We would like to thank the referee for reviewing our manuscript and providing constructive comments. The several issues raised by the referee are addressed one by one below. In addition to the changes in direct response to the reviewers, there are some small edits in the other parts of the manuscript. For these we refer to the marked-up version of the manuscript contained in this file.

Comments from referee #1

General comments

(1) Referee's comment: The Authors propose a suitable procedure for selecting calibration events of a hydrodynamic model of a predominantly green urban catchment. A two-stage calibration procedure is used for calibrated first the parameters related to impervious areas, using a set of rainfall events, followed by the pervious area parameters using another set of rainfall events. The selection of calibration events was carried out based on some characteristics such as precipitation intensity, runoff flow rate, flow volume, flow volume as percentage of rain and precipitation duration. The overall ranking of the different calibration scenarios in the validation period is estimated using the statistics of both NSE (Nash-Sutcliffe Efficiency) and RMSE (Root Mean Square Error). The paper address scientific questions within the scope of HESS even if it does not present new concepts or ideas but a rather useful procedure. The scientific methods and assumptions are clearly outlined and the overall presentation is well structured and clear.

Authors' response: we thank the referee for their generally supportive comments on the manuscript. Two points require some clarification:

First, we would like to clarify that the model performance in the validation period is also assessed on the basis of the flow volume error and the peak flow ratio, and these two statistics are actually the ones, for which some of the most notable differences between different calibration scenarios are visible.

Second, we believe that the suggestion that our paper "…does not present new concepts or ideas…" is open to discussion. In fact, the novelty of the paper consists in developing a calibration / validation procedure for a green urban catchment (i.e., with predominantly pervious areas). This is a unique class of catchments, because such green areas may receive water from both rainfall and runoff from adjacent impervious areas, and the resulting runoff is fed into a hydraulically efficient transport network of storm sewers. Although the ideas of using different types of calibration events and two-stage calibrations are not new in general (as pointed out by the referee or in our introduction section), we believe that the methodology for selection and execution of such procedures has not been addressed explicitly in any published article on urban drainage modelling, and the articles that may peripherally touch upon these issues are generally focused on urban drainage systems, of which runoff is controlled by impervious areas. Although the modelling of urban drainage has commonalities with general hydrologic modelling, there are also some key differences and even differences from modelling conventional urban catchments (Elliott and Trowsdale, 2007; Fletcher et al., 2013). Consequently, specific findings from modelling natural catchments or modelling conventional urban catchments dominated by directly connected impervious surfaces cannot be assumed to apply to green urban catchments without some caveats. This point can be clarified in the introduction.

Changes in manuscript: add on page 2, end of line 18: This second aspect also applies to investigations into other sources of uncertainty in urban drainage modelling, some of which have been investigated before, e.g. input and calibration data uncertainties (Dotto et al., 2014; Kleidorfer et al., 2009a) and spatial model resolution (Krebs et al., 2014; Petrucci and Bonhomme, 2014; Sun et al., 2014). However, these investigations used predominantly impervious catchments and it is, therefore, unknown to what extent their findings apply to greener urban catchments as well and how sensitive such results are to the calibration data set that was used.

Add to the objective statement at the end of the introduction: Two secondary objectives are to verify: (1) the findings from previous urban drainage modelling studies on a greener (less impervious) catchment, and (2) sensitivity of the earlier findings to the calibration data used.

Specific comments

(2) Referee's comment: While the calibration strategies (single- and two stage) was already presented by the Authors in a previous paper (1), the different metrics for selecting calibration events from a larger group of candidate events is rather innovative and well described.

Authors' response: we thank the referee for their supportive comment.

Changes in manuscript: none needed

(3) Referee's comment: The risk of using rainfall multipliers is to attribute to the rainfall all the errors due to an incorrect estimate of the model parameters as well as of the model itself. The Authors indicate that the rainfall multipliers compensate for discrepancies between the observed and best-fitting rainfall, rather than for other aspects of catchment runoff modelling by using the baseline model but it is not clear how they reach this conclusion.

Authors' response: two arguments support that the rainfall multipliers appear to work as intended, i.e. to compensate for a mismatch between observed rainfall and the rainfall that fits best with the observed outflow:

1. While there is high variability among the obtained parameter values between different calibration scenarios (CSs), there is little variability among the rainfall multipliers for each event as obtained by different CSs. If the multipliers had the effect of compensating for e.g. reduced runoff volumes caused by higher infiltration (e.g. if the calibration parameter saturated hydraulic conductivity was higher), then it would be expected to see inter-CS variation in the rainfall multipliers more similar to that found for the other model parameters.
2. When rainfall input was perturbed by -40% and +40% the rainfall multipliers changed by -37% and +33% respectively. The similarity shows that the rainfall multipliers are sensitive to mismatches between the observed and best-fitting rainfall volume.

Section 3.2.2 of the manuscript can be reorganized to present these arguments more clearly.

We would also like to point out that many studies include the catchment area or imperviousness (defined here as the 'directly connected' imperviousness) as a way of adjusting flow volumes. The calibrated area or imperviousness obtained from this will also be affected by the observed vs. best fitting rainfall mismatch. Since high-quality land cover information and field visits were used for catchment delineation in this study, we preferred not to further calibrate the catchment size parameter, so as to maintain its clear physical connection to the real system. It was still thought that a mismatch between observed and best-fitting rainfall could be present. Since the other hydrological model parameters (listed in Table 1 and Table 2 of the manuscript) do not have a large effect on the runoff volume, the rainfall multipliers presented a way of accounting for this mismatch.

Changes in manuscript: Reorganize the text in section 3.2.2 so that arguments 1 and 2 above are clearly identifiable as support for the conclusion on the rainfall multipliers. That the role of rainfall multipliers to adjust overall volumes is sometimes filled by calibrating catchment area is already mentioned in section 2.2 (p. 4 line 34 – p. 5 line 2), but the desired effect of maintaining the connection between physical and model catchment size will be added to this sentence.

(4) Referee's comment: It should however be considered that rainfall multipliers tend to treat the spatial variability of rain, which has a dynamic effect on the outflow, through a positive or negative variation of rainfall considered uniform on the single watershed and therefore treated in a static way.

Authors' response: We are fully aware of this issue, but with only one rain gauge and flow sensor being available (and lacking other information on the spatial variability of the rainfall and/or the effect of moving storms in the relatively small study area of 10.2 ha) there was no feasible alternative to assuming uniform rainfall over the catchment. Consequently, treating the rainfall error as being constant over the catchment seems fitting.

Changes in manuscript: add on page 4, line 30: Rainfall multipliers also do not address the spatial variability of the rainfall, but given the lack of multiple gauges or other information about the spatial variability of rainfall in the catchment no clear alternative was available.

(5) Referee's comment: Figure 5 is not clear and should be conceived in a new way.

Authors' response: We understand that the figure may be somewhat difficult to interpret, but also like that it contains a lot of information in a small amount of space. However, improvements can be made in several ways: (i) better explanation of the figure; (ii) better labelling of the different numbers in the figure, addition of units for parameter values, and more explanation in the figure caption can be added; or (iii) splitting the figure into multiple vertically aligned panels, each showing a subset of the calibration scenarios.

Changes in manuscript: new version of figure 5

Technical corrections

(6) Referee's comment: Table 8 does not contain bold characters as indicated in the text

Authors' response: the bold font was inadvertently left out.

Changes in manuscript: bold font will be added in the table to indicate the best value in each column. This will be done in all tables concerning calibration or validation performance.

We sincerely thank the referee for their extensive comments on the manuscript, which we reply to point-by-point below. The referee's comments have been numbered for easy reference. In addition to the changes in direct response to the reviewers, there are some small edits in the other parts of the manuscript. For these we refer to the marked-up version of the manuscript contained in this file.

Comments from referee #2

General comments

(10) Referee's comment: This manuscript presents an analysis of the impact of selecting different sets of calibration data for the SWMM urban hydrological model. Selection is based on a variety of hydro-meteorological characteristics of the available storm events. In addition, the calibration is performed either adjusting all calibration parameters simultaneously, or at two stages where parameters related to pervious and impervious areas are calibrated separately. Finally, the results are analyzed against a backdrop of other sources of uncertainty besides the calibration dataset.

Authors' response: this summarizes well the contents of the manuscript.

Changes in manuscript: -

(11) Referee's comment: The idea of calibrating impervious area parameters separately using such data where the role of previous areas is presumably insignificant is promising, and in my opinion the results related to this represent the most valuable contribution of the present manuscript. On the other hand, I struggle to find a novel scientific contribution in the analysis of the calibration event selection in combination with other causes of uncertainty. As argued in the specific comments below the results are inconclusive and it is hard to find any other take-home message than the fact that selection of calibration data has an impact on model parameter values and model performance. This has been established already in existing hydrological literature, as acknowledged also by the authors themselves.

Authors' response: In general, the novelty of our paper consists in: (i) drawing attention to the calibration/validation issues in green urban catchments, and (ii) proposing a calibration / validation procedure for a green urban catchment (i.e., with predominantly pervious areas). This is a unique class of catchments, because such green areas may receive water from both rainfall and runoff from adjacent impervious areas, and the resulting runoff is fed into a hydraulically efficient transport network of storm sewers. Consequently, specific findings from modelling conventional urban catchments dominated by directly connected impervious surfaces cannot be assumed to apply to green urban catchments without some caveats.

Further documentation of innovative aspect of our work is presented in our response to points 17-21, 24, 28 and a newly added emphasis on innovative aspects of our manuscript.

Changes in manuscript: see below (points 17-21, 24, 28)

(12) Referee's comment: The readability and the quality of the English language are at a very good level.

Authors' response: we thank the referee for their supportive comment.

Changes in manuscript: none needed

Specific comments

Study site and data

(13) Referee's comment: It would be useful to show somewhere a brief summary of the storm events (e.g. duration, cumulative rainfall depth, cumulative runoff, peak runoff, runoff percentage). The runoff percentage in particular would be interesting as it is used in selecting events for the two-stage calibration. Also, it would be interesting to see to which extend the permeable areas are activated during more intensive events (i.e. runoff-% > 12%).

Authors' response: we agree that a table summarizing rainfall-runoff events could be useful to the reader. The table can also contain a rough estimate of how many mm runoff was generated by the green areas. The extent to which green areas are activated can be estimated in a limited way from the data directly (see last column in table C1 in the supplement).

Changes in manuscript: such a table will be added in the methods section, see Table C1 in the supplement.

Table C1: characteristics of all rainfall events used in one or more calibration scenarios.

Event #	Precipitation sum in preceding 72 hr	Precipitation sum (P_sum)	Precipitation duration (D_prec)	Average precipitation intensity (PI_mean)	Highest 30-minute average precipitation intensity (PI_30m)	Runoff volume (QV)	Percentage runoff (QV_ppP)	Peak flow rate (Q_max)	Highest 60-minute average flow rate (Q_60m)	Runoff from green areas [b]	Of which originating from impervious areas [c]	Originating from green areas [d]	Average percentage runoff from green areas [e]
	mm	mm	hr	mm/hr	mm/hr	mm	%	L/s	L/s	mm	mm	mm	%
199	2.4	13.8	41.6	0.3	4.0	1.7	12.4	4.2	3.3	0.06	0.02	0.04	0.3
209	0.2	8.0	9.5	0.8	2.8	0.5	6.9	4.5	2.7				
211	8.3	9.7	22.8	0.4	6.9	1.1	11.1	29.2	11.1				
214	7.3	6.4	12.1	0.5	4.3	0.6	10.1	40.5	8.5				
222	1.1	9.8	12.8	0.8	7.5	0.7	7.2	26.4	13.3				
270	0.0	9.3	38.5	0.2	3.5	1.1	11.3	22.9	8.7				
306	10.1	8.6	9.1	0.9	7.1	0.7	8.5	27.5	9.3				
307	18.3	29.9	37.7	0.8	8.5	4.9	16.2	71.2	42.9	1.27	0.36	0.91	3.0
310	12.7	8.6	10.0	0.9	7.5	1.2	14.0	37.4	17.4	0.17	0.05	0.12	1.4
530	13.8	6.7	2.8	2.4	7.2	0.8	11.2	58.9	13.5				
939	0.6	7.0	25.6	0.3	1.0	0.4	5.7	2.1	1.8				
962	0.0	8.5	11.2	0.8	1.4	2.1	24.9	4.9	4.4	1.09	0.31	0.78	9.2
971	0.2	2.6	18.6	0.1	1.1	0.3	11.3	4.0	2.9				
978	12.7	25.0	65.8	0.4	5.8	4.8	19.1	64.5	16.6	1.77	0.50	1.27	5.1

982	0.0	5.6	3.4	1.7	7.0	0.9	15.8	49.5	17.2	0.21	0.06	0.15	2.7
984	13.1	2.4	6.3	0.4	4.6	1.4	59.1	71.7	14.0	1.12	0.32	0.80	33.7
995	4.8	2.1	8.5	0.2	1.8	0.6	28.6	32.0	9.7	0.35	0.10	0.25	11.9
997	2.2	24.6	49.0	0.5	2.4	5.1	20.7	15.0	6.9	2.14	0.61	1.53	6.2
1001	0.0	35.3	56.6	0.6	8.6	8.8	25.0	56.5	32.5	4.58	1.30	3.28	9.3
1004	22.5	4.2	13.9	0.3	5.9	1.1	25.2	33.3	10.6	0.56	0.16	0.40	9.5
1019	0.5	22.3	49.7	0.4	2.3	4.7	21.2	12.9	9.3	2.06	0.58	1.47	6.6
1028	6.2	2.8	7.0	0.4	1.3	1.2	43.5	6.3	4.2	0.89	0.25	0.64	22.5

a Calculated assuming 100% runoff from impervious areas: a = QV – 0.12 P_sum, where 0.12 is the percentage of directly connected impervious area. (Some of this runoff originated from impervious areas that drained to green areas).
b Calculated as b = a (25 / (25+63)), where 25 and 63 are the percentages of indirectly connected impervious surfaces and green surfaces respectively.
c Calculated as c = a – b
d Calculated as d = c / P_sum

Event selection

(14) Referee's comment: To me the most promising aspect of this manuscript lies in the idea of calibrating parameters related to pervious and impervious areas separately. It is obvious that with a greater runoff percentage than 12% other than just directly connected areas need to contribute. For events with less than 12% runoff it is not equally evident that ONLY directly connected areas contribute. Still, this is a feasible assumption and probably holds to a sufficient extent. There is ample evidence that in urban setting for small events (directly connected) impervious areas predominantly contribute to stormwater flow and for major events also permeable areas are activated.

Authors' response: it can indeed be the case that green areas contribute some runoff even when the percentage runoff is less than 12%. Even impervious surfaces will not generate 100% runoff, so if runoff is exactly 12% it is reasonable to expect that at least a small part of runoff has come from green areas instead. We agree with the referee that the amount of runoff from green areas is small enough that assuming it zero is a feasible assumption. In any case, it would be difficult to determine by how much the 12% threshold should be lowered to ensure that no green area runoff is included, since this would also depend on the antecedent conditions in the catchment. Given a lack of other measurements (e.g soil moisture, standing water in swales) in the catchment it is not possible to tell the initial wetness of the catchment from measurements. Estimating initial conditions using the model itself would lead to the undesirable situation where the value of the threshold (and therefore potentially the set of events to use) would be different for each model run. A fixed percentage is therefore much more workable and probably of more practical use.

Changes in manuscript: add the following sentence in section 2.3 on line 11: "(It is conceivable that there is some contribution of green areas when the percentage runoff is less than 12%, and in that case the threshold should be set at a lower value, but since the amount of green area runoff and the appropriate value of the threshold would be highly dependent on antecedent conditions this was not included here.)"

(15) Referee's comment: A couple of issues require further clarification. Did you check whether in the model any runoff was generated from permeable areas when the runoff-% was below 12%? If it is argued that no runoff is produced outside of the (directly connected) impervious areas for low runoff-% events it should be checked that the model result is consistent with this assumption.

Authors' response: there are several items to check here:

First, during the first stage calibration (i.e. with default values for green area parameters) there was no runoff from green areas for any of the calibration events in any of the calibration scenarios, and so the first stage calibration attributed all runoff to impervious areas.

