Reply to anonymous referee #1

February 29, 2012

Reviewer comment: The study is limited inasmuch as it does not consider the fit of extreme values in the assessment. It would be interesting if the fit to extremes were to be included in the objective function. This would be computationally intensive and the efficiency of the algorithms would become more pertinent.

Reply: Including the fit of extreme values in the objective function would indeed be very interesting to consider. However, extreme values cannot be calculated analytically. A time series of sufficient length (or an ensemble of time series) would thus have to be simulated whenever the objective function is evaluated.
This will lead to a serious increase in calibration time for two reasons. First, simulating is simply much more demanding than solving mere analytical expressions to obtain the moments in the objective function. Second, we know that certain parameter combinations result in extremely long simulation times. For example, when κ/ϕ is large, large amounts of tiny cells are created during each storm. However, when one can use analytical expressions in the objective function, calculating the moments for this scenario will take just as long as any other parameter combination.
Furthermore, somehow including the extremes in the calibration process means dealing with multiple objectives, or, if the evaluation of the extremes is included into the same objective function, it would require significant research to fine-tune the weights in this function.
Finally, the suspicion that this would highlight the algorithms' efficiency might not be completely accurate. The increase in computational intensity results from a slower objective function evaluation. The algorithm is not bothered by the duration of a function evaluation. The differences in computational efficience that are shown in this paper will only become more apparent.

Reviewer comment: Another general criticism is that the authors do not mention spatial versions of the clustered point process rainfall models.

Reply: This is because the article only focuses on the temporal version of the Bartlett-Lewis model. However, we agree that the existence of such spatial versions should at least be acknowledged in the introduction by including some relevant references.

Reviewer comment: P9709,26 The parameter eta is not defined, and when it is introduced on P9710 it is as a random variable. Furthermore, the reason given for introducing dimensionless parameters is not convincing as eta cancels in the ratio (line 29).

Reply: Indeed, the first mentioning of η and of μ_x was forgotten by the authors. On P9710, line 1 should be: "... having a random depth and duration, both drawn from exponential distributions characterized by parameters $1/\mu_x$ and η respectively".
As for the introduction of the dimensionless parameters κ and ϕ. This will be changed to a mere mentioning of these dimensionless parameters, which are used as model parameters in the MBL model. The reason why this was done by the developers of the Bartlett-Lewis model is irrelevant to the subject of the article. It will be mentioned after the definition of η and μ_x.

Reviewer comment: P9710, I can only find 5 parameters. What about the distribution of cell depth?

Reply: See previous comment, the definition of μ_x will be added together with the definition of η.

Reviewer comment: P9711,26 I think that describing stochastic models as deterministic when there is no hyper-distribution for the parameters is misleading.

Reply: What we meant with this phrasing is that usually only one parameter set will be used for practical purposes. Therefore, it is treated as a fully identifiable model. We will rephrase this sentence.

Reviewer comment: P9714 ,5 How is the variance of the observed statistic calculated?

Reply: The observed moments, used in the objective function, are calculated for the entire dataset. Each of these moments is also calculated for each year separately, resulting in a series of 105 repetitions of that particular moment. The variance of these repetitions is calculated and used in the objective function. This will be explained in the text.

Reviewer comment: P9714,8 Chandler (2004) is a somewhat inconvenient reference for the standard result of generalised least squares.

Reply: "in least squares problems with unequal variances, observations should be weighed according to the inverse of their variances" is a quote that was taken directly from Chandler (2004). Therefore, this reference seems necessary.

Reviewer comment: Section 4 Implementation of the optimization methods
This is specific to the Uccle site. It may provide good advice for fitting the B-L model in general, but this is not demonstrated. In Figure 1 the cooling rate scale is incorrect. The grey scale doesn't work well on my copy, but the three white squares surrounded by dark squares look rather odd.

Reply: Strictly speaking, these results cannot be extrapolated to other datasets or to other models. However, the results show that convergence by the algorithm is reached at a fairly wide range of algorithm parameters. So, it would not be unreasonable to assume that the algorithms will perform well when applied to other datasets or Bartlett-Lewis models, if the algorithm parameters are chosen in these regions. So, to our opinion, these results can certainly be used as a guideline for other Bartlett-Lewis models.
The scale of Figure 1 will be corrected.
It is normal that the greyscale looks rather odd as it is purposely designed like this. This is to show that the minimum has been found for different parameter combinations, (i.e. the black squares = convergence in the minimum). The boundaries of these areas are in grayscale. This kind of detail would not have been possible without an adjusted colour scale since the obtained objective function values differ considerably.

Reviewer comment: Section 5 Comparison of optimisation methods
P9728,15 How do the 30 repetitions vary? I guess in the initial parameter sets, but details should be given.
Reply: The 30 repetitions do indeed differ in the initial parameter sets. Each initial parameter set is randomly chosen within the boundaries of the parameter space. More information will be added to the article.

Reviewer comment: In Table 3, the duration of DSM is an order of magnitude less than for the other optimisation methods. If multiple starting points are used, more such starts would likely lead to lower minima.

