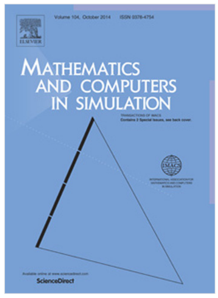
Journal Pre-proof



Fitting the Bartlett-Lewis rainfall model using approximate Bayesian computation

Nanda R. Aryal, Owen D. Jones

PII: S0378-4754(19)30331-3
DOI: https://doi.org/10.1016/j.matcom.2019.10.018
Reference: MATCOM 4889

To appear in: *Mathematics and Computers in Simulation*

Received date : 14 May 2018
Revised date : 14 October 2019
Accepted date : 26 October 2019

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, U.K.

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.

Because it has an intractable likelihood function, in the past the BL model has been fitted using the Generalized Method of Moments (GMM). The purpose 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.

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

1. Introduction

Stochastic rainfall models based on the Barlett-Lewis cluster process were introduced by Rodriguez-Iturbe et al. [1, 2], with later refinements by Cowpertwait et al. [3] and Kaczmarska et al. [4, 5]. Because the Barlett-Lewis (BL) rainfall model has an intractable likelihood function, it has in the past been fitted using the Generalized Method of Moments (GMM).

GMM requires a vector of summary statistics whose means have closed-form expressions, and the broader the variety of summary statistics used, the better the fitting. In contrast, Approximate Bayesian Computation (ABC) is a likelihood free method that compares the observed process to a model using simulations, and places no restrictions on the summary statistics used to compare them. As a result, ABC fitting allows us to use summary statistics that are potentially very informative, but do not have nice closed-form expressions for their means. Beyond that, ABC allows us to consider models whose complexity is not constrained by the need to obtain closed-form expressions for various summary statistics.

Our primary goal is to demonstrate that ABC fitting works as well as—or better than—GMM fitting for BL type rainfall models. Accordingly we will restrict ourselves to the simplest form of the BL model, namely the rectangular pulse model introduced by Rodriguez-Iturbe et al. [1]. To our knowledge this is the first time ABC has been used to fit rainfall models.

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) amount of time, which we call the storm duration. Assuming that we are working in a finite time window, denote the storm arrival times $T_1, T_2, \dots T_n$ and the storm durations $D_1, D_2, \dots, D_n$. Let the arrival times for the i-th cell arrival process be $S_1^i, S_2^i, \dots, S_{k(i)}^i \in [0, D_i]$, where $k(i)$ (possibly zero) is the number of cells in storm i. The cell arrival times are thus $\{T_i + S_j^i \, : \, i = 1, \dots, n, \; j = 1, \dots, k(i)\}$.

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

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

Rain gauges record accumulated rather than instantaneous rainfall. Accord-

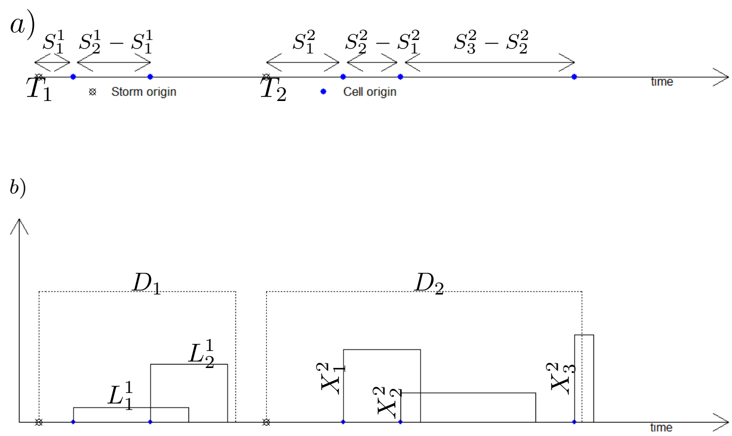


Figure 1: Constituent parts of a Bartlett-Lewis process: a) Storm process and cell processes b) Storm durations, cell durations, and cell intensities

ingly we define the rainfall for the i-th time period of length h to be

$$Y_i^h = \int_{(i-1)h}^{ih} Y(t)\, dt.$$

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

In the section Conclusions we discuss the implications of our findings for the modelling process, in particular regards the successful fitting of more complex (and thus more realistic) models.

1.1. Generalised Method of Moments

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 unknown 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. It can be shown that the optimal weights are proportional to $(\mathrm{Var}\,\mathbf{V})^{-1}$, and can be estimated iteratively.