Second, using the calibrated parameter values for both impervious and green areas, there were some first-stage events where some runoff was predicted from green areas:

1. When runoff was disabled from both directly and indirectly connected impervious areas /by setting their depression storage to 1000 mm) there were three calibrated models runs (2 for T32S_D_prec, 1 for T32s_Q_60m) that actually generated some runoff from green areas (i.e. the runoff did not originate on impervious areas draining to green areas), but since this was ≤2% of the total simulated runoff volume this was considered negligible.
2. When runoff was disabled only for directly connected impervious areas, a total of 12 calibrated model runs showed non-zero runoff from green areas. This was <5% of total simulated runoff volume for 4 runs, <10% for an additional 3 runs, and 11.6%, 11.7%, 21.7%, 22.9% and 25.7% respectively for the remaining 5 runs. However, almost all of this was runoff that was generated on impervious areas draining onto green areas (see point 1 above).
   Regarding the last mentioned 5 runs, it should be noted that these concerned 3 different events with a percentage runoff between 11% and 12%. Such events may be expected to include some green area runoff and it could be considered to exclude these from the first stage calibration as discussed in comment #14. In addition, all three events were also included in other first-stage calibrations that did not result in any significant simulated green area runoff (0, 0 and 3.4% of total simulated runoff, respectively). Removing these events from the first stage of calibration based on initial calibration results would therefore result in the same event being included in different stages for different calibration scenarios, which we considered undesirable.
   Overall we believe that, although the assumption that all runoff is from directly connected impervious areas when QV_ppP <12% is violated in some cases, the assumption that these events are suitable for calibrating impervious area parameters does hold to a sufficient degree, as also evidenced by the good first-stage calibration performance (mentioned on p 10, l. 2-3). In addition, checking for green area runoff as done here is only possible after calibration, and taking it into account when selecting events would thus create a more complex, iterative calibration procedure which limits the practical applicability of the approach. We considered this to be beyond the paper's original scope of examining different strategies for calibration event selection. It could however be considered as a potential avenue for further research on multi-stage calibration procedures.

Changes in manuscript: add a (shorter) version of our response above to section 3.1.1: For the two-stage calibrations the assumption that no runoff occurred from green areas during the first stage of the calibration was checked. During the actual 5 first-stage calibration (i.e. with green area parameters set to default values) there was no runoff from green areas for any of the calibration events in any of the calibration scenarios, so the first stage calibration attributed all runoff to impervious areas as assumed beforehand. However, some runoff occurred from green areas for first-stage events when the calibrated parameter values from the second stage were applied. This runoff was caused by impervious areas draining to green areas. The runoff from green areas was <5% of the total simulated runoff volume for 4 model runs, <10% for an additional 3 runs, and 11.6%, 11.7%, 21.7%, 22.9% and 25.7% respectively for 5 additional runs. These last 5 runs concerned 3 different events with a percentage runoff (calculated before applying rainfall multipliers) between 11% and 12%. Such events may be expected to include some green area runoff and it could be considered to exclude these from the first stage calibration (not done here to limit the complexity of the procedure as discussed in Sect 2.3). In addition, all three events were also included in other first-stage calibrations where they did not result in any significant simulated green area runoff. Removing these events from the first stage of calibration based on initial calibration results would therefore result in the same event being included in different stages for different calibration scenarios, which was considered undesirable. Overall we believe that, although the assumption that all runoff is from directly connected impervious areas when QV_ppP <12% is violated in some cases, the assumption that these events are suitable for calibrating impervious area parameters does hold to a sufficient degree, as also evidenced by the good first-stage calibration performance (see first paragraph of this subsection). In addition, checking for green area runoff as done here is only possible after calibration, and considering it when selecting events would thus create a more complex, iterative calibration procedure, which would limit the practical applicability of this approach. We considered this to be beyond the paper's original scope of examining different strategies for calibration event selection.

(16) Referee's comment: Second, the large range of rainfall multipliers (0.48 – 2.92) can make determining the runoff-% somewhat ambiguous. Presumably, the 12% runoff threshold was based on the measured values of precipitation and discharge before applying the rainfall multipliers. Did it happen that a smaller than the unity rainfall multiplier changed the initially below 12% runoff event to exceed the 12% threshold after rainfall multiplier calibration? If yes, should such an event be included in the first stage calibration?

Authors' response: the 12% runoff threshold was indeed applied directly to the measured values of precipitation and discharge.

There were two events where the rainfall multiplier was less than 1 and reduced rainfall so that the new percentage runoff exceeded 12%. This can also be displayed in an extended version of Table 4 from the manuscript, see Table C2 in the supplement. It is of course possible to exclude such events from their respective stages in the calibration and replace them with another event. Being consistent about considering the percentage runoff as calculated using the calibrated rainfall multipliers would also require the following three adjustments as well:

1. It would have to be applied 'in both directions', i.e. second-stage calibration events where the calibrated multiplier was large enough that runoff % was reduced below 12% would have to be excluded from the second stage. (This was the case for two events. For these events they would first have to be considered as replacement for a first-stage event, and the first stage calibration re-run, before redoing the second stage of the calibration. (Depending on the results from this the whole procedure might have to be repeated as well.)
2. All event characteristics related to rainfall (i.e. P_sum, PI_mean, PI_30m, QV_ppP) would have to be re-calculated and the related CSs determined and run again if the event set changed.
3. Out of the 32 events that were available for use in calibration scenarios, only 22 were actually selected by one or more CSs, so calibrated multipliers are not available for the other

10 events. It would be necessary to somehow obtain a calibrated multiplier value for them too so that they may be reconsidered for use in the calibration.

Although this might improve the overall results of the proposed calibration procedure, it would also increase the complexity and raise several new issues, such as how to obtain a calibrated rainfall multiplier for the 10 events that have not yet been used. We considered this to be beyond the paper’s original scope of examining different strategies for calibration event selection. It could however be considered as a potential avenue for further research on multi-stage calibration procedures.

Changes in manuscript: clarify that event selections were fixed beforehand and not adjusted based on initial calibration results. Add a short version of the explanation above in section 2.3, page 6, end of line 18: Applying the calibrated rainfall multipliers in the calibration (Sect. 2.2) means that event properties relating to rainfall and percentage runoff will change, and the percentage runoff can change from <12% to >12% and vice versa. Doing this consistently for all events in the calibration procedure would require (1) re-calculating which events should be available in each stage, (2) estimating in some way rainfall multipliers for all events, including those not initially selected by any calibration scenario, (3) re-calculating which events are used in each CS, and (4) repeating the calibration for any CS that has had any of its events changed. Although this might improve the overall results of the proposed calibration procedure, it would also increase the complexity and raise several new issues, such as how to obtain a calibrated rainfall multiplier for the 10 events that were not used in any CS. We considered this to be beyond the paper’s original scope of examining different strategies for calibration event selection and proposing a practically useable two-stage calibration procedure.

Table C2: calibrated rainfall multipliers and new percentages runoff.

Event #	N_T6	T32S_D_prec	T32S_P_sum	T32S_PI_mean	T32S_Q_60m	T32S_Q_max	T32S_QV_ppP	T6_D_prec	T6_P_sum	T6_PI_30m	T6_PI_mean	T6_Q_60m	T6_Q_max	T6_QV_ppP	Mean	New P	New QV_ppP	Swap stage
199								0.58	0.58						0.58	8.0	**21.4**	
209				0.48							0.48				0.48	3.8	14.3	gray > green
211		0.70	0.70		0.70	0.70									0.70	6.8	**15.8**	gray > green
214						1.16									1.16	7.4	8.7	
222			0.68		0.68					0.68					0.68	6.7	10.6	
270		1.24	1.22				1.28	1.26							1.25	11.7	9.1	
306				0.74						0.70	0.74				0.73	6.3	11.7	
307	1.48		**1.46**	1.48	1.48	1.48			1.48	1.44	1.44	1.52	1.48		1.47	44.0	11.0	green > gray
310				1.06	1.06					1.06	1.06	1.14			1.08	9.2	13.0	
530	1.14			1.10	1.10	1.12	1.04			1.08	1.08		1.14		1.10	7.4	10.2	

939		0.60													0.60	4.2	9.5	
962														0.98	0.98	8.3	**25.4**	
971							1.08								1.08	2.8	10.4	
978	1.38	1.38	**1.34**			1.34		1.40	1.42			1.36	1.38		1.38	34.4	13.9	
982	1.22			1.20							1.26	1.22	1.26		1.23	6.9	12.8	
984						2.02	1.94					2.12	2.00	1.90	2.00	4.8	**29.6**	
995							2.92							2.88	2.90	6.1	9.9	green > gray
997								1.24	1.26						1.25	30.8	16.6	
1001	1.70	1.66	**1.60**		1.64			1.66	1.66	1.60		1.64	1.70	1.64	1.65	58.2	15.1	
1004														0.78	0.78	3.3	**32.3**	
1019	1.46	1.48						1.46	1.44						1.46	32.6	14.5	
1028							1.30							1.30	1.30	3.7	**33.4**	

Other sources of uncertainty

(17) Referee's comment: The reasoning in including some of the uncertainty sources while leaving others out is not quite clear to me. Also, the take-home what readers should learn from this exercise should be clarified.

Authors' response: some of the issues described have been investigated before for urban drainage models (e.g. data uncertainties by Kleidorfer et al. (2009) and Dotto et al. (2014), model resolution by e.g. Krebs et al. (2014), Petrucci and Bonhomme (2014), Sun et al. (2014) and Tscheikner-Gratl et al. (2016)). The idea behind including other sources of uncertainty was (primarily) to see if different calibration event sets showed different sensitivity to these issues and (secondarily) to see if the findings also applied to a different data set and catchment (more dominated by green areas).

Although we considered it an interesting experiment at the time, the impact of what objective function is used in calibration of urban drainage models has not been investigated extensively before (Barco et al. (2008) made some short remarks), so we would remove this from the manuscript. (A thorough investigation of this would be an interesting topic for a different study.) However the different objective functions used for validation phase (e.g. volume error, peak flow) would still be included since they provide additional insight into the simulation results.

Removing the parts on objective function would also allow to describe in more detail the effect of the model resolution, since it is an interesting finding that some of the benefits of the two-stage calibration (better flow volume and peak flow in validation phase) are stronger for the low-resolution model.

The take-home messages from this are:

1. The impact of perturbed calibration data appears small (confirming the findings by Dotto et al. (2014)), but we do see interaction between the calibration data selection and the model discretization.
2. The two-stage calibration gives better results in terms of flow volume and peak flow in the validation phase, and this effect is much stronger for the low-resolution models.

Changes in manuscript:

1. Section 2.4 (other sources of uncertainty):
    a. Describe the aim of including other sources: i.e. check if earlier findings are sensitive to different calibration data sets and if they also apply for a different data set and a greener catchment.
    b. Add references to previous studies on rainfall input uncertainty effect on urban drainage modelling in lines 10-12 (Dotto et al., 2014; Kleidorfer et al., 2009).
    c. Lines 20-24: describe a bit more the Dotto and Kleidorfer papers that are referred to, including that they used more pipe-based drainage systems and a fixed set of events.
    d. Lines 28 and further: add references to articles dealing with model resolution (Krebs et al., 2014; Petrucci and Bonhomme, 2014; Sun et al., 2014; Tscheikner-Gratl et al., 2016)
2. Remove the parts that deal with the calibration using RMSE as alternative objective function:
    a. page 7 lines 25-27
    b. page 9 lines 2-5
    c. Section 3.1.2, including table 3.
    d. Table 5 column 3: "RMSE as obj. func." + update column "total"
    e. Section 3.3.3, including table 7 and figure 8.
    f. Conclusion page 24 lines 10-11
    g. mention in abstract

Rainfall input

(18) Referee's comment: The authors report that reducing flow measurements by 40% leads to 37% reduction in the mean value of rainfall multipliers, and increasing flow measurements by 40% results in a 33% increase in the rainfall multiplier mean value. This seems like rather a trivial result. A more justified description about the purpose of scaling the discharge by a constant multiplier, which causes a corresponding change in the rainfall depth scaling parameter, is needed.

Authors' response: since flow data is obviously an important part in the calibration process, we wanted to see if earlier results from Dotto (2014) and Kleidorfer (2009) would be sensitive to different sets of calibration events. Our findings mainly confirm their work. For urban catchments these issues have only been investigated to a limited extent (i.e. with a single set of events and for rather impervious catchments) so additional support of earlier findings is useful. Other disturbances of / errors in the calibration are conceivable, but were deemed beyond the scope of this study.

In addition, the correlation between the adjustment in rainfall and the adjustment in rainfall multipliers also supports the idea that the rainfall multipliers are compensating (even in the baseline run) for a mismatch between observed and best-fitting rainfall (as discussed in section 3.2.2), and therefore that they are a suitable way of accounting for this mismatch.

A better description of why this aspect is considered is also addressed in our response to comment 17 above.

Changes in manuscript: see above.

Calibration data measurement uncertainties

(19) Referee's comment: See comment above.

Authors' response: See response above.

Changes in manuscript: See response above.

Conceptualization / model discretization

(20) Referee's comment: While I agree that SWMM is a well established model for urban drainage I do not think that its applicability to areas clearly dominated by pervious areas is equally evident. Presumably in the SWMM runs of the current manuscript the groundwater module has been turned off and infiltration is based on the Green-Ampt equation with infiltration continuing with a rate appraoching assymptotically the hydraulic conductivity value. It can be questioned whether this is realistic for longer storm events when the soil becomes more saturated. Transpiration is also not accounted for but evaporation only occurs from the depression storage. I am not suggesting that it would feasible to take into account all aspects related to modelling uncertainty. But in my mind the authors' statement "… it is safe to assume that the SWMM conceptualization is appropriate for urban drainage modelling and there was no need to consider this issue further" is in the context of such a low density urban area questionable and does not constitute a valid argument for making a choice about which uncertainty sources are included/excluded in/from the analysis.

Authors' response: we appreciate the distinction between the application of SWMM to pervious and impervious areas. It is correct that the groundwater module in SWMM was not utilized in this study, and that therefore only the Green-Ampt equation + drying is used to account for infiltration. As pointed out by the referee, recovery of the infiltration capacity is not based on evapotranspiration, but is instead based on the soil's saturated hydraulic conductivity (Rossman and Huber, 2016).

Our original formulation was perhaps too optimistic, but we still believe that it is reasonable not to treat model structure as an uncertainty source in this article for the following reasons:

1. Unlike input and calibration data and model resolution, model structure uncertainty has not been addressed extensively in the urban drainage modelling literature.
2. There is a lack of methods for considering model structure uncertainty other than using different models, which is outside the scope of this study. The catchment and the high-resolution model also require certain features (e.g. routing runoff from one subcatchment to another subcatchment, support for automated runs) that are present in SWMM and not in other models. Model runtime is also a limiting factor.
3. The Green-Ampt method itself has been in use for many years. Other infiltration models are available (e.g. Horton or SCS curve number in SWMM) but going into these would be outside the scope of this study. Ideally a study on infiltration models in urban drainage modelling would also make use of infiltration and/or soil moisture measurements which are not available here.

Changes in manuscript: Replace p8, lines 5-7 with: Although model structure is also a recognized source of uncertainty (Deletic et al., 2012), it was not considered here since (a) there is a lack of previous research on this topic for urban drainage modelling that could be referred to and (b) there is a lack of methods to address this other than using different models in parallel, which was considered outside the scope of this study, and would in any case be difficult since the catchment model requires some SWMM features (e.g. routing runoff from one subcatchment to another, good support for automated runs) which are not always present in other models.

Calibration algorithm

(21) Referee's comment: The authors state that SCE-UA "… has been widely applied in hydrological applications with great success, so there was no need to subject it to scrutiny in this paper." While I agree that SCE-UA is a powerful tool with an extensive pool of hydrological modelling applications, it is not a sound, objective argument for leaving it out of study. The authors themselves admit that calibration against RMSE can yield a higher NSE than calibration against NSE itself, indicating that the algorithm does not always converge to the optimum value.

Authors' response: in relation to the improved description of why different sources of uncertainty are included it's good to mention that (like for objective functions) there is a lack of studies examining the effect of calibration algorithms on urban drainage modelling (Deletic et al., 2012). (And even to some extent in general hydrology (Houska et al., 2015)). A thorough examination of the effect of the calibration algorithm would require implementing many different algorithms. Since there is a lack of earlier studies here to refer to it is acceptable to leave the calibration algorithm out altogether.

Changes in manuscript: the statement on the exclusion of the calibration algorithm as source of uncertainty (page 8, lines 4-7) will be rephrased: Likewise, the calibration algorithm (Deletic et al., 2012; Houska et al., 2015) and numerical issues (Deletic et al., 2012; Kavetski 5 et al., 2006) are recognized as sources of uncertainty, but there is a lack of studies addressing these specifically for urban drainage modelling that could be referred to here. Since breaking new ground in these areas was considered beyond the scope of this paper, these sources of uncertainty are not considered here.

Validation performance

(22) Referee's comment: Validation performance should be the main argument for improved calibration strategy. If a calibration strategy leads to improved parameter identifiability this should be visible in better results against independent validation data. The authors state that "the two calibration strategies that performed best in the validation period were two-stage strategies" and "… calibrating impermeable and green area parameters in two separate steps may improve the model performance in the validation period…". I think that currently the results about the validation performance for one-stage and two-stage calibrations are inconclusive. The authors use the sum of ranks from several performance criteria as a proxy for overall performance. Are the results shown in any Table? If yes, I missed them.