Using multiple starting points does indeed lead to better results for the DSM. Therefore, the calibrations were repeated using the DSM with a total of 30 starting points. This number of starting points leads to roughly comparable calibration durations, which enables the comparison of the performances of the different algorithms. Table 1 shows a clear improvement of performance. It can be seen that the DSM is now capable of locating the same minima as the other optimization methods. However, the medians of the obtained calibration results for the different objective functions are somewhat higher in comparison with the other optimization methods. Together with the very low standard deviation, this indicates that the DSM is very proficient in finding near-optimal solutions in a robust manner. The robustness of the DSM with multiple starting points is further demonstrated by Fig. 1. This figure demonstrates that the DSM measures up to the other optimization methods. However, judging by Table 2, it seems that SCE-UA still has the lowest mean rank, indicating that it provides the most accurate results, on average. The conducted statistical tests are also repeated, taking into consideration the results of the DSM with multiple starting points. The Kruskal-Wallis test results in a p-value of 1.3e-017. This means that at a 5% significance level there is a significant difference between the populations´ medians. The pairwise Wilcoxon rank sum tests, shown in Table 3, still point to significant differences between the different methods.
Clearly, using a multistart DSM definitely shows good results. Furthermore, this method has great merit from a practical point of view. Especially because the duration of the calibration is comparable to the presented alternatives. Therefore, the DSM with single starting point will be replaced by the DSM with multiple starting points in the article.

Reviewer comment: P9729,29 Why are identifiability issues important for a conceptual model? It wouldn't matter if a good fit were obtained with one or two parameters fixed in advance. P9730,1 A reference for "mentioned as a stumbling block in the literature" would be useful.

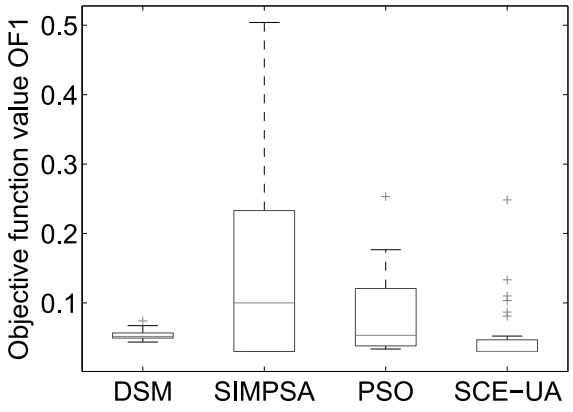
Reply: We fully agree with this comment. These remarks will be omitted from the article.

Reviewer comment: 6 Comparison of objective functions P9733, 12 How long are the simulations?

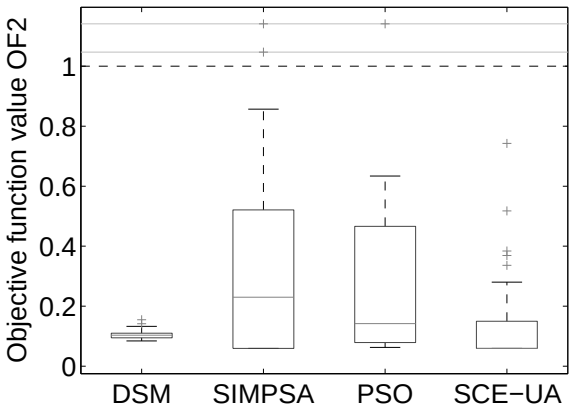
Reply: The simulations are each 105 years long, i.e. equal to the length of the observations.

Table 1: Descriptive statistics of the performance of different optimization methods for the calibration of the Modified Bartlett-Lewis Rectangular Pulses model

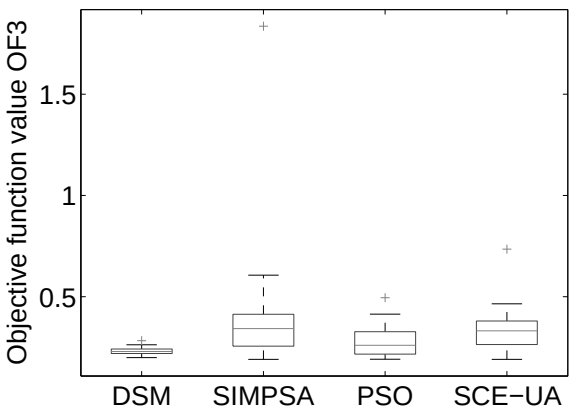
		minimum	median	StDev	duration (min)
DSM	OF1	0.0300	0.0467	0.0247	9
	OF2	0.0594	0.0922	0.0580	15
	OF3	0.1908	0.2063	0.0524	9
SIMPSA	OF1	0.0300	0.0300	0.4696	18
	OF2	0.0594	0.0594	18.2159	100
	OF3	0.1908	0.1908	0.6206	4
PSO	OF1	0.0283	0.0311	0.1531	13
	OF2	0.0594	0.0624	0.8168	31
	OF3	0.1908	0.1908	0.2359	14
SCE-UA	OF1	0.0300	0.0300	0.1150	14
	OF2	0.0594	0.0594	0.3552	24
	OF3	0.1908	0.1908	0.3433	11



(a) OF1



(b) OF2



(c) OF3

Figure 1: Box plots comparing calibration results of DSM, SIMPSA, PSO and SCE-UA. All months and repetitions are lumped together, no averages have been taken.

Table 2: Mean ranks of the different optimizaton method's performances

	DSM	SIMPSA	PSO	SCE-UA
Mean rank	2391	2085	2236	1929

Table 3: p-values for pairwise Wilcoxon rank sum tests between different optimization methods

	DSM	SIMPSA	PSO	SCE-UA
DSM	1	1.14E-08	0.0043	9.20E-18
SIMPSA		1	0.0059	0.0031
PSO			1	1.08E-08
SCE-UA				1