Rodriguez-Iturbe et al. [1] 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, autocorrelation 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.) For specific details of GMM fitting for BL models, we refer the reader to [6, 4], for example.

1.2. Approximate Bayesian Computation

ABC was introduced by Pritchard et al. [7], and was later extended to incorporate Markov Chain Monte Carlo (MCMC) [8], or alternatively Sequential Monte Carlo (SMC) [9, 10, 11]. We will use the ABC-MCMC methodology in what follows. Wegmann et al. [12] provide some practical advice on implementing ABC-MCMC, and give a proof of the central result in Marjoram et al. [8].

We assume that we have an observation D, supposedly from some model $f(\cdot|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$, and that we are able to simulate from f. D could be actual data from a rain gauge, or simulated data when validating the procedure. 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 $D^* \sim f(\cdot|\boldsymbol{\theta}^*)$, $\boldsymbol{\theta}^*$ is a proposal for $\boldsymbol{\theta}$ distributed according to π, 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 $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$ and return to step 1
4. Simulate data $D^* \sim f(\cdot|\boldsymbol{\theta}^*)$
5. If $\rho(S(D^*), S(D)) \le \epsilon$ then set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$, otherwise set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$

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. 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.

Our first three parameters give a top down description of the process, more directly relatable to the observed rainfall. The total intensity at time t has mean $I_T = \lambda\gamma^{-1}\beta\eta^{-1}\mu_x$; the percentage of time covered by storms has mean roughly proportional to $\lambda\gamma^{-1}$; and the percentage of a storm covered by rain cells has mean roughly proportional to $\beta\eta^{-1}$. Our final two parameters were chosen to be roughly orthogonal to these three, with respect to the posterior. Our new parameters are

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

Posterior plots of these parameters show less dependence than the originals. We also found that I_T is much easier to estimate than μ_x, and that the new parameterisation avoids the very strong negative correlation between λ and γ^{-1} (which exists because 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).

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. Ideally, summary statistics should discriminate between the effects of all the model parameters. For example, a process with small storm arrival rate λ and large storm duration $1/\gamma$ may have the same expected rainfall intensity (proportional to λ/γ) as a process with high storm arrival rate and small storm duration. Thus the average rainfall will tell us a lot about λ/γ, but to distinguish between λ and γ we need other statistics, for example statistics that can measure the clustering of rainfall. We can also choose summary statistics to support our research objectives. For example, if it is important that our model accurately reproduces rainfall extremes, then we can include statistics that measure extremal behaviour, such as the number of exceedances over a given threshold. In our case, 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 unnecessary statistics will reduce the performance of the ABC estimator, essentially by introducing noise that makes it harder to distinguish between good and bad simulations. (By a good simulation we mean one which was generated using parameters close to the true parameters.) This can be mitigated somewhat using the post-hoc analysis of Beaumont et al. [13], which uses regression to determine which summary statistics are most significant. We also determined experimentally that leaving out any one of our summary statistics gave less precise 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 too is the choice of weights. Various authors have found that choosing w_i inversely proportional to the variance of $S^*(i)$ works well, formally giving equal importance to each component of S. In the context of least-squares estimation this weighting is optimal when the $S^*(i)$ are independent [14]. When the $S^*(i)$ are dependent, as is certainly the case here, the Mahalanobis distance has been observed to work well [15]. However in practice the Mahalanobis distance is problematic for more than a small number of summary statistics, as the covariance matrix is often close to singular.