Authors' response: the overall ranking is shown for the calibration phase in Table 3, and shown for both calibration and validation phase in Table 9.

Presentation of validation results could be presented better by having all results relating to the baseline calibration (HR model) in one table, i.e. combining tables 6, 9 and 11. A similar table could also be made for all results with the low-resolution model (i.e. replacing table 8) to better illustrate the benefits that the two-stage calibration offers there in terms of flow volume and peak flow performance in the validation phase, since these are more pronounced for the low-resolution model. The results from the low-resolution model were not described extensively in the manuscript but we believe they can strengthen a take-home message for the readers and should therefore be included in more detail. The results are currently best illustrated by Table 4.2 in the corresponding author's licentiate thesis (Broekhuizen, 2019), so the re-organizing of the tables with validation results should also include the data shown there. The table is included below (Table C3 in the supplement) for easy reference, but the data will be organized differently (i.e. one table for the HR model and one for the LR model) in the paper.

Changes in manuscript: re-organize tables as described above, and describe in more detail in the text the effects of the single- and two-stage calibrations for the low-resolution model.

(23) Referee's comment: Also, I would prefer a more quantitative statistic than a sum of class variables (ranks). As NSE is used as the objective criterion for the baseline calibrations it would be a logical choice also for comparing the validation performance.

Authors' response: the validation NSE is already presented in Table 9 to allow for comparison of the different CSs. The problem with any single validation characteristic is that it would either ignore some aspects of model performance or it would have to combine different statistics (i.e. NSE, volume error, peak flow error) in some arbitrary way. E.g. using only the NSE for validation performance would ignore that two-stage strategies perform better in terms of total flow volume and peak flow. We think it is interesting that different statistics give a different view of which calibration strategies perform better and this should be reflected in the manuscript.

Changes in manuscript: The discussion will be focused more on discussing the individual performance statistics to highlight that different criteria give a different picture of the effects of calibration data selection.

(24) Referee's comment: The authors state in Section 3.5. about the validation performance "In terms of NSE, the single-stage calibrations performed better…". On the other hand the 'NSE joint' criterion, typically used for validation (performance over the entire validation data set), seems to be higher for two-stage strategies in Table 6. It is hard for the reader to find guidance here what would be the preferred calibration strategy.

Authors' response: although the single-stage performs better in terms of mean NSE (i.e. NSE calculated for each event, then averaged), it performs worse in terms of joint NSE (i.e. all events collated into a single time series for which NSE is then calculated), joint volume error and mean peak flow ratio. As discussed in section 3.3.2 the downside of the joint NSE is that it can give good scores even when several events are poorly predicted. Therefore joint NSE may be considered too optimistic which is why we did not use it extensively in this paper.

In terms of a take-home message it is important to point out that the two-stage calibration is much faster since it reduces the dimensionality of the calibration problem compared to the single-stage calibration. In addition to this it has sometimes slightly poorer validation performance in terms of NSE but typically better performance according to other characteristics.

The take-home message can also be strengthened by highlighting more the differences between HR and LR models (or rather that the benefits of the two-stage calibration are stronger for the LR model). This is currently best illustrated by Table 4.2 in the corresponding author's licentiate thesis (Broekhuizen, 2019), so the re-organizing of the tables with validation results (see comment #22) should also include the data shown there. The table is included in the supplement as table C3 for easy reference, but the data will be organized differently (i.e. one table for the HR model and one for the LR model) in the paper:

Table C3: Calibration and validation performance of single and two-stage calibration scenarios. HR denotes the high-resolution model, LR the low resolution model. The names of the calibration scenarios are explained in paper III.

	Calibration (6 events)		Validation (19 events)							
Calibration scenario	NSE		# events NSE > 0.5		Mean NSE [a]		Volume error		Peak flow ratio	
	HR	LR	HR	LR	HR	LR	HR	LR	HR	LR
N_T6	0.80	0.84	12	7	0.45	0.21	-0.24	-0.43	0.91	0.50
T6_D_prec	0.74	0.81	11	6	0.43	0.34	-0.25	-0.44	0.91	0.51
T6_P_sum	0.75	0.75	11	8	0.45	0.22	-0.23	-0.38	0.91	0.60
T6_PI_30m	0.74	0.74	9	9	0.29	0.43	-0.24	-0.34	0.98	0.74
T6_PI_mean	0.77	0.77	10	6	0.33	0.38	-0.24	-0.43	0.96	0.59
T6_Q_60m	0.79	0.81	8	6	0.37	0.29	-0.29	-0.46	0.81	0.49
T6_Q_max	0.85	0.86	12	10	0.44	0.49	-0.24	-0.36	0.92	0.64
T6_QV_ppP	0.68	0.65	12	10	0.47	0.37	-0.24	-0.40	0.90	0.66
T32S_D_prec	0.76	0.84	12	10	0.34	0.38	-0.02	-0.05	1.00	0.86
T32S_P_sum	0.83	0.68	10	13	0.34	0.51	-0.15	-0.27	0.99	0.60
T32S_PI_mean	0.83	0.78	13	13	0.44	0.46	-0.16	-0.22	1.00	0.80
T32S_Q_60m	0.79	0.73	10	10	0.33	0.28	-0.13	-0.04	0.99	1.02
T32S_Q_max	0.82	0.80	11	12	0.34	0.33	-0.13	-0.07	0.96	1.03
T32S_QV_ppP	0.70	0.67	11	12	0.46	0.46	-0.26	-0.18	0.87	0.79

[a] mean NSE was calculated after setting NSE of individual events to -1 if NSE was lower than -1, to avoid large influence from negative NSE values.

Changes in manuscript: changes according to the previous two paragraphs.

Recommendation

(25) Referee's comment: In its current form the manuscript is not in my mind publishable in HESS. The following major changes would be required:

Authors' response: we believe that the major changes requested by the referee can be implemented in a new version of the manuscript, as detailed for the individual comments.

Changes in manuscript: see individual points below.

(26) Referee's comment: A more informative description of the hydrometeorological data to allow the readers to understand differences between different calibrations

Authors' response: see our response to comment #13 above.

Changes in manuscript: Include table C1 (supplement) in the manuscript's methods section.

(27) Referee's comment: A better justified reasoning for inclusion/exclusion of different error sources

Authors' response: this is addressed in our response to points 17-21, 24, 28.

Changes in manuscript:

(28) Referee's comment: Most importantly, a clear statement about the scientific novelty value of the manuscript where it becomes obvious what are the new findings over just showing that different calibration data lead to different model parameter values and validation performance

Authors' response: aspects to highlight in the conclusion and abstract:

- Two-stage calibration is faster, and can provide some performance benefits: e.g. better match of flow volume and peak flow in validation phase.
- Benefits of two-stage calibration are stronger for the LR model.
- Confirmation of earlier findings regarding input and calibration data from Dotto et al. (2014) and Kleidorfer et al. (2009) for a different data set and site (more green area). Findings are independent of the calibration event selection which provides support for their general applicability.

Changes in manuscript: changes according to the response above.

Technical corrections

(29) Referee's comment: Mostly technical comments. The comment for Figure 4 also relates to the content of the manuscript.

(30) Referee's comment: Figure 1. Remove the text below the figure (1 map catchment.png). Increase the font size/figure resolution. The legend is hard to read.

Authors' response: The text below the figure is added automatically by the Copernicus Latex template used for the submission and would not appear in the final published version of the article. This applies to the other figures as well.

The font size in the legend can be increased.

Changes in manuscript: increase font size in legend.

(31) Referee's comment: Remove the text below the figure (2 example hydrographs run130.pdf).

Authors' response: see above.

Changes in manuscript: -

(32) Referee's comment: Figure 3. Remove the text below the figure (3 VE PFR histograms.pdf). In the figure caption it is stated peak flow ratios to be on the left whereas in the figure the left panel shows the volume error. Please correct.

Authors' response: -.

Changes in manuscript: upon further consideration the section and figure in question do not contribute much to the papers goals and they have therefore been removed.

(33) Referee's comment: Figure 4. Remove the text below the figure. It is hard to interpret with the given information what is causing the negative NSE for the right panel. Is there a timing difference invisible to the eye? Why does the modelled flow stay at zero for the beginning of the event? Clearly there is rain (left panel), so is the diminished rainfall multiplier and/or increased depression storage value causing all rain falling on the directly connected impervious area to be captured in the depression storage?

Authors' response: The main reason why NSE is low is that the low flow rates in the event mean that the variance of observations is low, see also section 2.5, lines 13-14. For the baseline run (left panel), the variance of the observation is 1.2 L^2 s^{-2} while for the right panel it is just 0.33 L^2 s^{-2}. The variance of the errors meanwhile is 0.25 L^2 s^{-2} (left) resp. 0.38 L^2 s^{-2} (right). NSE is calculated as NSE = 1 – (var err. / var obs.) so the variance of the observations is used as a scaling factor and it is mainly the difference in this factor that causes the degradation in NSE in this example.

Changes in manuscript: Add the variance of observations in figure and refer to section 2.5 in the caption for explanation of NSE and discuss this in the text of section 3.1.4, page 12, line 7.

(34) Referee's comment: Figure 5, 6, 7, 8. Remove the text below the figure.

Authors' response: see above.

Changes in manuscript: -

(35) Referee's comment: Table 11. Mistake in the NSE single-stage value for D_prec (0.41)? The corresponding value in Table 6 is 0.43?

Authors' response: the correct value is 0.43.

Changes in manuscript: correct this to 0.43.

Barco, J., Wong, K.M., Stenstrom, M.K., 2008. Automatic Calibration of the U.S. EPA SWMM Model for a Large Urban Catchment. Journal of Hydraulic Engineering 134, 466–474. https://doi.org/10.1061/(ASCE)0733-9429(2008)134:4(466)

Broekhuizen, I., 2019. Uncertainties in rainfall-runoff modelling of green urban drainage systems: Measurements, data selection and model structure (Licentiate thesis). Luleå University of Technology, Luleå.

Deletic, A., Dotto, C.B.S., McCarthy, D.T., Kleidorfer, M., Freni, G., Mannina, G., Uhl, M., Henrichs, M., Fletcher, T.D., Rauch, W., Bertrand-Krajewski, J.L., Tait, S., 2012. Assessing uncertainties in urban drainage models. Physics and Chemistry of the Earth, Parts A/B/C 42–44, 3–10. https://doi.org/10.1016/j.pce.2011.04.007

Dotto, C.B.S., Kleidorfer, M., Deletic, A., Rauch, W., McCarthy, D.T., 2014. Impacts of measured data uncertainty on urban stormwater models. Journal of Hydrology 508, 28–42. https://doi.org/10.1016/j.jhydrol.2013.10.025

Elliott, A., Trowsdale, S., 2007. A review of models for low impact urban stormwater drainage. Environmental Modelling & Software 22, 394–405. https://doi.org/10.1016/j.envsoft.2005.12.005

Fletcher, T.D., Andrieu, H., Hamel, P., 2013. Understanding, management and modelling of urban hydrology and its consequences for receiving waters: A state of the art. Advances in Water Resources 51, 261–279. https://doi.org/10.1016/j.advwatres.2012.09.001

Houska, T., Kraft, P., Chamorro-Chavez, A., Breuer, L., 2015. SPOTting Model Parameters Using a Ready-Made Python Package. PLOS ONE 10, e0145180. https://doi.org/10.1371/journal.pone.0145180

Kleidorfer, M., Deletic, A., Fletcher, T.D., Rauch, W., 2009. Impact of input data uncertainties on urban stormwater model parameters. Water Science and Technology 60, 1545–1554. https://doi.org/10.2166/wst.2009.493

Krebs, G., Kokkonen, T., Valtanen, M., Setälä, H., Koivusalo, H., 2014. Spatial resolution considerations for urban hydrological modelling. Journal of Hydrology 512, 482–497. https://doi.org/10.1016/j.jhydrol.2014.03.013

Petrucci, G., Bonhomme, C., 2014. The dilemma of spatial representation for urban hydrology semi-distributed modelling: Trade-offs among complexity, calibration and geographical data. Journal of Hydrology 517, 997–1007. https://doi.org/10.1016/j.jhydrol.2014.06.019

Rossman, L.A., Huber, W.C., 2016. Storm Water Management Model Reference Manual. Volume I: hydrology (Revised). U.S. Environmental Protection Agency, Cincinnati.

Sun, N., Hall, M., Hong, B., Zhang, L., 2014. Impact of SWMM Catchment Discretization: Case Study in Syracuse, New York. Journal of Hydrologic Engineering 19, 223–234. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000777

Tscheikner-Gratl, F., Zeisl, P., Kinzel, C., Leimgruber, J., Ertl, T., Rauch, W., Kleidorfer, M., 2016. Lost in calibration: why people still do not calibrate their models, and why they still should – a case study from urban drainage modelling. Water Science and Technology 74, 2337–2348. https://doi.org/10.2166/wst.2016.395

Note to reviewers: most of the tables in the manuscript have been updated since the previous version of this manuscript. The automated tool used to create this version with changes marked struggles to display this in an intelligible way, so for most tables just the new version is shown. This is indicated in the table captions as well.

~~Calibration event selection~~ Selection of calibration events for modelling green urban drainage~~modelling~~

Ico Broekhuizen[1], Günther Leonhardt[1], Jiri Marsalek[1], and Maria Viklander[1]

[1]Luleå University of Technology, Department of Civil, Environmental and Natural Resources Engineering, Urban Water Engineering. Luleå, Sweden

Correspondence: Ico Broekhuizen (ico.broekhuizen@ltu.se)

Abstract. Calibration of urban drainage models is typically performed based on a limited number of observed rainfall-runoff events, which may be selected from a ~~longer time-series of measurements~~ larger dataset in different ways. In this study, 14 single- and two-stage strategies for selecting ~~these~~ the calibration events were tested ~~for~~ in calibration of a SWMM model of a predominantly green urban area. The event selection was considered in relation to ~~other~~ runoff contributions from green pervious areas and such sources of uncertainty such as ~~measurement uncertainties, objective functions,~~ rainfall/runoff measurement uncertainties and catchment discretization. Even though all 14 strategies resulted in successful model calibration, the difference between the best and worst strategies reached 0.2 in Nash-Sutcliffe Efficiency (NSE) and the calibrated parameter values notably varied. Most, but not all, calibration strategies were robust to ~~changes in objective function,~~ perturbations in calibration data and the use of a ~~low spatial resolution~~ coarse catchment discretization model in the calibration phase. The various calibration strategies satisfactorily predicted 7 to 13 out of 19 validation events. The two-stage strategies performed better than the single-stage strategies when~~measuring performance using the Root Mean Square Error, flow volume error or peak flow error (but not using NSE) ; when~~: (1) perturbing flow data in the calibration ~~period had been perturbed~~ events by +-40%; and ~~when using a lower model resolution~~(2) using a coarser catchment discretization, especially in terms of total flow volume and peak flow rates. The two calibration strategies that performed the best in the validation ~~period~~ phase were two-stage strategies. The findings in this paper show that ~~different~~ various strategies for selecting calibration events ~~may~~ lead in some cases to different results ~~for the validation period~~in the validation phase, and that calibrating ~~impermeable~~ impervious and green area parameters in two separate steps may ~~improve model performance in the validation period, while also~~ increase the effectiveness of model calibration/validation by reducing the computational demand in the calibration phase and improving model performance in the validation phase.

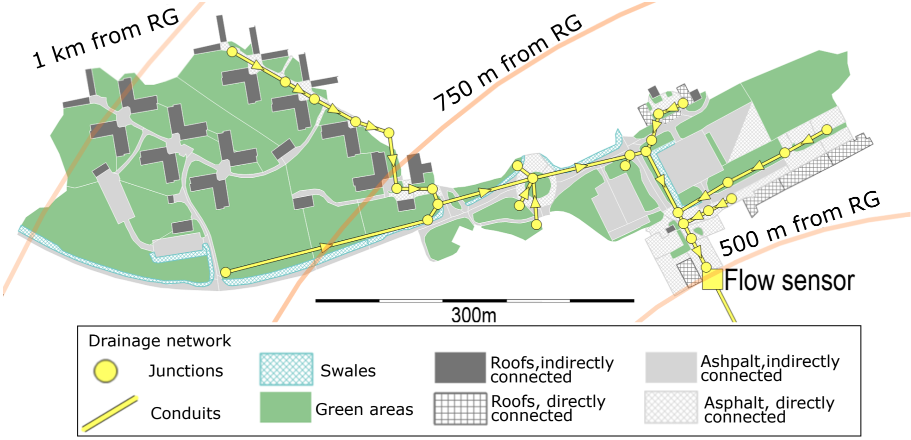
1 Introduction

Calibration of generic urban drainage model codes is usually required to obtain a model representing an actual site with sufficient accuracy. In the calibration process, the information contained in records of relevant variables, such as rainfall and flow rates at the catchment outlet, is used for estimating model parameter values that produce results consistent with the data (Mancipe-Munoz et al., 2014). It can be expected that the best parameter estimates will be obtained when they are inferred from the largest amount of information, i.e. by using all data from a long series of measurements. However, the availability of calibration data may be limited and the nature of the calibration process, by trial and error, requires model iterations for many different parameter sets, which means that the runtime of the model has to be kept short and the length of the simulated periods should be limited. Therefore, calibration may have to be performed on a limited number of rainfall events from a longer record. As each of the available events will differ from the others, it can be expected that the choice of a specific event (or an event set) will influence the results of calibration (Tscheikner-Gratl et al., 2016).

