this will converge to the posterior $f(\boldsymbol{\theta}|D)$. ABC-MCMC adds a proposal chain with density q and a rejection step, to generate a sample $\{\boldsymbol{\theta}_i\}$. The algorithm is as follows:

Algorithm 1 ABC-MCMC

for i=1 to N **do**

1. Given current state $\boldsymbol{\theta}_i$ propose a new state $\boldsymbol{\theta}^*$ using $q(\cdot|\boldsymbol{\theta}_i)$
2. Put $\alpha = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_i|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)}\right)$
3. Go to step 4 with probability α, otherwise set $\theta_{i+1} = \theta_i$ and return to step 1
4. Simulate data $D^* \sim f(\cdot|\theta^*)$
5. If $\rho(S(D^*), S(D)) \leq \epsilon$ then set $\theta_{i+1} = \theta^*$, otherwise set $\theta_{i+1} = \theta_i$

end for

Note that the MCMC rejection at step 3 comes before the ABC comparison in step 5. This is to avoid unnecessarily running the simulation in step 4.

2.1 Applying ABC-MCMC to the BL model

Firstly we reparameterise the model, to reduce the dependence between the parameters. In addition we use a log transformation to map them from $\mathbb{R}_+$ to $\mathbb{R}$, which simplifies the choice of the proposal chain.

The total intensity at time t has mean $I_T = \lambda\gamma^{-1}\beta\eta^{-1}\mu_x$. Our new parameters are

$$\begin{aligned}
\theta(1) &= \log(I_T) \\
\theta(2) &= \log(\lambda\gamma^{-1}) \\
\theta(3) &= \log(\lambda\gamma) \\
\theta(4) &= \log(\beta\eta^{-1}) \\
\theta(5) &= \log(\beta\eta)
\end{aligned}$$

These parameters are still dependent, for example a low storm arrival rate and long storm duration can give the same total intensity as a high storm arrival rate and short storm duration. None-the-less we found that this reparameterisation improved estimation; in particular I_T is much easier to estimate than μ_x.

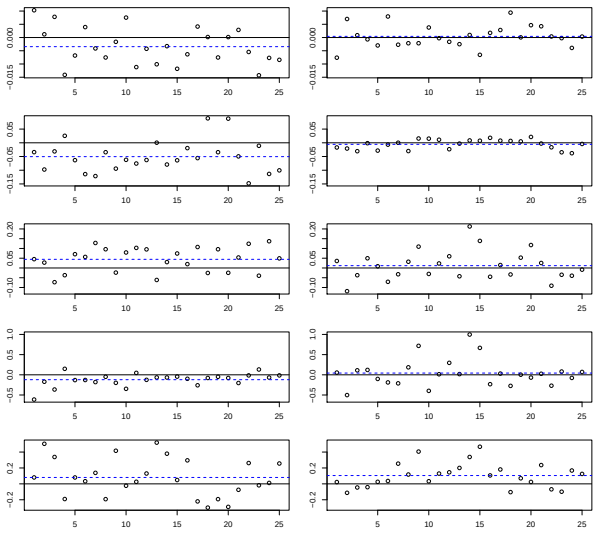

Figure 2. From top to bottom, plots are errors in the estimates of parameters λ, γ, β, η and μ_x, for 25 separate simulations. Left figures are GMM estimation errors and right figures are ABC-MCMC estimation errors. The blue dotted lines give the average bias.

.

Vague normal priors were used for all the $\theta(i)$; that is $\pi(\boldsymbol{\theta}) \sim N(\mathbf{0}, \sigma^2 I)$ for σ^2 large. For the proposal chain we just used a random walk. Note that as the proposal distribution is symmetric, α will depend only on the prior.

The choice of good summary statistics is important to the success of ABC fitting. In addition to the statistics used for GMM (mean, standard deviation, auto-correlation at lags 1 and 2, probability of no rain), we also used mean length of wet and dry periods, standard deviation of wet and dry periods, and the total number of wet and dry periods, again for six-minute and hourly aggregated data. This gave us a vector of 19 summary statistics for ABC-MCMC fitting.