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

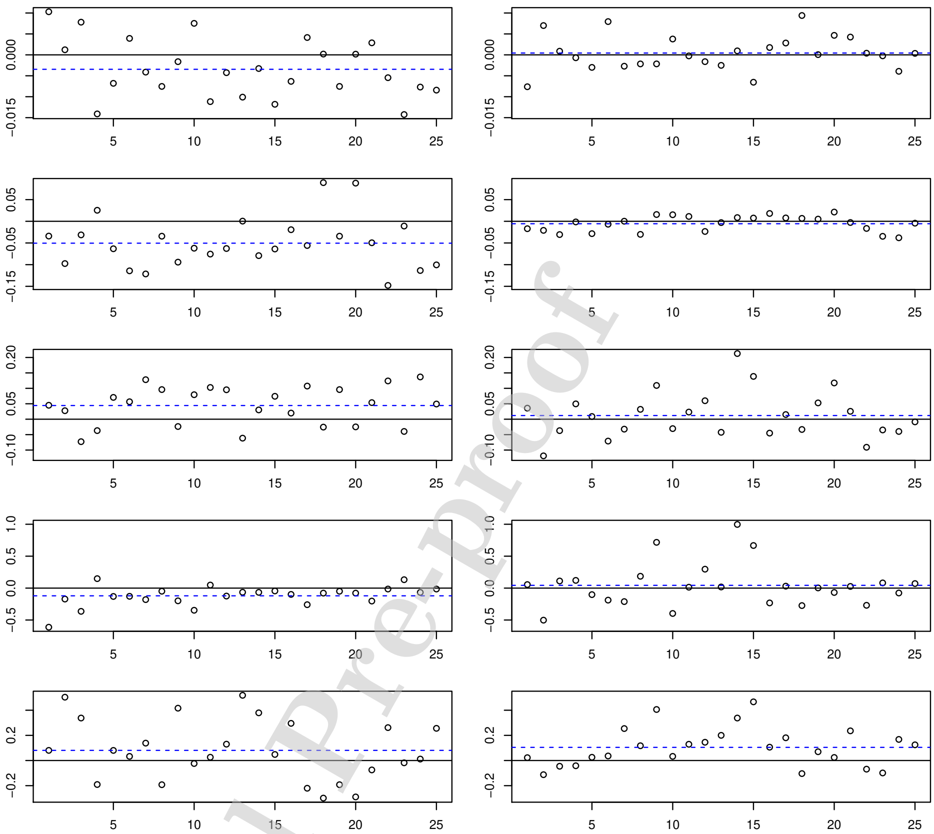


Figure 2: From top to bottom, plots are errors in the estimates of parameters λ, γ, β, η and μ_x, for 25 separate simulations. Left figures are GMM estimation errors and right figures are ABC-MCMC estimation errors. The blue doted lines give the average bias.

.

and then used GMM and ABC-MCMC to estimate the parameters. This was repeated 25 times to gauge 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 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 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%, and an efficiency of around 0.003 for each parameter (the effective sample size over the number of simulations).

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. [13]. The results are given in Figure 2. These 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.

A preliminary version of these results appeared in [16]. The results here benefit from increased ABC simulation time, and the estimates given correct a previous scaling error (the previous estimates were all out by a factor of ten). By comparing various parameter combinations we show where the ABC and GMM fits agree and where they differ. Then by considering how well the fitted models describe wet and dry periods, we show that the ABC fit is doing a better job.