Tscheikner-Gratl et al. (2016) studied such influence by calibrating water level in the outflow pipe of a catchment using ten different rain events. They found that two of them could not be reproduced in calibration and the others, while successful in calibration, could only predict up to six of the remaining events. When applying the calibrated models with design storms, they found that the calibrated models predicted different flooding volumes. In calibration of combined sewer overflow (CSO) volumes, Kleidorfer et al. (2009b) compared calibration results obtained for (~~i~~1) the five longest duration events and (~~ii~~2) the five highest peak flow events, finding that using the longest duration events reduced the number of measurement sites required for successful calibration. Schütze et al. (2002) demonstrated that calibration based on discrete events saved time compared to calibrating for a complete time series, but also that this introduced additional uncertainty. Mourad et al. (2005) showed that calibration of a stormwater quality model was sensitive to: (~~i~~1) which randomly selected events were used, and (~~ii~~2) how many events were used.

While the above papers helped elucidate some aspects of the sensitivity of urban drainage model calibration to the calibration events used, such findings possess some limitations: firstly, only a limited number of generally available options for selecting calibration events has been considered; secondly, the modelling focused on traditional urban drainage systems~~where~~, in which generation of runoff is dominated by impervious surfaces, but the current trend towards green urban drainage infrastructure creates the need to pay more attention to runoff processes on green areas ~~Fletcher et al. (2013)~~ ~~.~~ (Elliott and Trowsdale, 2007; Fletcher et al., 2013). This second aspect also applies to investigations into other sources of uncertainty in urban drainage modelling, some of which have been investigated before, e.g. input and calibration data uncertainties(Dotto et
and spatial model resolution (Krebs et al., 2014; Petrucci and Bonhomme, 2014; Sun et al., 2014). However, these investigations used predominantly impervious catchments and it is, therefore, unknown to what extent their findings apply to greener urban catchments as well and how sensitive such results are to the calibration data set that was used.

~~The~~ Considering the above findings, the primary objective of the paper that follows is to advance the knowledge of calibration processes for green urban areas by examining different strategies for ~~calibration event selection and their effects~~ selecting calibration events and assessing the effects of such selections on the performance of a calibrated hydrodynamic model of a


1 km from RG
750 m from RG
500 m from RG
Flow sensor
300m
Drainage network
Junctions
Conduits
Swales
Green areas
Roofs,indirectly connected
Roofs, directly connected
Ashpalt,indirectly connected
Asphalt, directly connected

catchment larger legend.pdf

Figure 1. Map of the studied catchment showing elements of the high-resolution rainfall-runoff model and the distance of the catchment to the rain gauge (RG). The diameters of the pipes range from 400 mm for the main trunk where the flow sensor is located to 200 mm for the smaller branches.

predominantly green urban catchment. ~~Since uncertainties in~~ Part of this is a proposal for a practical two-stage calibration strategy. Two secondary objectives are to verify: (1) the findings from previous urban drainage modelling ~~arise from other sources as well (Deletic et al., 2012), the calibration event selection is considered in relation to some of them~~studies on a greener (less impervious) catchment, and (2) sensitivity of the earlier findings to the calibration data used.

2 Materials and methods

2.1 Study site and data

The study site is a 10.2 ha catchment in the city of Luleå, Sweden (see Figure ~~1~~1). The catchment area comprises 63% of green areas, 12% of impervious areas ~~draining~~ connected directly to the storm sewer system, and 25% of impervious areas draining ~~to~~ onto adjacent green areas. The green areas include a number of vegetated swales that are connected to the storm sewer system at their lowest point.

Precipitation was measured at 1-minute intervals with a Geonor T200B weighing-bucket precipitation gauge located outside of the study catchment, about 500 and 1,000 metres from the nearest and furthest borders of the catchment, respectively (see circles in Figure 1). The gauge was tested in the field and confirmed to work well twice a year in 2016 and 2017, and before 2016, such tests were also performed occasionally. Laboratory and field tests (by others) found this design of precipitation sensor to be a reliable instrument (Duchon, 2002; Lanza et al., 2010). Records were available for individual rain events in 2013-2015 and continuously for 2016 and 2017.

~~The flow~~ Flow rates in the storm sewer draining the catchment were ~~performed,~~ measured at 1-minute intervals ~~,~~by means of an ISCO 2150 AV sensor (a combination of an acoustic Doppler velocimeter and a pressure transducer) installed in the catchment outlet formed by a 400 mm diameter concrete sewer pipe. This type of sensor was assessed in the laboratory by Aguilar et al. (2016) and found to have a combined uncertainty (consisting of bias, precision and benchmark uncertainty) of ±19.0 mm for the water depth measurements (the test range was 10-150 mm) and ±0.0985 m/s for the velocity measurement (test range 0.1-0.6 m/s). These tests were carried out in a 0.46 m wide square channel, so the stage-discharge relationship was different from the study site described herein. It was also reported that the field performance of this type of sensors can suffer from the presence of too few (Teledyne ISCO, 2010) or too many particles suspended in the water (Nord et al., 2014).

While the difficulties in estimating all the uncertainties at the actual field site prevented a precise determination of the uncertainties~~'~~' magnitude, the general lab tests of the sensors used confirmed the acceptability of their records for the study purpose. Finally, it was also confirmed by Dotto et al. (2014) that errors in the calibration data can be compensated for in the calibration process.

The available precipitation record was divided into rainfall events with ~~at least six hourswithout precipitation between them~~a minimum inter-event time of no precipitation of six hours. Events deemed suitable for use in calibration were selected using the following criteria:

1. A minimum total precipitation of 2 mm (Hernebring, 2006).

2. No or small gaps in rain and flow data , i.e. both have to be available for >90% of the event duration.

3. Sufficient in-pipe water depths for the flow sensor to work reliably: >10 mm during at least 50% of the event and >25 mm at least once in the event, based on recommendations from the manufacturer (Teledyne ISCO, 2010).

4. Peak flow >2 L s^{-1}, since relative measurement uncertainties are high below this point.

5. No snowfall or -melt, since these would introduce additional processes in the hydrological behaviour and model of the catchment.

Calibration and validation periods were separated by using the 19 observed events from 2016 for the validation period, and the 32 events from 2013-2015 and 2017 for the calibration period. In this way, all the calibration scenarios (see section 2.3) were tested (validated) against the same dataset and no calibration scenarios could benefit from including calibration events that also appeared in the validation set. The year 2016 was selected as the validation period for two reasons: it was the year with total precipitation closest to the annual mean, and the measured data records were continuous. Table 1 contains an overview of all events that were used in at least one calibration scenario as well as an initial estimate of the runoff from green areas.

2.2 Runoff model and calibration approach

The US EPA Storm Water Management Model (SWMM) was selected since it is a commonly used semi-distributed urban drainage model ~~and it~~ that allows to route runoff from one sub-catchment to another. This routing feature was needed since it allows for a high-resolution model setup in which each subcatchment (146 were used in total) features a single land cover. The high resolution input data needed for this approach was available in the form of GIS data, aerial photographs, and observations

Table 1. Characteristics of all rainfall events used in one or more calibration scenarios. Note to reviewers: this table is new in this revised version of the manuscript.

Event #	Precipitation sum in preceding 72 hr	Precipitation sum (P_sum)	Precipitation duration (D_prec)	Average precipitation intensity (PI_mean)	Highest 30-minute average precipitation intensity (PI_30m)	Runoff volume (QV)	Percentage runoff (QV_ppP)	Peak flow rate (Q_max)	Highest 60-minute average flow rate (Q_60m)	Runoff from green areas [a]	Of which originating from imperv. areas [b]	Originating from green areas [c]	Average percentage runoff from green areas [d]
	mm	mm	hr	mm hr^{-1}	mm hr^{-1}	mm	%	L s^{-1}	L s^{-1}	mm	mm	mm	%
199	2.4	13.8	41.6	0.3	4.0	1.7	12.4	4.2	3.3	0.06	0.02	0.04	0.3
209	0.2	8.0	9.5	0.8	2.8	0.5	6.9	4.5	2.7				
211	8.3	9.7	22.8	0.4	6.9	1.1	11.1	29.2	11.1				
214	7.3	6.4	12.1	0.5	4.3	0.6	10.1	40.5	8.5				
222	1.1	9.8	12.8	0.8	7.5	0.7	7.2	26.4	13.3				
270	0.0	9.3	38.5	0.2	3.5	1.1	11.3	22.9	8.7				
306	10.1	8.6	9.1	0.9	7.1	0.7	8.5	27.5	9.3				
307	18.3	29.9	37.7	0.8	8.5	4.9	16.2	71.2	42.9	1.27	0.36	0.91	3.0
310	12.7	8.6	10.0	0.9	7.5	1.2	14.0	37.4	17.4	0.17	0.05	0.12	1.4
530	13.8	6.7	2.8	2.4	7.2	0.8	11.2	58.9	13.5				
939	0.6	7.0	25.6	0.3	1.0	0.4	5.7	2.1	1.8				
962	0.0	8.5	11.2	0.8	1.4	2.1	24.9	4.9	4.4	1.09	0.31	0.78	9.2
971	0.2	2.6	18.6	0.1	1.1	0.3	11.3	4.0	2.9				
978	12.7	25.0	65.8	0.4	5.8	4.8	19.1	64.5	16.6	1.77	0.50	1.27	5.1
982	0.0	5.6	3.4	1.7	7.0	0.9	15.8	49.5	17.2	0.21	0.06	0.15	2.7
984	13.1	2.4	6.3	0.4	4.6	1.4	59.1	71.7	14.0	1.12	0.32	0.80	33.7
995	4.8	2.1	8.5	0.2	1.8	0.6	28.6	32.0	9.7	0.35	0.10	0.25	11.9
997	2.2	24.6	49.0	0.5	2.4	5.1	20.7	15.0	6.9	2.14	0.61	1.53	6.2
1001	0.0	35.3	56.6	0.6	8.6	8.8	25.0	56.5	32.5	4.58	1.30	3.28	9.3
1004	22.5	4.2	13.9	0.3	5.9	1.1	25.2	33.3	10.6	0.56	0.16	0.40	9.5
1019	0.5	22.3	49.7	0.4	2.3	4.7	21.2	12.9	9.3	2.06	0.58	1.47	6.6
1028	6.2	2.8	7.0	0.4	1.3	1.2	43.5	6.3	4.2	0.89	0.25	0.64	22.5

[a] Calculated assuming 100% runoff from impervious areas: a = QV - 0.12 P_sum, where 0.12 is the percentage of directly connected impervious area. (Some of this runoff originated from impervious areas that drained to green areas).

[b] Calculated as b = a (25 / (25+63)), where 25 and 63 are the percentages of indirectly connected impervious surfaces and green surfaces respectively.

[c] Calculated as c = a - b

[d] Calculated as d = c / P_sum

from site visits. The advantage of these single ~~land-use~~ land-cover subcatchments is that their parameter values maintain their physical meaning and can be calibrated (or appropriate values found in the literature) for each land use or cover. The traditional approach of using larger subcatchments with multiple land uses/covers usually necessitates calibration to estimate the values of parameters that then represent a weighted average value over multiple land uses/covers. Some spatial characteristics, such as the slope and the width of subcatchments, can also be estimated more easily for smaller, uniform subcatchments. This approach has been used successfully by e.g. Krebs et al. (2014, 2016), Petrucci and Bonhomme (2014) and Sun et al. (2014). Within SWMM the Green-Ampt infiltration method was selected since it can be calibrated with just two parameters (Rossman, 2016).

Whenever feasible, parameters for ~~the~~ different subcatchments were set directly from the available GIS data and site visits, i.e. the sizes and slopes of all subcatchments and sewer pipes, as well as the catchment widths of small and disconnected roofs. For other subcatchments the catchment width was calibrated together with the other model parameters. To reduce the scope of the calibration problem, parameters were grouped based on land cover, yielding a total of thirteen calibration parameters for the hydrodynamic model. Parameter values were limited based on values reported in the literature (see Table 2). The precipitation gauge was situated a few hundred metres outside of the actual catchment, and may have provided a biased estimate of the catchment rainfall. Therefore, a rainfall multiplier for each individual rainfall event was included in the calibration. This approach has been used with satisfactory results e.g. by Datta and Bolisetti (2016), Fuentes-Andino et al. (2017, and Vrugt et al. (2008), although it is limited by assuming a simple multiplicative difference between the gauge and catchment-average rainfall, which is not necessarily the case (Del Giudice et al., 2016). Furthermore, rainfall multipliers do not address the spatial variability of the rainfall, but in the absence of multiple rain gauges or other information about the spatial variability of rainfall in the study catchment, there were no feasible alternatives in this case. The rainfall multipliers create a way of adjusting the rainfall volume in the calibration so that the simulated runoff volume can better match the observed runoff volume. ~~It is, however, not possible to distinguish between~~ However, the multipliers do not allow distinguishing between (1) deviations between rainfall at the gauge and the catchment-averaged rainfall, (2) errors in the rainfall measurement, and (3) errors in the runoff measurement. A more traditional approach would be to calibrate the percentage of impervious areas, but in view of the availability of high-resolution land-cover information, it was preferred to apply rainfall multipliers instead.

Green surfaces like those in the study area have a long hydrological memory for antecedent rainfall, and this had to be accounted for in the simulations. Neglecting this memory would increase the risk of green areas allowing unrealistically high infiltration in some rainfall events. Since SWMM does not allow for setting the initial values of state variables directly, such adjustments can be done by choosing an appropriate warm-up period for modelling runs. When sufficiently long warm-up periods are used, this approach offers an advantage consisting of treating the first rainfall/runoff peak of an event the same as way as any following peaks, i.e., with initial conditions corresponding to a continuous simulation. The required length of this warm-up period was estimated by finding the last time before each rainfall event when the study area was dry. This was calculated for all rainfall events using the actual precipitation data and for various values for the maximum depression storage and infiltration rate. The last antecedent time when the study area was dry was then used as the starting point of the warm-up period. This lookup procedure was applied to every event for each iteration in the calibration process, so that all events were treated the same way as in a continuous simulation.

Table 2. Calibration parameters and their ranges.

Parameter	Abbr.	Groups	Range	Reference
Subcatchment width [m]	width	Asphalt parking lots (AP)	20-200	Physical dimensions of subcatchments
		Grass areas (GR)	1-200	
		Swales (SW)	0-5	
Subcatchment length [m]	length	Asphalt roads[a]	0.5-5	
Manning's number [-]	n	Impervious surfaces (IMP)	0.005 - 0.015	(Krebs et al., 2016; Rossman, 2016)
		Grass areas (GR)	0.1 - 0.5	
		Swales (SW)	0.1 - 0.5	
		Pipes	0.010 - 0.015	
Depression storage [mm]	s	Impervious surfaces (IMP)	0 - 2.5	
		Grass areas (GR)[b]	0 - 20	
		Swales (SW)[c]	0 - 150	(Rujner et al., 2018)[d]
Saturated hydraulic ~~conductivity~~ conduc-tivity [mm ~~hr-1~~ hr^{-1}]	ksat	Grass areas (GR)[e]	1 - 200	(Rawls et al., 1983)
Initial moisture deficit [-]	imd	Grass areas (GR)[e]	0.10 - 0.35	

[a] In SWMM, the subcatchment width is an input, but in this group of subcatchments, the length (in the flow direction) showed more similarity among the subcatchments, so it was calibrated instead of the width.

[b] Includes vegetation and trees as well.

[c] The maximum value was intentionally set high since the swales' outlets are not always located exactly at the lowest points and the swales can be observed with larger ponds after heavy rain events.

[d] Field experiments on similar swales in the same city.

[e] Used for both grass areas and swales.