Note that while it is important that our summary statistics are sufficient, including unnessesary statistics will reduce the performance of the ABC estimator, essentially by introducing noise that makes it harder to distinguish between good and bad simulations. This can be mitigated somewhat using the post-hoc analysis of Beaumont et al. (2002), which reweights the component statistics to give more importance to those that can better predict the quality of a simulation. However we also determined experimentally that using a smaller set of summary statistics gave poorer estimates.

For the distance measure ρ we use a weighted Euclidean metric,

$$\rho(S(D^*), S(D)) = \sum_i w_i (S^*(i) - S(i))^2,$$

where $S^*(i)$ and $S(i)$ are respectively the i-th component of $S(D^*)$ and $S(D)$. Just as the choice of summary S is important, so to is the choice of weights. Various authors have found that choosing w_i inversely propotional to the variance of $S^*(i)$ works well, formally giving equal importance to each component of S. In practice we estimate $\mathrm{Var}\,(S^*(i))$ using a sample generated from $f(\cdot|\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is a preliminary estimate of $\boldsymbol{\theta}$.

3 Simulation study

In this section we use a simulated data set to compare GMM and ABC-MCMC parameter estimation for the BL model. Using $\lambda = 0.04$, $\gamma = 0.20$, $\beta = 0.50$, $\eta = 2.00$ and $\mu_x = 1.50$, we simulated rainfall for a two week period and then used GMM and ABC-MCMC to estimate the parameters. This was repeated 25 times to guage the bias and variability of each estimator.

ABC-MCMC requires tuning to perform well. We need to choose ϵ small enough that we get a good approximation to the posterior, but large enough that the chain has a reasonable acceptance rate. Also, for very small

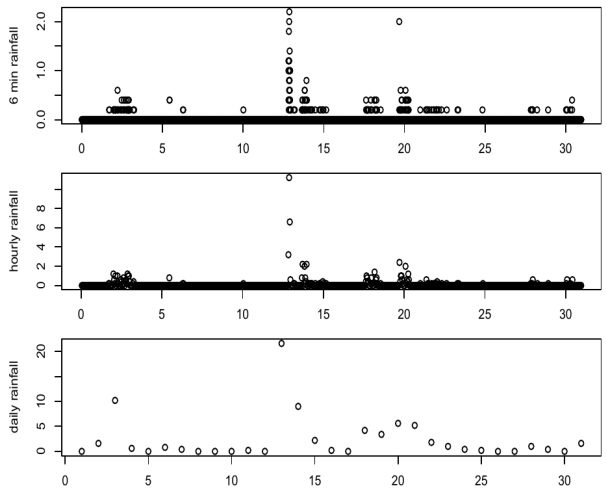

Figure 3. Rainfall measurements from Bass River, Victoria, July 2010. The x-axis is measured in days and the y-axis in mm. Data obtained from the Australian Bureau of Meteorology.

ϵ it can be difficult to get the chain started, particularly if your starting point is in a region of low posterior probability. The practical solution to this problem is to run a short initial ABC estimate (using i.i.d. samples from the prior π, instead of using the proposal density q). This allows us to roughly estimate the distribution of $\rho(S(D^*), S(D))$ and thus choose ϵ. It also allows us to choose a starting point for the MCMC chain that has high posterior probability (removing the need for a burn-in period), and to refine the weights w_i used in ρ.

For priors we used the $N(0, 3.0^2)$ distribution for each $\theta(i)$. As for any MCMC procedure, the proposal chain needs to be chosen so that it mixes well and explores the whole parameter space. We used a random walk with $N(\mathbf{0}, 0.2^2 I)$ increments. Combined with a threshold of $\epsilon = 2.5$, this gave an acceptance rate of around 3%.

We used the posterior mean of the ABC-MCMC sample to get a point estimate that we could compare directly to the GMM estimate. We used local linear regression to calculate the posterior mean, as suggested by Beaumont et al. (2002). The results are given in Figure 2. This graphs clearly show that ABC-MCMC gives less biased and less variable estimates than GMM.

4 APPLICATION TO REAL DATA

Figure 3 gives rainfall for Bass River, Victoria, July 2010. Rainfall is measured in increments of 0.2 mm every 6 minutes using a tipping bucket. Rainfall of less than 0.2 mm is considered as no-rainfall. In this section we fit a Bartlett-Lewis model to these data.