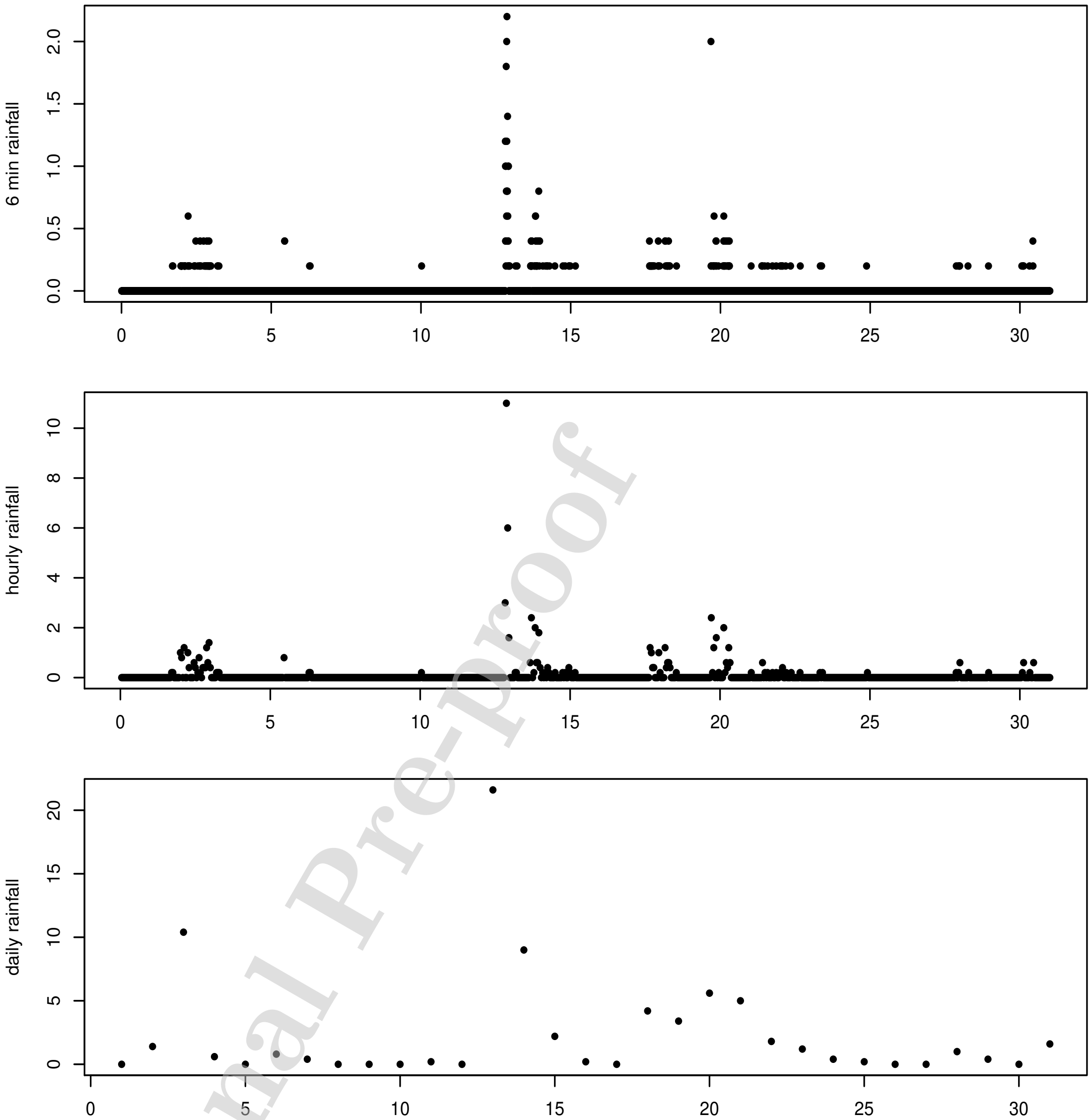


Figure 3: Rainfall measurements from Bass River, Victoria, July 2010. The x-axis is measured in days and the y-axis in mm. Data obtained from the Australian Bureau of Meteorology.

Parameter	GMM	ABC-MCMC Posterior mean	ABC-MCMC 95% credible interval	ABC-MCMC Posterior median
λ	0.010	0.007	(0.005, 0.010)	0.007
γ^{-1}	26.96	32.35	(18.57, 56.38)	30.73
β	0.019	0.097	(0.056, 0.156)	0.095
η^{-1}	14.10	0.279	(0.057, 0.634)	0.256
μ_x	0.140	1.994	(0.699, 6.450)	1.531

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

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). Figures 5 and 6 give the estimated posterior densities for $\boldsymbol{\theta}$ and 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. The ABC and GMM estimates for λ and γ^{-1} are roughly similar, and if you take the mean storm coverage (storm rate by mean duration $\lambda\gamma^{-1}$) you get an even closer match, with 0.256 for the GMM fit and 0.226 using the ABC posterior means. The estimates for β, η^{-1} and μ_x are quite different, however if you take the mean storm intensity ($\beta\eta^{-1}\mu_x$) you get 0.038 using the GMM fit and 0.054 using the ABC posterior means, which are again roughly similar. Formally, storm coverage and storm intensity are essentially directly observable quantities, making their means easier to estimate.

To compare the models fitted using GMM and ABC-MCMC, we used simulation to generate 95% Monte-Carlo predictive intervals for a variety of statistics (at different levels of temporal aggregation), and compared these to their observed values. The results are given in Figures 7 and 8. 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. That is, the extra statistics—such as the number of wet/dry periods—has helped ABC to distinguish β, η^{-1} and μ_x more successfully than GMM.