In the calibration process, the Shuffled Complex Evolution - University of Arizona algorithm (SCE-UA; Duan et al. (1994)) was used to estimate the optimal values of the parameters. The algorithm was selected because it is commonly used in hydrological studies and allows for parallel computing. The Python library SPOTPY (Houska et al., 2015), which includes this algorithm, was used to carry out the entire calibration process.

2.3 Event selection

This paper investigates single- and two-stage calibration scenarios (CS), with each CS using six rainfall events. The single-stage CSs used the six events with the highest values of a certain event characteristic, and calibrated all parameters simultaneously. Two-stage calibration scenarios calibrated first the parameters related to impervious areas, using a set of three rainfall events, followed by the pervious area parameters using another set of three rainfall events. Since only 12% of the total catchment surface is impervious and connected directly to storm sewers, it was assumed that the events, for which runoff volume was less than 12% of rainfall volume, produced runoff only from impervious areas. (It is conceivable that there is some contribution of green areas when the percentage runoff is less than 12%, and in that case the threshold should be set at a lower value, but since the amount of green area runoff and the appropriate value of the threshold would be highly dependent on antecedent conditions this was not included here.) Therefore, these events were suitable for calibration of impervious area parameters in the first stage of the calibration process. Following this step, events with more than 12% runoff were assumed to also include runoff from green areas and were used to estimate pervious area parameters in the second stage of the calibration. When calibrating the green area parameters, the parameters related to impervious areas were kept fixed at their values from the first stage. This procedure splits the optimization problem into two smaller problems that have fewer parameters and shorter run times. The smaller number of parameters (reduced dimensionality) can ease the search for optimal parameter sets, while the shorter run time per iteration allows shortening the total time needed, increasing the number of iterations used, or including more events in the calibration.

Characteristics related to the rainfall, flow depths and flow rates were calculated for each event. For the single-stage calibration scenarios, the six highest ranking events for each characteristic were selected. For the two-stage calibration scenarios, the three highest ranking events with less than 12% runoff were selected for the first stage and the three highest ranking events with more than 12% runoff were selected for the second stage. Applying the calibrated rainfall multipliers in the calibration (Sect. 2.2) means that event properties relating to rainfall and percentage runoff will change, and the percentage runoff can change from <12% to >12% and vice versa. Doing this consistently for all events in the calibration procedure would require (1) re-calculating which events should be available in each stage, (2) estimating in some way rainfall multipliers for all events, including those not initially selected by any calibration scenario, (3) re-calculating which events are used in each CS, and (4) repeating the calibration for any CS that has had any of its events changed. Although this might improve the overall results of the proposed calibration procedure, it would also increase the complexity and raise several new issues, such as how to obtain a calibrated rainfall multiplier for the 10 events that were not used in any CS. We considered this to be beyond the paper's original scope of examining different strategies for calibration event selection and proposing a practically useable two-stage calibration procedure.

To avoid making the comparison too large in scope, a limited number of calibration scenarios (eight single-stage and six two-stage) was selected for use in this study. This selection was made so that it included a range of different characteristics and avoided multiple CSs with the exact same set-up of events. The names of the CSs consist of two or three elements:

- T6 (Top 6) for single-stage or T32S (Top 3 - 2 stages) for two-stage scenarios.
- The relevant event characteristic: precipitation (P), precipitation intensity (PI), runoff flow rate (Q), flow volume (QV), or flow volume as percentage of rain QV_ppP, precipitation duration D_prec.
- The duration over which the characteristics were calculated: sum, mean and max refer to the whole event. 30 and 60 min refer to the time interval used to calculate an average rainfall intensity or flow rate (i.e. the highest value found within the event for a 30 or 60 minute moving average). Calculating rainfall intensities and average flow rates over these windows rather than the entire event suppresses the effects of e.g. dry periods within events on such calculations.

The calibration scenario N_T6 consists of the six events that were selected most often in other calibration scenarios with the goal of obtaining a set of events that score highly on a variety of characteristics.

2.4 Other sources of uncertainty

Calibration data selection is not the only source of uncertainty in urban drainage modelling. Deletic et al. (2012) ~~identify~~ identified nine sources: (1) input data, (2) model parameters, (3) calibration data measurements, (4) calibration data selection, (5) calibration algorithm, (6) objective functions, (7) conceptualisation (e.g. discretization), (8) process equations and (9) numerical methods and boundaries. As described above, calibration data selection is the focus of this paper~~, however, it~~. However, earlier findings regarding the other sources of uncertainties were based on predominantly impervious catchments and they should not be ~~viewed in isolation from the other~~ assumed to apply equally to greener catchments. The nature of the catchment in this paper provides an opportunity to (1) check if these findings apply to greener catchments as well and (2) check if these findings are sensitive to the calibration data set that is used. It was beyond the scope of this paper to break new ground in all of the nine sources listed above~~. Therefore, different strategies for selecting calibration events were considered in relation to the~~ ; therefore, we focused on uncertainty sources that have been covered in earlier literature. The uncertainties arising from objective functions, calibration algorithms and numerics are not considered explicitly in this paper. The choice of objective function can be expected to affect the calibration results, but this issue has received hardly any attention in urban drainage modelling, except for some short remarks by Barco et al. (Barco et al., 2008). Likewise, the calibration algorithm (Deletic et al., 2012; Houska et al., 2015) and numerical issues (Deletic et al., 2012; Kavetski et al., 2006) are recognized as sources of uncertainty, but there is a lack of studies addressing these specifically for urban drainage modelling that could be referred to here. Since breaking new ground in these areas was considered beyond the scope of this paper, these sources of uncertainty are not considered here. The inclusion of other sources of uncertainty ~~as discussed below~~is described in the remainder of this section.

Rainfall input uncertainty. ~~Since the rain~~ Earlier studies of the Geonor T200B rain gauge used have reported wind-induced undercatch of 4-5% (Duchon and Essenberg, 2001; Lanza et al., 2010). Additionally, there may be some deviations between the rainfall at the gauge ~~is located outside of the catchment and the maintenance of the gauge was carried out by different people, it is possible that there are structural errors~~ and in the catchment. It is therefore possible that structural errors exist in the rainfall measurements. This aspect was investigated by examining the rainfall multipliers that were included for each event in the calibration (see Sect. 2.2). It should be noted that the rainfall multipliers are used to adjust flow volumes and that they may therefore also reflect uncertainties in e.g. subcatchment delineation and runoff routing.

Parameter uncertainty. The uncertainty of urban drainage model parameter estimates has been investigated extensively earlier, e.g., by Del Giudice et al. (2016), Dotto et al. (2009, 2011, 2012), Kleidorfer et al. (2009a) and Muleta et al. (2013). Therefore, this issue is addressed herein just by comparing the parameter values obtained in different calibration scenarios.

Calibration data measurement uncertainties. Measurement uncertainties of flow rates in storm sewer pipes have been described by a number of researchers, e.g., Aguilar et al. (2016), Blake and Packman (2008), Bonakdari and Zinatizadeh (2011), Heiner and Vermeyen (2012), Lepot et al. (2014), Maheepala et al. (2001). In this paper, structural flow measurement errors are considered by testing calibration after reducing or increasing all flow observations by 40%. This value was chosen on the basis of uncertainties reported by Aguilar et al. (2016) and applied to the ~~current~~ study outflow measurement location~~and is slightly higher than the value of 30% used by~~ . . This is a rather simple approach and other ways of simulating errors in the measured data may be considered: e.g. Dotto et al. (2014) ~~and Kleidorfer et al. (2009a). The flow data from the validation period was not adjusted. Other researchers (e.g. ibid)~~ also tested the effect of random errors; ~~such effects and their thorough investigation were~~ However, since many different ways of perturbing flow data can be used it was deemed outside of the scope of this paper to examine them all, and only the constant offset was used as a simple way of introducing errors in the flow measurement. However, it should be noted that the use of measured flow rates, implemented in this study, involves the presence of random errors in the calibration data sets used. The flow data from the validation period was not adjusted.

~~*Objective functions*. The calibration process strives to find the optimal value of the specified objective function, so the choice of such a function can be expected to affect the calibration results. This was addressed here by assessing all calibration scenarios using both Nash-Sutcliffe model efficiency (NSE) and Root Mean Square Error (RMSE) as objective functions (see Sect. 2.5).~~

Conceptualisation / model discretization. ~~The model code (SWMM) employed in this study has been widely used for many years, with some improvements made to those parts of its conceptualisation that were deemed unsatisfactory. Therefore, it is safe to assume that the SWMM conceptualization (Rossman, 2016) is appropriate~~ Although model structure is also a recognized source of uncertainty (Deletic et al., 2012), it was not considered here since: (a) there is a lack of previous research on this topic for urban drainage modelling ~~and there was no need to consider this issuefurther. However, the~~ that could be referred to, and (b) there is a lack of methods to address this issue, other than using different models in parallel, which was considered outside the scope of this study, and would be difficult since the catchment model requires some SWMM features (e.g. routing runoff from one subcatchment to another, good support for automated runs), which are not always available in other models.

The choice of catchment discretization into the subcatchments in the model ~~is done, somewhat subjectively, by the modellers for individual studies: therefore~~has been investigated by several authors. Tscheikner-Gratl et al. (2016) found that a lumped model was not able to reproduce the shapes of storm runoff hydrographs as well as a more detailed model, even though total runoff volumes were similar. Sun et al. (2014) and Krebs (2014) found that a finer discretization resulted in parameter values that were more applicable to other study sites and events. Petrucci and Bonhomme (2014) found that using additional geographic information to increase the spatial resolution could improve model performance, since some model parameters can then be estimated directly from geographic data (see also Dongquan et al., (2009); Warsta et al., (2017). To investigate the impact of calibration data selection on these findings and to check them for a predominantly green urban catchment, two levels of discretization were compared: (~~i~~1) the basic model set-up (the high-resolution model described in Sect. 2.2), and (~~ii~~2) a simpler, more traditional set-up using five subcatchments. In the latter case, each subcatchment was created by aggregating multiple smaller subcatchments from the high-resolution model. The area and percentage imperviousness of each aggregated subcatchment were calculated from its constituent smaller catchments. The calibration parameters were modified accordingly, as shown in Table 3, with the total number of calibration parameters (including rainfall multipliers) being the same.

Table 3. Calibration parameters and their ranges for the low-resolution model.

Parameter	Abbr.	Groups	Range	Reference
Subcatchment width [m]	width	5 individual subcatchments	20 - 200	Physical dimensions of subcatchments
Manning's coefficient [-]	n	Impervious surfaces (IMP)	0.005 - 0.015	
		Pervious surfaces (GR)	0.1 - 0.5	
		Pipes	0.010 - 0.015	(Krebs et al., 2016; Rossman, 2016)
Depression storage	s	Impervious surfaces (IMP)	0 - 2.5	
		Pervious surfaces (GR)	0 - 20	
Percentage runoff routed from impervious to pervious (%)		See footnote [a]	1-99	
Saturated hydraulic conductivity [mm hr^{-1}]	ksat	Grass areas (GR)	1 - 200	(Rawls et al., 1983)
Initial moisture deficit [-]	imd	Grass areas (GR)	0.10 - 0.35	

[a] For two subcatchments the percentage routed was estimated at 0% and 100% respectively. A single percentage was calibrated and shared by the three remaining subcatchments.

~~*Sources of uncertainty not considered.* The calibration algorithm used in this study (SCE-UA) has been widely applied in hydrological applications with great success, so there was no need to subject it to scrutiny in this paper. Similarly, since SWMM is a well-established mature model, there was no need to examine the equations, numerical methods and boundaries used in the model.~~

2.5 Objective functions

~~Each calibration scenario was run with two different objective functions, of which values were first calculated for individual events and the average of those values for the whole scenario served as the target for optimization.~~ The objective function used for ~~all except one~~ the calibrations was the Nash-Sutcliffe model efficiency:

$$\mathrm{NSE} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(O_i - \bar{O})^2} \qquad (1)$$

Where O denotes observed values and S simulated values. The NSE measures the variance of the model errors (the numerator) as a fraction of the variance of the observations (the denominator). This fraction is then scaled so that it extends from -infinity (i.e., the worst possible fit) via 0 (the score that would be achieved by using the average of observations) to 1, for a perfect fit. The NSE is dimensionless, so it allows comparing runoff events of different magnitudes. However, when the variance of the observations is small (e.g. for small runoff events), it can become quite sensitive to small changes in the simulated hydrograph. ~~To examine the impact of different objective functions, one calibration used Root Mean Square Error (RMSE):~~

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$$

~~RMSE has the same units as the observations (in this case L s^{-1} for the flow rate)~~The NSE was calculated for each individual event and the average used as the calibration objective. For further assessment of the modelled hydrographs, two metrics related to the peak flow and the hydrograph volume were used. The peak flow ratio (PFR) was defined as the ratio of the highest simulated to the highest observed flow rates, regardless of the times when they occurred:

$$\mathrm{PFR} = \frac{\max S_i}{\max O_i} \tag{2}$$

Where values >1 indicate overestimated simulated peak flows and values <1 indicate underestimated simulated peak flows. Finally, the relative volume error (VE) considers total flow volumes throughout the event:

$$\mathrm{VE} = \frac{\sum_{i=1}^{n}(S_i - O_i)}{\frac{1}{n}\sum_{i=1}^{n} S_i} \tag{3}$$

It is positive when the simulated total flow volume exceeds the observed one and vice versa. Note that the above formula is only valid if the observation interval is constant. The peak flow ratio and volume error were used here since peak flow rates and storage volumes are often the targets that drainage systems are designed for.

The quick response of the studied catchment means that low flow rates may cover a significant part of the event. Measurements in this range have relatively high uncertainties and may be considered less relevant than periods with higher flows. Therefore, it should be avoided that low flows dominate the analysis, which was achieved by including only time steps with observed flow rates >1 L s^{-1} in calculating these metrics.

3 Results and discussion

3.1 Calibration performance

3.1.1 Baseline calibration

The baseline calibration (i.e. ~~with NSE as objective function,~~ using the high resolution model without flow data perturbations) was successful for all calibration scenarios, with average NSE for all events ranging from 0.68 to 0.85 (see Table 4). The lowest NSE corresponded to the two CSs based on the percentage runoff (T6_QV_ppP and T32S_QV_ppP). This result can be attributed to one event (see ~~Figure 2~~2), for which both CSs resulted in simulated hydrographs with low NSE, in spite of a visually good fit of the observed data. In this case, low NSE resulted from a small timing error and from low flow rates in the event, which lead to a low variance of the observations and, therefore, an NSE that is more sensitive to small simulation

Table 4. Calibration results. Bold font indicates the best value in each column. Note to reviewers: this table has been updated in this revised version of the manuscript.

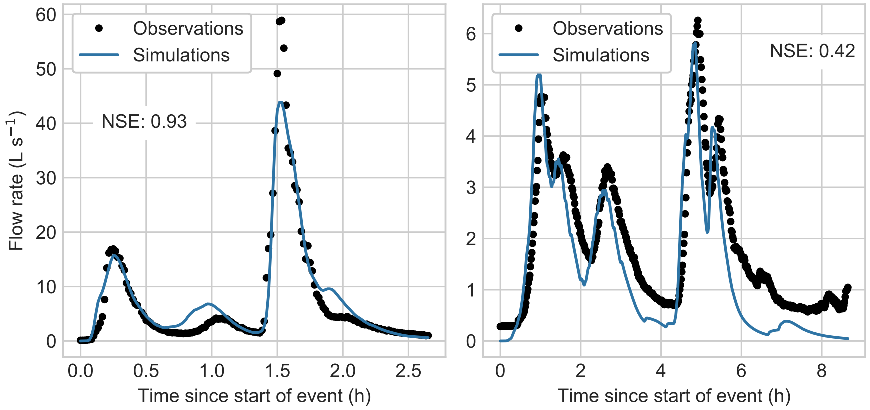
	High resolution model					Low resolution model			Mean NSE
	Baseline			Flow-40%	Flow +40%				
	NSE	VE	PFR	NSE	NSE	NSE	VE	PFR	
N_T6	0.80	-0.07	0.93	0.77	0.76	0.84	0.03	0.85	0.78
T6_P_sum	0.75	-0.11	0.96	0.65	0.65	0.75	-0.07	0.90	0.68
T6_PI_mean	0.77	-0.04	0.90	0.63	0.78	0.77	0.02	0.86	0.73
T6_PI_30m	0.74	-0.09	0.95	0.72	0.72	0.74	-0.05	**0.95**	0.72
T6_Q_max	**0.85**	-0.03	0.89	0.82	**0.84**	**0.86**	0.04	0.86	**0.84**
T6_Q_60m	0.79	-0.09	0.91	0.77	0.77	0.81	**0.01**	0.90	0.78
T6_QV_ppP	0.68	-0.11	0.89	-0.10	0.65	0.65	-0.09	0.94	0.41
T6_D_prec	0.74	-0.10	0.92	0.72	0.69	0.81	-0.02	0.86	0.72
T32S_P_sum	0.83	0.03	0.90	0.77	0.83	0.68	0.08	0.74	0.81
T32S_PI_mean	0.83	0.03	0.96	0.75	0.80	0.78	0.05	0.84	0.79
T32S_Q_max	0.82	0.06	0.86	0.79	**0.84**	0.80	0.07	0.78	0.81
T32S_Q_60m	0.79	0.04	**0.98**	0.73	0.76	0.73	0.02	0.93	0.76
T32S_QV_ppP	0.70	0.06	0.85	0.62	0.73	0.67	0.11	0.75	0.68
T32S_D_prec	0.76	**0.02**	0.97	**0.83**	0.73	0.84	0.03	0.85	0.77

errors. For the two-stage calibration scenarios, the individual stages also produced successful calibrations (stage 1 NSE 0.70 ~~-~~- 0.87, stage 2 NSE 0.78-0.87), except for the second stage in T32S_QV_ppP for the reasons explained above. The NSE for the individual calibration events in the different calibration scenarios is similar to that reported by Krebs et al. (2013). Overall, the two scenarios based on peak flow performed best (being the only CSs with mean NSE > 0.8) while the two scenarios based on percentage runoff performed worst (only CSs with mean NSE < 0.7).