We used independent $N(0, 3.0^2)$ priors for the $\theta(i)$. For the proposal chain we used a random walk with $N(\mathbf{0}, 0.2^2 I)$ increments. Trace plots were used to verify that the chain was mixing nicely. Figure 4 gives the estimated posterior densities for the original (untransformed) parameters. The diagonals are marginal densities and the off-diagonals pairwise densities.

In Table 1 we give the posterior mean, median and 95% credible intervals for each parameter, together with a GMM estimate. Note that the GMM estimate sits within the credible interval in each case.

To judge the models fitted using GMM and ABC-MCMC, we used simulation to generate 95% confidence intervals for a variety of statistics (at different levels of temporal aggregation), and compared these to their observed values. The results are given in Figure 5. We see that the GMM fitted model only gives a good correspondence between the fitted model and the data for those statistics used in the GMM fit, but the ABC-MCMC fitted model gives a good correspondence for all the statistics considered.

5 CONCLUSIONS

Using both a simulation study and real data, we have seen that ABC-MCMC gives better fits than GMM, for fitting a Bartlett-Lewis rainfall model. An important advantage of ABC fitting over GMM fitting is that we can use summaries of the data that capture useful information, whether or not we have an expression for their

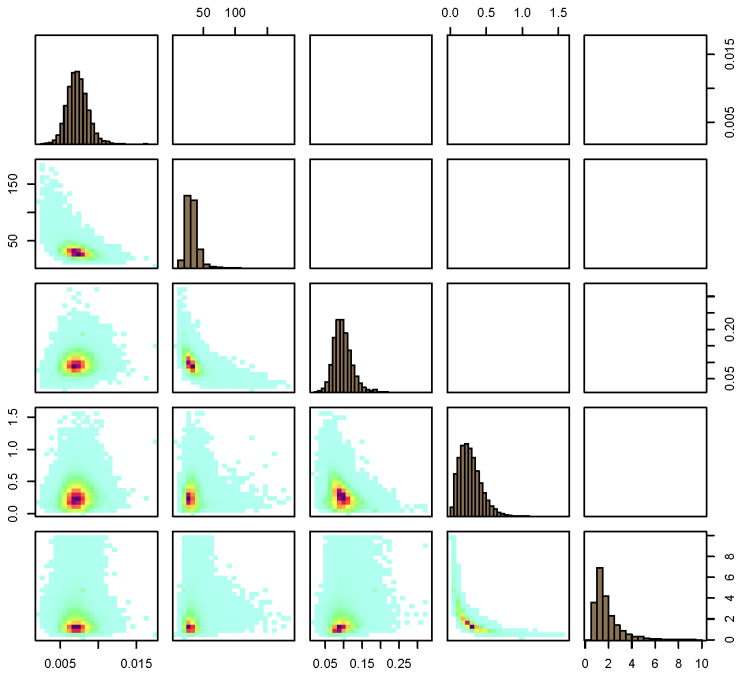

Figure 4. Posterior distributions (marginals and pairs) for λ, γ^{-1}, β, η^{-1} and μ_x. From the BL model fitted to the Bass River data.

Table 1. Parameter estimates for the BL model fitted to the Bass River data. All estimated parameter values are per hour except μx, which is mm per hour.

Parameter	GMM	ABC-MCMC Posterior mean	ABC-MCMC 95% credible interval	ABC-MCMC Posterior median
λ	0.0950	0.0707	(0.0461, 0.0981)	0.0701
γ	0.3709	0.2819	(0.1585 , 0.4858)	0.2978
β	0.1911	0.9697	(0.5526, 1.5360)	0.9369
η	0.7091	0.4456	(0.1987, 2.1642)	0.4830
μ_x	1.3957	2.3999	(0.8642, 6.7225)	1.9149

expectation. Moreover, this means that ABC can be used for models for which GMM fitting is not available. For example, if we used a gamma distribution for the duration of a rain cell, rather than an exponential distribution, then we would not be able to calculate the second order statistics of the $\{Y_i^h\}$, making GMM fitting impossible. However ABC fitting would proceed as before, with the addition of a single parameter. This opens up the possibility of fitting much more realistic stochastic rainfall models.

Finally we note that unlike GMM, ABC fitting provides credible intervals and not just point estimates.

ACKNOWLEDGEMENTS

The authors thank the Australian Bureau of Meteorology for the data used. Nanda acknowledges with thanks the support received from an Australian Government Research Training Program Scholarship and a Melbourne Research Scholarship.