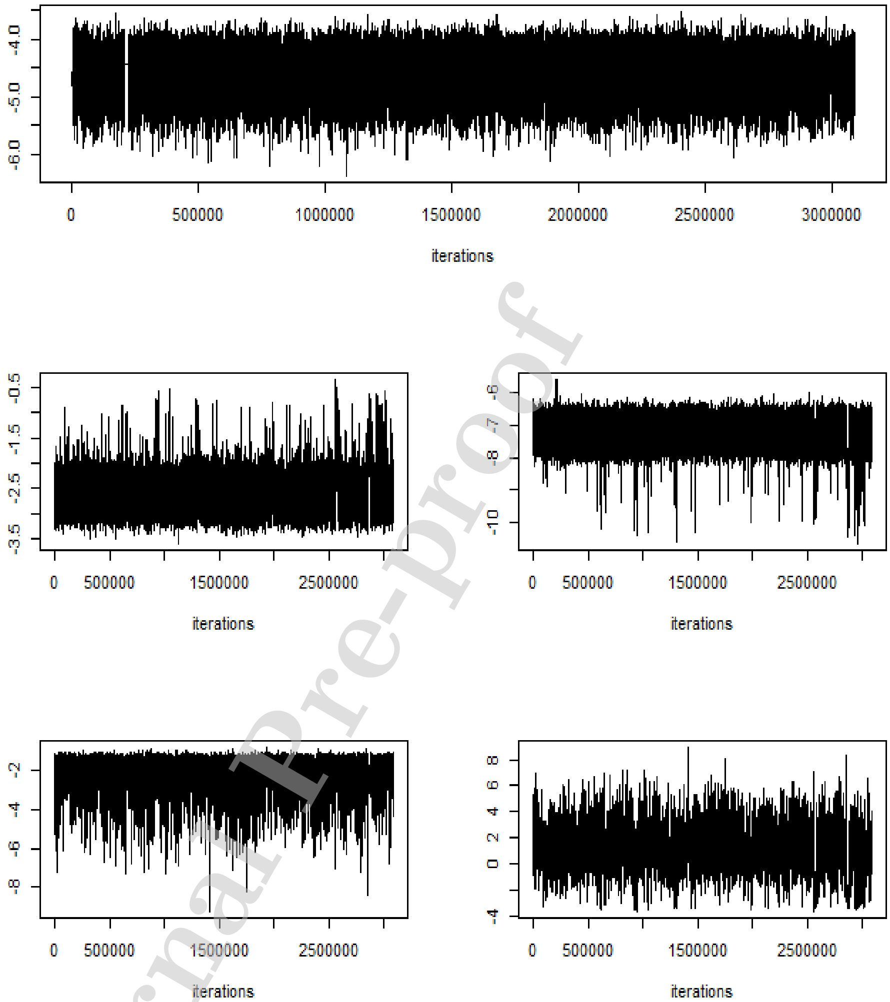


Figure 4: Trace plots for $\theta(1), \ldots, \theta(5)$.

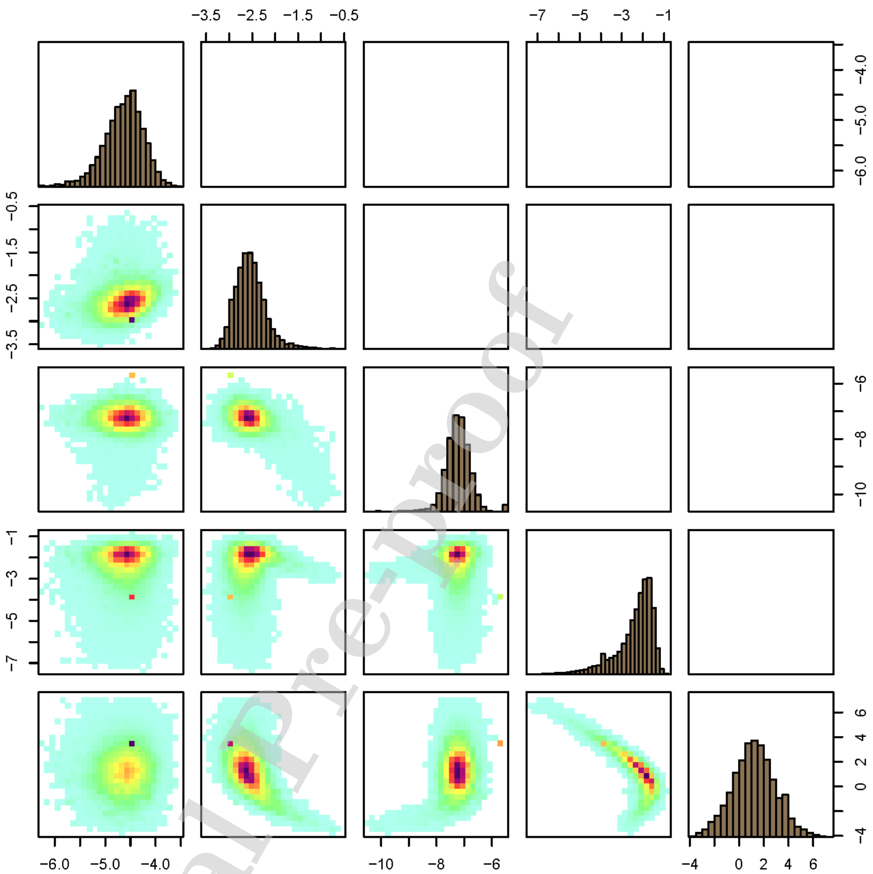


Figure 5: Posterior distributions (marginals and pairs) for $\theta(1), \ldots, \theta(5)$. From the BL model fitted to the Bass River data.