~~Across the different calibration scenarios and events, the most common source of error was flow underestimation, with respect to both the total flow volume (see Figure 3, left panel) and the peak flow (see Figure 3, right panel). Volume errors for individual events were large in some cases (ranging from 35% underestimation to 30% overestimation), but the average VE for each calibration scenario was limited to underestimation by 1-11%. The magnitudes of the peak flow and volume errors are comparable to those found in previous studies on calibration of SWMM (Barco et al., 2008; Krebs et al., 2016).~~

~~Histograms of peak flow ratios (left) and volume errors (right) for all individual events in all calibration scenarios.~~

3.1.2 ~~Sensitivity to objective functions~~



2 example hydrographs run130.pdf

Figure 2. Examples of hydrographs for events with high (left) and low (right) objective function (NSE) values.

~~The differences between calibrations using NSE and RMSE as objective functions were small (see Table 4), with the largest differences being 0.05 (NSE) and 0.4 (RMSE)for T32S_QV_ppP. For three calibration scenarios the NSE calibration found a better RMSE than the RMSE calibration and for four CSs the RMSE calibration found a better NSE than the NSE calibration. This indicates that~~For the two-stage calibrations the assumption that no runoff occurred from green areas during the first stage of the calibration was checked. During the actual first-stage calibration (i.e. with green area parameters set to default values) there was no runoff from green areas for any of the calibration events in any of the calibration scenarios, so the first stage calibration attributed all runoff to impervious areas as assumed beforehand. However, some runoff occurred from green areas for first-stage events when the calibrated parameter values from the ~~algorithm does~~ second stage were applied. This runoff was caused by impervious areas draining to green areas. The runoff from green areas was <5% of the total simulated runoff volume for 4 model runs, <10% for an additional 3 runs, and 11.6%, 11.7%, 21.7%, 22.9% and 25.7% respectively for 5 additional runs. These last 5 runs concerned 3 different events with a percentage runoff (calculated before applying rainfall multipliers) between 11% and 12%. Such events may be expected to include some green area runoff and it could be considered to exclude these from the first stage calibration (not done here to limit the complexity of the procedure as discussed in Sect 2.3). In addition, all three events were also included in other first-stage calibrations where they did not result in any significant simulated green area runoff. Removing these events from the first stage of calibration based on initial calibration results would therefore result in the same event being included in different stages for different calibration scenarios, which was considered undesirable. Overall we believe that, although the assumption that all runoff is from directly connected impervious areas when QV_ppP <12% is violated in some cases~~find a local rather than a global optimum. However, the differences between them are small.~~, the assumption that these events are suitable for calibrating impervious area parameters does hold to a sufficient degree, as also evidenced by the good first-stage calibration performance (see first paragraph of this subsection). In addition, checking for green area runoff as done here is only possible after calibration, and considering it when selecting events would thus create a more complex, iterative calibration procedure, which would limit the practical applicability of this approach. We considered this to be beyond the paper's original scope of examining different strategies for calibration event selection.

3.1.2 ~~Sensitivity to~~ Low-resolution model~~discretization~~

Calibration runs with a model setup consisting of five instead of 140 subcatchments showed NSE similar to that of the baseline run (Table 4): the change in performance ranged from +0.08 (T32S_D_prec) to -0.06 (T32S_Q_60m), with only T32S_P_sum showing a larger loss of 0.15. The peak flows predicted by the low-resolution models were most often lower than in the high-resolution model and as a result, peak flow ratios were worse. This effect was stronger for the two-stage calibrations than for the single-stage calibrations. Overall runoff volume was higher in the low-resolution models, which resulted in a smaller volume error. These findings on peak flows and total flow volumes confirm earlier findings by Tscheikner-Gratl et al. (2016). The changes in peak flow performance were smaller than reported by Krebs et al. (2016), but the changes in NSE and volume errors were comparable.

3.1.3 Sensitivity to structural flow measurement errors

Calibration results (NSE) are shown in Table 4 ~~3~~ for the cases of structural flow data errors of -40% and +40%. For most calibration scenarios there was a small loss in NSE, except for T6_QV_ppP, which failed to calibrate with an NSE of -0.1 when the flow data was reduced by 40%. Three of the events in that scenario calibrated well (NSE 0.76 ~~–~~- 0.95), but the other three produced negative NSE values. These latter three events all missed the first runoff peak; for two of ~~these events~~ them the quality of fit, judged visually, was the same as in the baseline run, but since the flow rates were low, the variance of the observations was low and thus the NSE values were unsatisfactory (see Figure ~~4~~ 3 for an example). T6_PI_mean included one event, for which the reduction of flow observations by 40% resulted in a hydrograph where large parts fell below the 1 $L\ s^{-1}$ threshold. Except for the events described above, the flow errors could be compensated for in calibration~~. This issue is influenced by the use of~~, confirming the earlier findings in the literature (Dotto et al., 2014). In the paper by Dotto et al. the perturbations in flow data resulted in different calibrated values for the percentage imperviousness of the catchment, while in the current paper the perturbations resulted in different values for the rainfall multipliers as discussed in Sect. 3.2.2.

3.2 Calibrated parameter values

3.2.1 Hydrologic model parameters

Figure ~~5~~ 4 shows the calibrated parameter values (for the baseline run), normalized with respect to their calibration ranges (see Table 2). There is considerable variation among the calibrated values obtained in different calibration scenarios, demonstrating that even for parameters with a clear physical interpretation, identification of the best (ideal) value is not straightforward. Gupta et al. (1998) also found considerable variation in the parameter values obtained when using different years as calibration periods for a natural catchment model. Nonetheless, the span of parameter values is considerably reduced compared to the

4 example hydrographs flow errors varobs.pdf

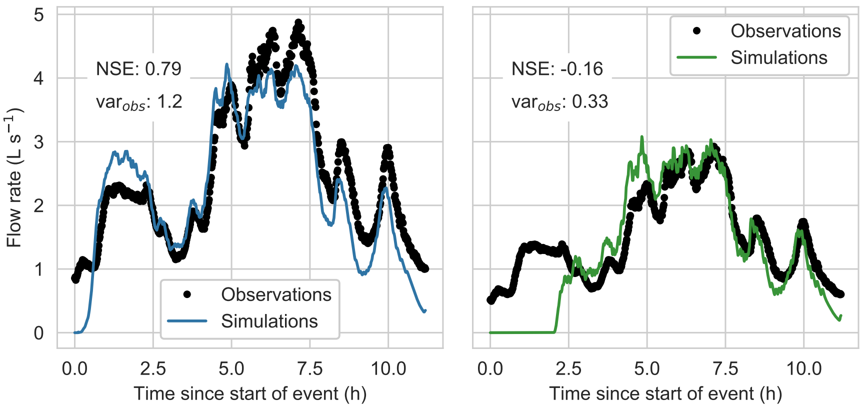


Figure 3. Calibrated hydrographs for T6_QV_ppP in the baseline run (left) and after reducing all flow measurements by 40% (right). The low NSE in the right panel is caused by the low variance of the observations.

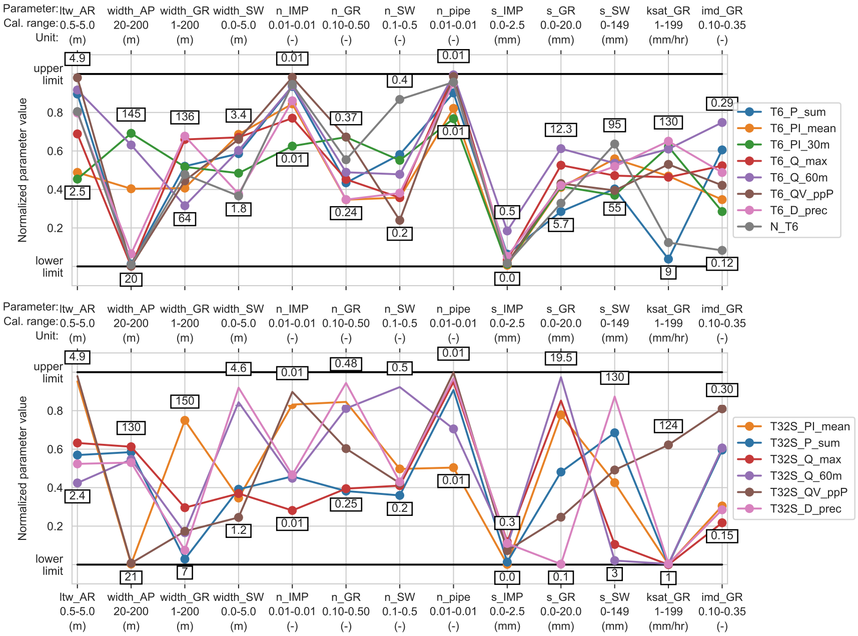
range imposed during calibration, showing that the boundaries were not set too tightly and that the calibration procedure does offer benefits over estimating parameter values directly.

Calibrated parameter values are always uncertain estimates. This uncertainty has been investigated for urban drainage models and shown to be dependent on parameter type, study catchments, model structures, catchment discretization and measurement errors (Dotto et al., 2009, 2011, 2014; Kleidorfer et al., 2009a; Sun et al., 2014). The variation found here among the optimum parameter values obtained in different calibration scenarios suggests that the selection of calibration events could also affect the uncertainty of parameter estimates and this influence should be investigated further.

3.2.2 Rainfall multipliers

The values of rainfall multipliers found in the calibration process ranged from 0.48 to 2.92, showing that there could be significant measurement errors (in precipitation and/or flow) and/or differences between the gauge rainfall and the catchment average rainfall ~~fitting best with~~ matching best the observed flow rates. For rainfall events that were included in multiple calibration scenarios, the calibrated multipliers from different scenarios were close to each other (see Table 5). This variation ~~is~~ was much smaller than that for the hydrological model parameters (see Sect. 3.2.1). ~~This indicates~~ The average value of the rainfall multipliers across all events was 1.2.

When all flow data was decreased by 40%, prior to calibration, the different CSs remained in agreement with each other, except for T6_QV_ppP, which failed in this run. The average rainfall multiplier across all events was 0.76 (i.e., 37% lower than in the run without any perturbation of flow data). When all flow data was scaled up by 40%, T32S_P_sum and T32S_Q_max produced deviating multipliers (compared to the other calibration scenarios) for three events each, but the quality of fit was the



param values.pdf

Figure 4. Normalized calibrated parameter values for different calibration scenarios and the baseline run. The highest and lowest values found for each parameter are indicated. Note to reviewers: this figure has been changed to display single and two-stage CS in separate panes and to clarify the figure.

same across all CSs (according to both the NSE and visual comparison). The average value of the multipliers across all events was 1.59 (i.e., 33% higher than in the baseline run).

The close inter-CS agreement and the similarity in between the magnitude of perturbations in flow data and the magnitude of the corresponding change in rainfall multipliers indicate that the rainfall multipliers ~~compensate~~ work as intended, i.e. compensating for discrepancies between the observed and best-fitting rainfall, rather than for other aspects of catchment runoff modelling. ~~The average value of the rainfall multipliers across all events is~~ In this respect, the average multiplier of 1.2 in the baseline run suggests that there was some structural disagreement between the observed rainfall and flows.

~~When all flow data was decreased by 40%, prior to calibration, the different CSs remained in agreement with each other, except for T6_QV_ppP, which failed in this run. The average rainfall multiplier across all events was 0.76 (i.e., 37% lower than in the run without any perturbation of flow data). When all flow data was scaled up by 40%, T32S_P_sum and T32S_Q_max produced deviating multipliers (compared to the other calibration scenarios) for three events each, but the quality of fit was the same across all CSs (according to both the NSE and visual comparison). The average value of the multipliers across all events was 1.59 (i.e., 33% higher than in the baseline run). This finding suggests that the rainfall multipliers were responsible~~

Table 5. Baseline run calibrated rainfall multipliers for events that were used in at least three CSs. Note to reviewers: this table has been updated in this revised version of the manuscript.

Event #	N_T6	T32S_D_prec	T32S_P_sum	T32S_PI_mean	T32S_Q_60m	T32S_Q_max	T32S_QV_ppP	T6_D_prec	T6_P_sum	T6_PI_30m	T6_PI_mean	T6_Q_60m	T6_Q_max	T6_QV_ppP	Mean	New P	New QV_ppP
199								0.58	0.58						0.58	8.0	21.4
209				0.48							0.48				0.48	3.8	14.3[a]
211		0.70	0.70		0.70	0.70									0.70	6.8	15.8[a]
214						1.16									1.16	7.4	8.7
222			0.68		0.68					0.68					0.68	6.7	10.6
270		1.24	1.22				1.28	1.26							1.25	11.7	9.1
306				0.74						0.70	0.74				0.73	6.3	11.7
307	1.48		1.46	1.48	1.48	1.48			1.48	1.44	1.44	1.52	1.48		1.47	44.0	11.0[b]
310				1.06	1.06					1.06	1.06	1.14			1.08	9.2	13.0
530	1.14			1.10	1.10	1.12	1.04			1.08	1.08		1.14		1.10	7.4	10.2
939		0.60													0.60	4.2	9.5
962														0.98	0.98	8.3	25.4
971							1.08								1.08	2.8	10.4
978	1.38	1.38	1.34			1.34		1.40	1.42			1.36	1.38		1.38	34.4	13.9
982	1.22			1.20							1.26	1.22	1.26		1.23	6.9	12.8
984						2.02	1.94					2.12	2.00	1.90	2.00	4.8	29.6
995							2.92							2.88	2.90	6.1	9.9 [b]
997								1.24	1.26						1.25	30.8	16.6
1001	1.70	1.66	1.60		1.64			1.66	1.66	1.60		1.64	1.70	1.64	1.65	58.2	15.1
1004														0.78	0.78	3.3	32.3
1019	1.46	1.48						1.46	1.44						1.46	32.6	14.5
1028							1.30							1.30	1.30	3.7	33.4

[a] Event percentage runoff switches from <12% to >12% when applying rainfall multiplier.

[b] Vice versa.

~~for much (if not all) of the model adjustment to the perturbed flow data. In this respect, the average multiplier of 1.2 in the baseline run suggests that there was some structural disagreement between the observed rainfall and flows.~~

~~With the~~ In runs with the low-resolution model, ~~in contrast to~~ contrarily to those with the high-resolution model, there was considerable variation in the values of the rainfall multipliers for each event found by the different calibration scenarios, ~~see Figure 6. The values obtained were~~ as shown in Figure 5. The multiplier values obtained ranged from 25% lower to 50% higher~~(~~, for the same event in the same calibration scenario~~) than in~~ , compared to the baseline calibration. Three of the low-

6 rainfall multis.png

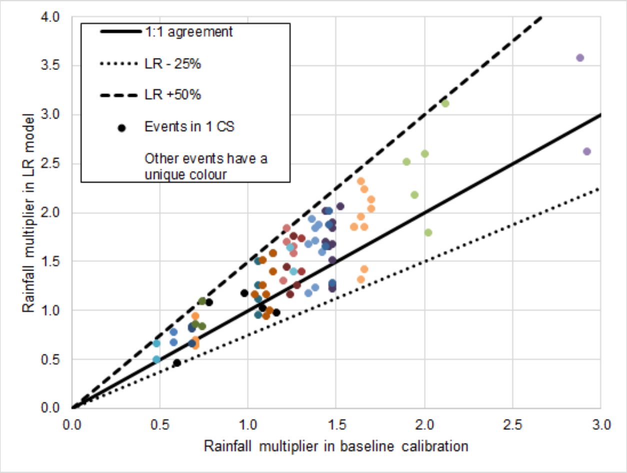


Figure 5. Rainfall multipliers in baseline calibration (horizontal axis) compared to the LR-model calibration (vertical axis). Each dot is a rainfall multiplier calibrated by one calibration scenario for one event. Identical events appearing in multiple calibration scenarios share the same colour.

resolution two-stage calibrations (T32S_D_prec, T32S_Q_60m, T32S_Q_max) found lower multipliers than in the baseline calibration, T32S_QV_ppP had three higher and three lower multipliers and other CSs had all higher multipliers. This behaviour indicates that~~(despite similar resulting performance)~~ , in spite of yielding similar results, the rainfall multipliers in the LR-model were used to compensate (within a single event) for the effects of the specific parameter set found in calibration, rather than to compensate for a structural discrepancy between the observed rainfall and flow data as in the baseline calibration ~~.~~ (as was the case for the HR models). That the rainfall multipliers appear to behave in a more physical way in the high-resolution model is in line with earlier findings about more transferable parameter values resulting from high-resolution models (Krebs et al., 2014; Sun et al., 2014).