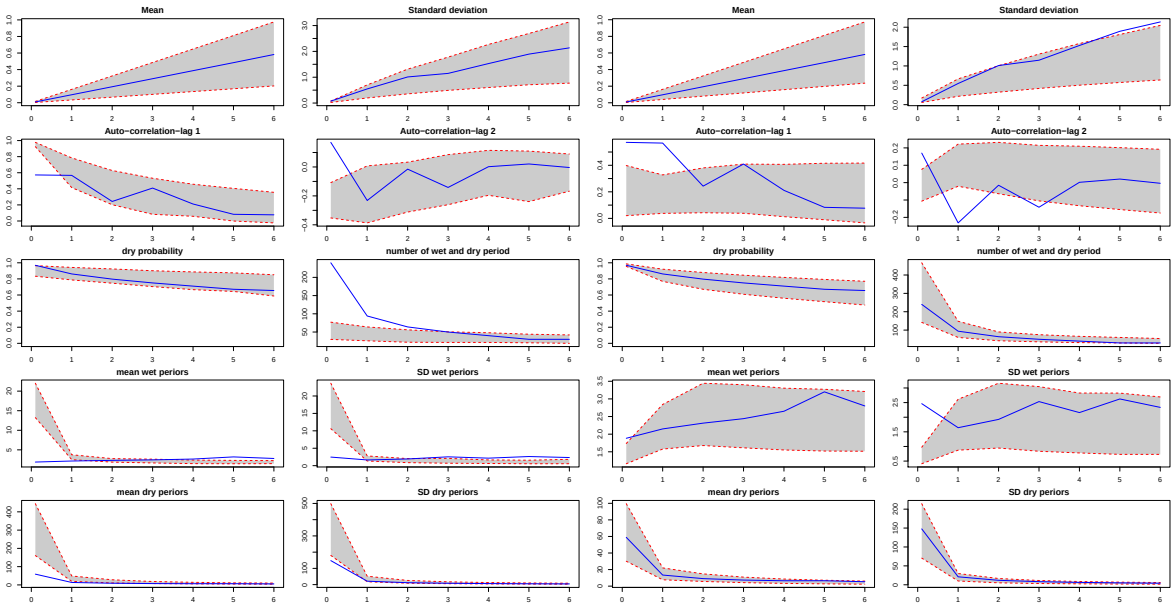

Figure 5. 95% confidence intervals for various statistics; the left two columns use GMM estimates and the right two columns use ABC-MCMC estimates. For each figure, the statistics were calculated using rainfall aggregated over intervals of 0.1, 1, 2, 3, 4, 5, 6 hours, and were calculated from 3000 independent simulations. The solid blue lines give the observed statistics.

REFERENCES

Beaumont, M. A., J.-M. Cornuet, J.-M. Marin, and C. P. Robert (2009). Adaptive approximate Bayesian computation. *Biometrika 96*(4), 983–990.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian computation in population genetics. *Genetics 162*(4), 2025–2035.

Cowpertwait, P., V. Isham, and C. Onof (2007). Point process models of rainfall: developments for fine-scale structure. *Proc. Roy. Soc. London 463*, 2569–2587.

Kaczmarska, J. (2011). Further development of Bartlett–Lewis models for fine-resolution r ainfall. Research Report No. 312, Department of Statistical Science, University College London.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov Chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences 100*(26), 15324–15328.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution 16*, 1791–1798.

Rodriguez-Iturbe, I., D. R. Cox, and V. Isham (1987). Some process for rainfall: further develepoment. *Proc. Roy. Soc. London 410*, 269–288.

Sisson, S., Y. Fan, and M. M. Tanaka (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States 104*, 1760–1765.

Sisson, S., Y. Fan, and M. M. Tanaka (2009). Correction for "Sequential Monte Carlo without likelihoods". *Proceedings of the National Academy of Sciences of the United States 106*, 16889.

Wheater, H. S., R. E. Chandler, C. J. Onof, V. S. Isham, E. Bellone, C. Yang, D. Lekkas, G. Lourmas, and M. L. Segond (2005). Spatial-temporal rainfall modelling for flood risk estimation. *Stochastic Environmental Research and Risk Assessment 19*(6), 403–416.