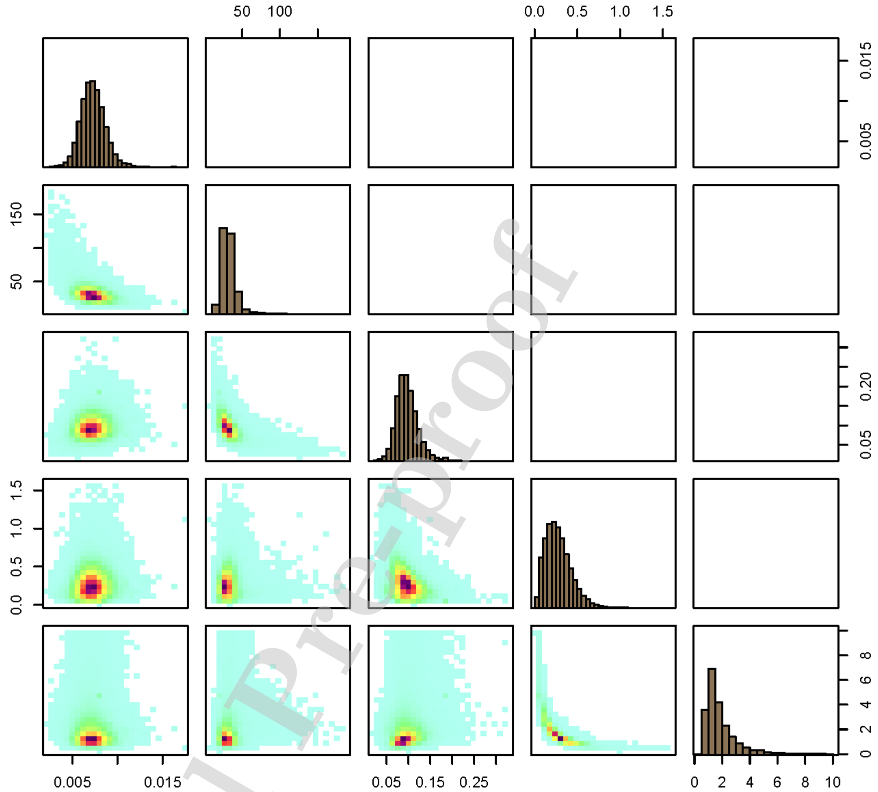


Figure 6: Posterior distributions (marginals and pairs) for λ, γ^{-1}, β, η^{-1} and μ_x. From the BL model fitted to the Bass River data.

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. From the modelling perspective, 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 $\{Y_i^h\}$, 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.

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

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, autocorrelation 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.

Finally we note that other simulation-based model fitting approaches are available, in particular the Simulated Method of Moments (SMM) of McFadden [17] and Pakes and Pollard [18]. The advantage of ABC over SMM is that it gives access to all the advantages of Bayesian modelling.

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.