3.3 Validation performance

3.3.1 Individual events

The successful calibrations predicted ~~7-13~~ 8-13 out of the 19 validation events satisfactorily (NSE ~~>~~≥0.5), see Table 6. T6_PI_30m (9 events) and T6_Q_60m (8 events) performed worst while T32S_PI_mean performed best. Perturbations of the flow data in the calibration period led to a lower number of satisfactorily predicted events for most CSs. The two-stage calibration scenarios were less sensitive to perturbations of the flow data in the calibration period~~and to~~ , i.e. they predicted more validation events satisfactorily than their single-stage counterparts. When switching from the high resolution to the low-resolution model ~~,~~ the single-stage CSs were no longer able to predict up to 5 events, while from the two-stage CSs only T32S_D_prec lost two events, and T32S_P_sum, T32S_Q_max, and T32S_QV_ppP actually predicted a higher number of

Table 6. Number of validation events with NSE >0.5 out of 19 total events. Bold font indicates the best value in each column. Note to reviewers: this table has been updated in this revised version of the manuscript

	Baseline	Cal. flow -40%	Cal. flow +40%	Low-res. model	Total
N_T6	12	10	8	7	37
T6_D_prec	11	9	9	6	35
T6_P_sum	11	9	9	8	37
T6_PI_30m	9	9	9	9	36
T6_PI_mean	10	6	**12**	6	34
T6_Q_60m	8	9	9	6	32
T6_Q_max	12	9	11	10	42
T6_QV_ppP	12	7[a]	9	10	31
T32S_D_prec	12	**12**	**12**	10	46
T32S_P_sum	10	9	10	**13**	42
T32S_PI_mean	**13**	**12**	**12**	**13**	**50**
T32S_Q_60m	10	9	9	10	38
T32S_Q_max	11	8	10	12	41
T32S_QV_ppP	11	**12**	10	12	45

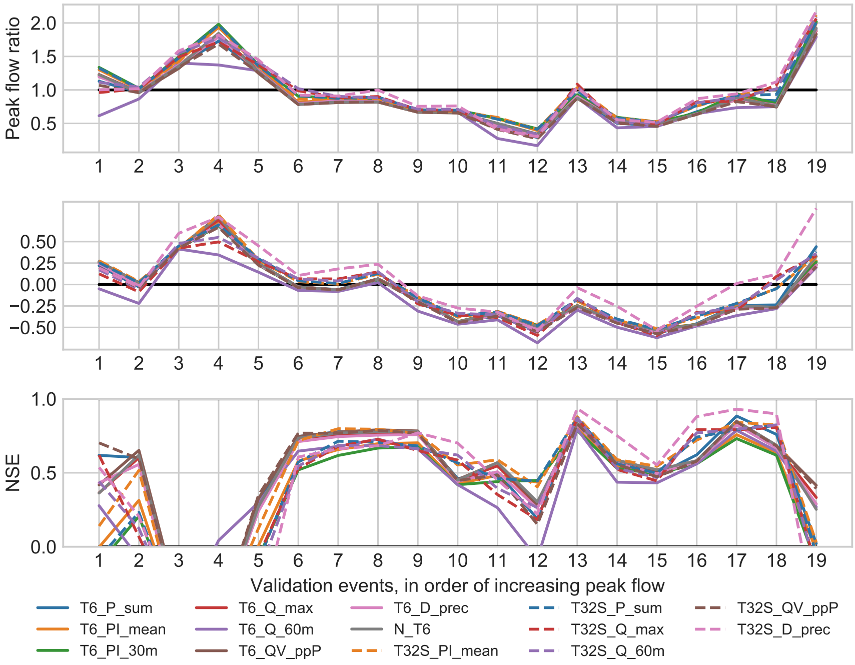
[a] Run was unsuccessful in calibration

events satisfactorily~~with the low-resolution model than with the calibrated high resolution model~~. Over all four calibration runs, the two-stage calibrations were able to predict more events satisfactorily than their single-stage counterparts.

The events that most often caused failure in validation were four events with peak flow rates of 10 L s^{-1} or less, and therefore, such failures may be attributed to: (~~i~~1) relatively high measurement uncertainties, and (~~ii~~2) high sensitivity of the NSE to even small changes in the hydrographs. However, it should be noted that the two smallest events (both with a peak flow rate of 4.6 L s^{-1}) were predicted with NSE>0.5 by some calibration scenarios. For the other CSs, examination of the hydrographs showed that they predict well the magnitude of events, but produce wrong timing.

Another event that failed in validation for all CSs was that with the highest peak flow rate (53 L s^{-1}, see Table A1), which was overestimated by a factor of up to three. This event was dominated by an intense, single-peak burst of rainfall (the highest 30-minute average rainfall intensity was 11.1 mm hr^{-1}), so it could have suffered from high spatial variation of the rainfall.

The ~~volume errors were similar for all high-resolution calibrated models and showed a general tendency to underestimate flow volumes by 25%. When using the low-resolution model, the single-stage CSs underestimated runoff volume by around 40%, while two-stage scenarios underestimated it by a maximum of 27%. Across all CSs, two-stage versions had similar or better performance in terms of total runoff volume. Peak flow ratios were <1 for most events, but for the events that generally did poorly in validation (see above) peak flows (as well as flow volumes) were over predicted instead. The results for both total~~



7 error stats validation events b.pdf

Figure 6. Error statistics for individual validation events for all calibration scenarios in the baseline runs.

~~volumes and peak flows indicate that for most events flows were underestimated, which may be (at least partially) attributed to the discrepancies between observed rainfall and flow found in the calibration phase (see Sect. 3.2.2).~~

~~The peak flow~~ peak flow ratios obtained for the 19 validation events using the calibrated models from the baseline are shown in the upper panel of Figure ~~7.~~ 6. Under- or overestimation of peak flows and runoff volumes by the model could lead to an under- or over-dimensioned system design, and it is therefore relevant to consider these aspects alongside the NSE. Underestimation of peak flows was most frequent, but the largest errors ~~occured~~ occurred when the flow was overestimated. The variation among CSs was generally larger when the prediction error was larger. The corresponding figure for volume errors is shown in the middle panel of Figure ~~7.~~ 6. Again, underestimation was more common, but overestimation did occur for a limited number of events. For both peak flows and total volumes, the variation among events was generally larger than the variation among different calibration scenarios, showing that selecting a limited number of validation events may also influence the results of the model evaluation. Across all CSs, two-stage versions had similar or better performance in terms of total runoff volume. Peak flow ratios were <1 for most events, but for the events that generally did poorly in validation (see above) peak flows (as well as flow volumes) were over predicted instead. The results for both total volumes and peak flows indicate that for most events flows were underestimated, which may be (at least partially) attributed to the discrepancies between observed rainfall and flow found in the calibration phase (see Sect. 3.2.2).

When examining the NSE of the validation events (see the bottom panel of Figure 7), more variation among the different CSs became visible, although the amount of variation was still event-dependent: inter-CS variation for the same events varies from 0.15 to 1.25. This shows that some events can have a much larger impact on the overall validation results than others. Out of the 19 events, 6 were predicted satisfactorily (NSE>0.5) by some CSs but not by others; 5 events failed for all CSs, and 8 were predicted satisfactorily by all CSs.

3.3.2 Overall performance of the high-resolution model

To assess the overall performance of different calibration scenarios for the validation period, several ways of combining the individual events were considered (see Table 7). The simplest metric is obtained by using the NSE means, which ranged from 0.13 (T6_PI_30m) to 0.42 (T32S_QV_ppP). There are two conceptual problems with this metric: First, since NSE ranges from negative infinity to plus one, one poorly fitting event can offset multiple well-fitting events. Second, two simulated hydrographs of equally poor fit can have rather different (negative) NSE values, producing different impacts on the overall results, which is not justified by a visual comparison. Therefore, this mean metric is not considered a reliable metric for comparisons, when poorly fitting events are present.

The exclusion of low flow (<10 L s^{-1} peak) events avoids this issue, but does not reward calibration scenarios that do manage to predict these events satisfactorily. Another option is to set all NSE values <-1 to -1 before calculating the mean, which results in NSE ranging from 0.29 to 0.47. Adoption of the median NSEs (insensitive to outliers) lead to a higher range of 0.43 to 0.61, showing that the average or overall validation performance depends more on the outlier events than on typical events.

A more commonly used approach is to combine all the events into a single time series prior to calculating the NSE on the joint time series. This procedure indicated satisfactory performance for all CSs (NSE 0.57 - 0.70). The discussion of various metrics shows that caution is needed when averaging performance over multiple events, as metrics may not reflect the fact that a significant number of events is poorly predicted in all CSs (see Table 6).

The considerations in the previous paragraph concern the NSE and are not necessarily applicable to other statistics in the same way. The ~~RMSE is calculated in flow units (L s^{-1}) and tends towards larger values for larger events, even if the fit is visually better. Because of this taking the mean across events is somewhat conceptually unsatisfactory, but the resulting values differ from the RMSE calculated on a joint time series only by an offset that is almost the same for all CSs. Therefore, all CSs show the same relative performance. The~~ volume error (VE) was included in this study to yield some indication of the overall difference between the modelled and observed runoff volumes over longer time periods. Therefore, this statistic was summarized over all events using the joint time-series approach.

~~To obtain an overall ranking of the different CSs in the baseline run, they were ranked by five characteristics (see Table 7) and then the sum of the individual ranks was taken. This shows that the two-stage CSs performed better in the validation period than the~~ The volume errors were similar for all high-resolution single-stage ~~CSs.~~

3.3.3 ~~Sensitivity to the objective function~~

~~For most calibration scenarios, the models that were calibrated with different objective functions (NSE in the baseline run, RMSE in the alternative) retained a similar performance in the validation phase. However, there are differences for some of the~~

Table 7. Summarized performance for all 19 validation events for the high-resolution model. Bold font indicates the best value in each column. Note to reviewers: this table has been updated in this revised version of the manuscript

	Mean NSE	Clip mean NSE	Median NSE	Joint NSE	# neg NSE	# good NSE	Joint VE	Mean PFR
N_T6	0.33	0.45	0.58	0.65	2	12	-0.24	0.91
T6_P_sum	0.39	0.45	0.60	0.66	2	12	-0.23	0.91
T6_PI_mean	0.18	0.33	0.51	0.59	4	10	-0.24	0.96
T6_PI_30m	0.13	0.29	0.49	0.57	**5**	9	-0.24	0.98
T6_Q_max	0.34	0.44	0.58	0.65	2	12	-0.24	0.92
T6_Q_60m	0.37	0.37	0.43	0.60	3	8	-0.29	0.81
T6_QV_ppP	0.36	**0.47**	0.58	0.67	2	12	-0.24	0.90
T6_D_prec	0.34	0.43	0.56	0.64	2	11	-0.25	0.91
T32S_P_sum	0.19	0.34	0.56	0.68	**5**	10	-0.15	0.99
T32S_PI_mean	0.26	0.44	0.59	**0.70**	2	**13**	-0.16	**1.00**
T32S_Q_max	0.31	0.34	0.53	0.67	4	11	-0.13	0.96
T32S_Q_60m	0.26	0.33	0.53	0.68	4	10	-0.13	0.99
T32S_QV_ppP	**0.42**	0.46	0.58	0.65	2	11	-0.26	0.87
T32S_D_prec	0.22	0.34	**0.61**	**0.70**	4	12	**-0.02**	1.01

calibrated models and showed a general tendency to underestimate flow volumes by 25%. For the two-stage ~~CSs, see Table ?? for a description and Figure 8 for an example.~~

calibrated models volume errors were smaller with underestimation of around 15% (except for T32S_QV_ppP).

~~Examples of hydrographs showing typical (left panel, N_T6) and differing (right panel, T32S_D_prec) behaviour when calibrated for different objective functions.~~

3.3.3 ~~Low-resolution~~ Overall performance of the low-resolution model

The effect of the low-resolution model depended on the calibration scenario considered, see Table 8. Some scenarios scored better in terms of NSE (gains of up to 0.17 and 3 events predicted with NSE >0.5), while others lost performance by the same metrics (up to 0.24 and 5 events). This is a more ~~mixed result~~ less consistent than that found by Krebs et al. (2016), who tested high- and low-resolution models of three catchments and found the high-resolution models to perform better in validation for all three. All but one of the two-stage scenarios predicted more events satisfactorily with the low-resolution model than with the high-resolution model.

Table 8. Summarized validation performance (over 19 events) for the low-resolution models. Bold font indicates the best value in each column. Note to reviewers: this table is completely new and not the same as Table 8 in the previous version of the manuscript

	Mean NSE	Clip mean NSE	Median NSE	Joint NSE	# neg NSE	# good NSE	Joint VE	Mean PFR	LR visually better than HR (# events)
N_T6	0.12	0.21	0.36	0.52	5	7	-0.43	0.50	2
T6_P_sum	0.05	0.22	0.42	0.57	6	8	-0.38	0.60	3
T6_PI_mean	0.38	0.38	0.37	0.50	**0**	6	-0.43	0.59	4
T6_PI_30m	0.43	0.43	0.50	0.58	2	9	-0.34	0.74	5
T6_Q_max	0.49	0.49	0.56	0.59	**0**	10	-0.36	0.64	5
T6_Q_60m	0.29	0.29	0.36	0.49	4	6	-0.46	0.49	3
T6_QV_ppP	0.37	0.37	0.51	0.54	3	10	-0.40	0.66	4
T6_D_prec	0.34	0.34	0.38	0.50	4	6	-0.44	0.51	4
T32S_P_sum	**0.51**	**0.51**	0.55	0.66	2	**13**	-0.27	0.60	4
T32S_PI_mean	0.44	0.46	0.60	0.69	2	**13**	-0.22	0.80	5
T32S_Q_max	0.05	0.33	0.64	0.70	5	12	-0.07	1.03	**12**
T32S_Q_60m	0.13	0.28	0.52	0.66	4	10	**-0.04**	**1.02**	11
T32S_QV_ppP	0.44	0.46	**0.65**	0.72	2	12	-0.18	0.79	7
T32S_D_prec	0.29	0.38	0.56	**0.76**	4	10	-0.05	0.86	4

[a] calculated after setting individual event values <-1 to -1.

~~The volume errors~~ For the single-stage calibration scenarios, the volume errors in the LR were twelve to nineteen percent points higher~~for the single-stage calibration scenarios~~. The two-stage scenarios showed both worsened performance (T32S_P_sum, T32S_PI_mean) and improved performance (T32S_Q_60m and T32S_Q_max, T32S_QV_ppP). When comparing the hydrographs from the two different model discretizations per event, the high-resolution model usually performed better. However, for the last three CSs mentioned, the low-resolution performed better compared to the other CSs. For T32S_Q_60m and T32S_Q_max, the low-resolution model predicted the observed hydrographs better for most validation events. These three calibration scenarios were also the only ones where the low-resolution model resulted in lower values for the calibrated rainfall multipliers.

3.3.4 Sensitivity to structural flow errors

The introduction of structural flow measurement errors ~~in~~ into the calibration data had little effect on performance in the validation phase. Although there were some changes (compared to the baseline calibration) in the overall NSE values, volume errors and peak flow ratios were almost the same for the baseline and disturbed flow data runs. For T6_D_prec, T6_P_sum, T6_Q_60m, and T6_QV_ppP, runoff started later in the validation event when calibration flow data was increased by 40%, but this had a limited influence on the overall performance metrics (NSE, VE and PFR). Only T6_PI_mean was more sensitive to reducing calibration flow data by 40%. This resulted in lower flows (and therefore better fits) in validation events for the five events that caused problems for most other CSs (i.e. the four lowest and the single highest peak flow rate(s), see Sect. 3.3.1).