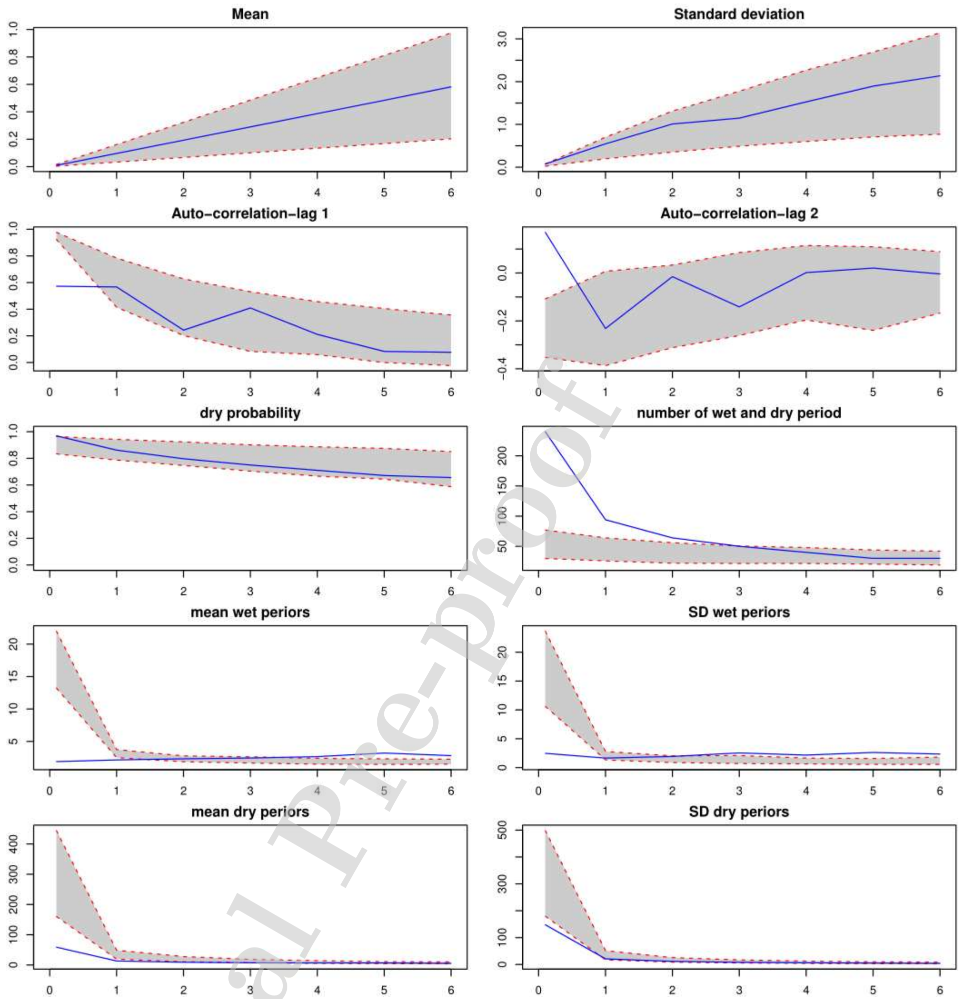


Figure 7: 95% Monte-Carlo prediction intervals from GMM estimates, for various statistics. In each figure, the statistics were calculated using rainfall aggregated over intervals of 0.1, 1, 2, 3, 4, 5, 6 hours, and were calculated from 3000 independent simulations. The solid blue lines give the observed statistics.

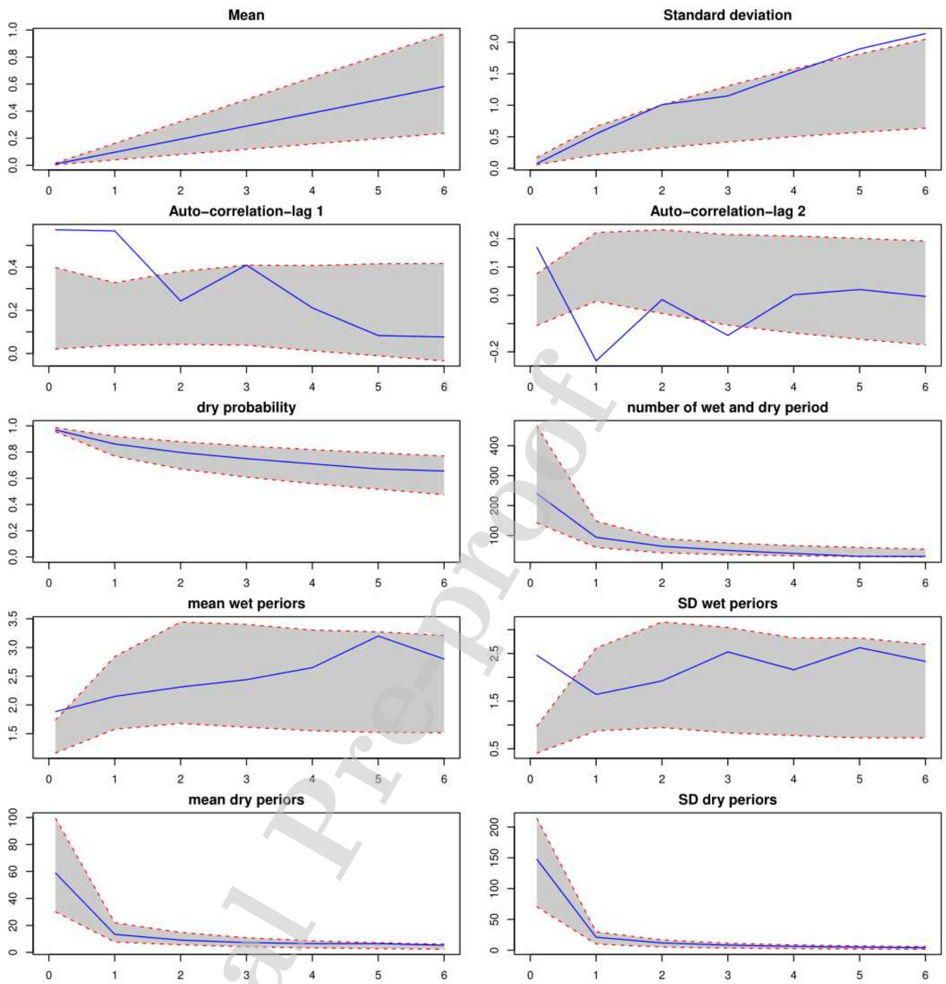


Figure 8: 95% Monte-Carlo prediction intervals from ABC-MCMC estimates, for various statistics. In each figure, the statistics were calculated using rainfall aggregated over intervals of 0.1, 1, 2, 3, 4, 5, 6 hours, and were calculated from 3000 independent simulations. The solid blue lines give the observed statistics.

[1] I. Rodriguez-Iturbe, D. Cox, V. Isham, Some models for rainfall based on stochastic point processes, Proc. R. Soc. Lond. A 410 (1987) 269–288.

[2] I. Rodriguez-Iturbe, D. R. Cox, V. Isham, A point process model for rainfall: further developments, Proc. Roy. Soc. London 417 (1988) 283–298.

[3] P. Cowpertwait, V. Isham, C. Onof, Point process models of rainfall: developments for fine-scale structure, Proc. R. Soc. Lond. A 463 (2007) 2569–2587.

[4] J. Kaczmarska, Further development of Bartlett-Lewis models for fine-resolution rainfall, Tech. Rep. 312, Department of Statistical Science, University College London (2011).

[5] J. Kaczmarska, V. Isham, C. Onof, Point process models for fine-resolution rainfall, Hydrological Sciences Journal 59 (11) (2014) 1972–1991.

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

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

[8] P. Marjoram, J. Molitor, V. Plagnol, S. Tavaré, Markov chain Monte Carlo without likelihoods, Proc. National Academy of Sciences 100 (2003) 15324–15328.

[9] S. Sisson, Y. Fan, M. Tanaka, Sequential Monte Carlo without likelihoods, Proc. National Academy of Sciences 104 (2007) 1760–1765.

[10] S. Sisson, Y. Fan, M. Tanaka, Correction for "Sequential Monte Carlo without likelihoods", Proc. National Academy of Sciences 106 (2009) 16889.

[11] M. Beaumont, J.-M. Cornuet, J.-M. Marin, C. Robert, Adaptive approximate Bayesian computation, Biometrika 96 (4) (2009) 983–990.

[12] D. Wegmann, C. Leuenberger, L. Excoffier, Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood, Genetics 182 (4) (2009) 1207–1218.

[13] M. Beaumont, W. Zhang, D. Balding, Approximate bayesian computation in population genetics, Genetics 162 (2002) 2025–2035.

[14] A. C. Aitken, On least squares and linear combination of observations, Proceedings of the Royal Society of Edinburgh 55 (1936) 42–48.

[15] S. A. Sisson, Y. Fan, Likelihood-free markov chain monte carlo, arXiv preprint arXiv:1001.2058 (2010).

[16] N. Aryal, O. Jones, Fitting the Bartlett-Lewis rainfall model using Approximate Bayesian Computation, in: G. Syme, D. H. MacDonald, B. Fulton, J. Piantadosi (Eds.), MODSIM2017, 22nd International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand, 2017, pp. 1579–1585.

[17] D. McFadden, A method of simulated moments for estimation of discrete response models without numerical integration, Econometrica 57 (1989) 995–1026.

[18] A. Pakes, D. Pollard, Simulation and the asymptotics of optimization estimators, Econometrica 57 (1989) 1027–1057.