3.3.5 ~~Overall ranking for validation~~

~~For an overall ranking of the different calibration scenarios in the validation period the baseline runs were ranked by each of the following statistics: mean NSE (limited to -1), number of events with NSE >0.5, RMSE (calculated over the joint time series of all events), volume error (see RMSE), and mean peak flow ratio. The ranks for each characteristic were then summed to obtain an overall ranking, see Table ??. T32S_PI_mean and T32S_D_prec performed best, with T6_PI_30m and T6_Q_60m bringing up the rear.~~

3.4 Degradation of performance from calibration to validation

In calibration, the NSE for the different calibration scenarios ranged from 0.68 to 0.85, while in validation it ranged from 0.29 to 0.47~~Table ??~~. The CSs that did better in calibration lost more performance (measured by NSE) when switching to the validation ~~period, see figure 9. Considering the change in overall rank from calibration to validation, the~~ phase (see Figure 7). The range of performance loss for the different calibration scenarios was larger for the low-resolution model than for the high-resolution model. For the high resolution model all but one of the two-stage ~~scenarios showed smaller changes then the~~ calibrations lost more performance when switching to the validation phase than their single-stage ~~scenarios. Several scenarios showed large gains (+10 for T6_QV_ppP, +7 for T6_P_sum, +5 for T32S_PI_mean) while the largest losses were smaller (-7 for T6_Q_60m, -6 for T6_Q_max)~~counterparts, whereas for the low-resolution model all but one of the two stage calibrations had a smaller performance loss. The findings in this ~~Sect.~~ section demonstrate that good calibration performance is not necessarily indicative of good validation performance and vice versa, and therefore, whenever feasible, validation should be performed~~, if at all possible.~~ . Previous studies found that high-resolution models lead to more transferable parameter estimates (e.g. less loss of performance when switching to validation, Sun et al. (2014), Krebs et al. (2014)), but in the current study this seems dependent on the calibration data set used. For the two-stage calibrations the low-resolution model usually has less loss in performance than the high resolution model.

3.5 Single-stage vs. two-stage calibrations

For those selection criteria, for which both single and two-stage calibrations were performed, the results of the two options ~~were~~ can be compared directly (see ~~Table ??). In terms of NSE and volume error,~~ Figure 8). For the high-resolution model, calibration performance of the two-stage ~~calibrations performed better than~~ CSs was somewhat better than for the single-stage ~~calibrations, except for Q_max. In terms of peak flow ratio the results were mixed. For D_prec and PI_mean the two-stage variant outperformed the single-stage across all metrics, for Q_max~~ CSs. By contrast, in the validation phase the NSE was better for the single-stage ~~variant performed better and for other CSsthe results depended on the metric used. In validation the~~

~~Loss of performance (NSE) when switching from calibration to validation~~ 9 v20190710 loss val performance.pdf

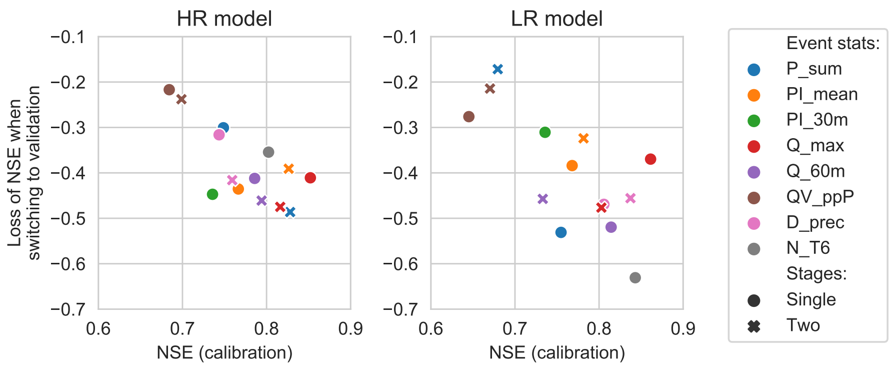


Figure 7. Loss of performance (NSE) when switching from calibration to validation.

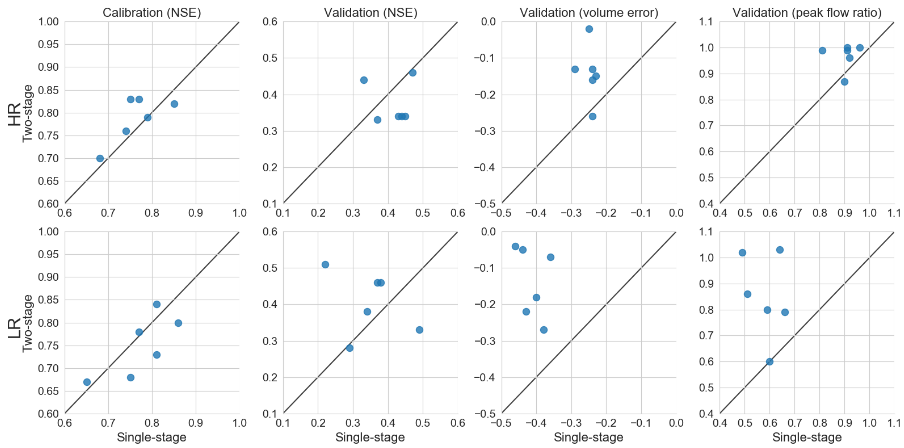


Figure 8. Comparison of single-stage and two-stage calibration strategies.

~~differences between single and~~ CSs. However, the volume error and peak flow ratio were better for the two-stage ~~calibration were less pronounced, see Table ??. In terms of NSE, the single-stage calibrationsperformed better, but they had the same number of satisfactorily predicted events as~~ calibrations. For the low-resolution model performance was similar or worse for the two-stage calibrations, but in the validation phase the two-stage calibrations ~~. In terms of RMSE, VE and PFR~~ most often had higher NSE. In addition, the two-stage calibrations ~~performed better , except for QV_ppP. This is also the only criterion where all metrics indicated the same, i.e. that the~~ resulted in much better performance in terms of volume error and peak flows than their single-stage ~~calibration had better results in the validation period~~counterparts.

4 Conclusions

The primary objective of this study was to compare different strategies for the selection of calibration events for a hydrodynamic model of a predominantly green urban area. Two secondary objectives were to verify (1) whether earlier findings on other sources of uncertainty in urban drainage modelling also apply to a greener urban catchment, and (2) whether they are sensitive to the calibration data set used. Calibration strategies consisted of single- and two stage calibrations and considered a number of different metrics by which calibration events can be selected from a larger group of candidate events. Calibration strategies were tested with ~~two different objective functions,~~ high and low spatial resolution models and on data sets with structural flow data errors~~, and with high and low spatial resolution models~~. The conclusions drawn below are strictly valid for the specific data and catchment characteristics used in this study.

In the baseline run (high resolution model, ~~Nash-Sutcliffe as objective function,~~ no structural flow data errors), all calibration scenarios produced successful calibrations ~~,~~ (i.e., NSE $>$ 0.5), albeit with varying performance: NSE values ranged from 0.68 to 0.85. For the two-stage calibrations, both stages gave satisfactory results (NSE 0.70-0.87). The two-stage calibrations performed better than their single-stage counterparts in terms of NSE and runoff volume error. The ~~choice of NSE or RMSE as the objective function~~ two-stage calibrations also were faster since they reduced the dimensionality (number of simultaneously calibrated parameters) of the calibration problem. Although the obtained values of the SWMM model parameters varied between the different CSs, they found highly similar values for the rainfall multipliers included in the calibration. Switching from a high-resolution to a low-resolution model discretization had only a small impact on ~~the results.~~ calibration performance metrics. However, the values of the rainfall multipliers for each event showed much more variation than with the high-resolution models. Most high-resolution calibration models produced higher values of the multipliers, except for three two-stage CSs, which produced lower values instead. These observations on the rainfall multipliers in low and high-resolution models are in line with previous studies (Krebs et al., 2014; Sun et al., 2014).

The robustness of the calibration scenarios to structural flow errors was tested by calibrating them after uniformly reducing or increasing all flow observations by 40%. Most calibration scenarios were able to adjust to this with only small effects on the calibration performance, except for T6_QV_ppP (six events with highest percentage runoff), which failed in calibration (NSE -0.1) when flow data was reduced by 40%. This can be attributed to two low-flow events, which produced negative NSE values, even though they visually indicated a good fit.

~~Switching from a high-resolution to a low-resolution model discretization has only a small impact on calibration performance metrics. However, the values of the rainfall multipliers for each event show much more variation than with the high-resolution models. Most high-resolution calibration models find higher values for the multipliers, but three two-stage CSs find lower values instead~~This compensation for errors in the calibration data confirms earlier findings from a predominantly impervious catchment(Dotto et al., 2014) for a predominantly green catchment, and confirms that. these findings were insensitive to calibration data selection except in the case of T6_QV_ppP.

The calibrated scenarios were validated against an independent set of 19 validation events. All calibrated scenarios predicted 7 to 13 of these events satisfactorily (NSE ~~>~~>0.5). A group of four events with peak flow rates of less than 10 L s^{-1} caused problems in most calibration scenarios, as did the event with the highest observed peak flow rate. Although most calibration scenarios yielded similar results for the validation events with respect to the overall volume error and the ratio between the modelled and observed peak flow rates, there were considerable differences between the CSs when performance for the validation events was measured by NSE. In terms of NSE the single-stage CSs proved more successful in the validation phase, but for ~~RMSE,~~ volume error and peak flow error the two-stage CSs performed better. Better performance in regards to flow volumes and peak flows bears more significance for engineering design.

~~In the validation phase, there were again (as in the calibration) only small differences between the two considered objective functions.~~ Concerning model discretization, the low-resolution single-stage calibration scenarios ~~show~~ showed significantly larger volume errors than their high-resolution counterparts, while most two-stage calibration scenarios ~~show~~ showed either the same or even improved volume errors. Two of the two-stage CSs (that also deviated from the others in terms of the calibrated rainfall multipliers) were also the only ones to obtain visually better fitting hydrographs with the low-resolution model setup than with the ~~high-resolution~~ high resolution model setup. Two-stage calibrations also predicted more validation events satisfactorily when the calibration flow data was perturbed.

~~An overall ranking of the different scenarios across the different influential factors (objective function, flow data errors, model discretization) showed that T6_Q_max, T32S_D_prec and N_T6 performed the best in calibration. However, in the validation phase this order was changed considerably with T32S_PI_mean, T32S_D_prec and T6_P_sum forming the top three. The ranking of~~ Earlier studies found that high-resolution models lost less performance when switching to the validation phase (Krebs et al., 2014; Sun et al., 2014), but, in the ~~two-stage scenarios was more consistent between calibration and validation than that of the single-stage scenarios~~current paper, this depended on the set of calibration data that was selected.

Author contributions. Ico Broekhuizen maintained the field measurements, validated the data, designed and carried out the simulation experiments, analyzed the results, and drafted the paper. Günther Leonhardt, Jiri Marsalek and Maria Viklander provided feedback on the design of the simulation experiments and reviewed the paper drafts.

Competing interests. The authors declare that they have no conflicts of interest.

Acknowledgements. This study was funded by the Swedish Research Council Formas (grant number 2015-121). The authors would like to thank CHI/HydroPraxis for providing a license for PCSWMM. The authors would also like to thank Helen Galfi, Ralf Rentz and Karolina Berggren for their work in setting up and maintaining the field measurements.

References

Aguilar, M. F., McDonald, W. M., and Dymond, R. L.: Benchmarking laboratory observation uncertainty for in-pipe storm sewer discharge measurements, Journal of Hydrology, 534, 73–86, https://doi.org/10.1016/j.jhydrol.2015.12.052, https://linkinghub.elsevier.com/retrieve/pii/S0022169415010008, 2016.

Barco, J., Wong, K. M., and Stenstrom, M. K.: Automatic Calibration of the U.S. EPA SWMM Model for a Large Urban Catchment, Journal of Hydraulic Engineering, 134, 466–474, https://doi.org/10.1061/(ASCE)0733-9429(2008)134:4(466), http://ascelibrary.org/doi/10.1061/%28ASCE%290733-9429%282008%29134%3A4%28466%29, 2008.

Blake, J. R. and Packman, J. C.: Identification and correction of water velocity measurement errors associated with ultrasonic Doppler flow monitoring, Water and Environment Journal, 22, 155–167, https://doi.org/10.1111/j.1747-6593.2007.00089.x, http://doi.wiley.com/10.1111/j.1747-6593.2007.00089.x, 2008.

Bonakdari, H. and Zinatizadeh, A. A.: Influence of position and type of Doppler flow meters on flow-rate measurement in sewers using computational fluid dynamic, Flow Measurement and Instrumentation, 22, 225–234, https://doi.org/10.1016/j.flowmeasinst.2011.03.001, http://linkinghub.elsevier.com/retrieve/pii/S0955598611000288, 2011.

Datta, A. R. and Bolisetti, T.: Uncertainty analysis of a spatially-distributed hydrological model with rainfall multipliers, Canadian Journal of Civil Engineering, 43, 1062–1074, https://doi.org/10.1139/cjce-2015-0413, http://www.nrcresearchpress.com/doi/10.1139/cjce-2015-0413, 2016.

Del Giudice, D., Albert, C., Rieckermann, J., and Reichert, P.: Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation, Water Resources Research, 52, 3162–3186, https://doi.org/10.1002/2015WR017871, http://doi.wiley.com/10.1002/2015WR017871, 2016.

Deletic, A., Dotto, C., McCarthy, D., Kleidorfer, M., Freni, G., Mannina, G., Uhl, M., Henrichs, M., Fletcher, T., Rauch, W., Bertrand-Krajewski, J., and Tait, S.: Assessing uncertainties in urban drainage models, Physics and Chemistry of the Earth, Parts A/B/C, 42-44, 3–10, https://doi.org/10.1016/j.pce.2011.04.007, http://linkinghub.elsevier.com/retrieve/pii/S1474706511000623, 2012.

Dongquan, Z., Jining, C., Haozheng, W., Qingyuan, T., Shangbing, C., and Zheng, S.: GIS-based urban rainfall-runoff modeling using an automatic catchment-discretization approach: a case study in Macau, Environmental Earth Sciences, 59, 465–472, https://doi.org/10.1007/s12665-009-0045-1, http://link.springer.com/10.1007/s12665-009-0045-1, 2009.

Dotto, C., Kleidorfer, M., Deletic, A., Rauch, W., McCarthy, D., and Fletcher, T.: Performance and sensitivity analysis of stormwater models using a Bayesian approach and long-term high resolution data, Environmental Modelling & Software, 26, 1225–1239, https://doi.org/10.1016/j.envsoft.2011.03.013, http://linkinghub.elsevier.com/retrieve/pii/S1364815211000880, 2011.

Dotto, C., Mannina, G., Kleidorfer, M., Vezzaro, L., Henrichs, M., McCarthy, D. T., Freni, G., Rauch, W., and Deletic, A.: Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling, Water Research, 46, 2545–2558, https://doi.org/10.1016/j.watres.2012.02.009, http://linkinghub.elsevier.com/retrieve/pii/S0043135412000978, 2012.

Dotto, C., Kleidorfer, M., Deletic, A., Rauch, W., and McCarthy, D.: Impacts of measured data uncertainty on urban stormwater models, Journal of Hydrology, 508, 28–42, https://doi.org/10.1016/j.jhydrol.2013.10.025, http://linkinghub.elsevier.com/retrieve/pii/S0022169413007440, 2014.

Dotto, C. B. S., Deletic, A., and Fletcher, T. D.: Analysis of parameter uncertainty of a flow and quality stormwater model, Water Science and Technology, 60, 717–725, https://doi.org/10.2166/wst.2009.434, https://iwaponline.com/wst/article/60/3/717/15644/Analysis-of-parameter-uncertainty-of-a-flow-and, 2009.

Duan, Q., Sorooshian, S., and Gupta, V. K.: Optimal use of the SCE-UA global optimization method for calibrating watershed models, Journal of Hydrology, 158, 265–284, https://doi.org/10.1016/0022-1694(94)90057-4, http://linkinghub.elsevier.com/retrieve/pii/0022169494900574, 1994.

Duchon, C. E.: Results of Laboratory and Field Calibration-Verification Tests of Geonor Vibrating Wire Transducers from March 2000 to July 2002, Tech. rep., School of Meteorology University of Oklahoma. Prepared for U.S. Climate Reference Network Management Office, 2002.

Duchon, C. E. and Essenberg, G. R.: Comparative rainfall observations from pit and aboveground rain gauges with and without wind shields, Water Resources Research, 37, 3253–3263, https://doi.org/10.1029/2001WR000541, http://doi.wiley.com/10.1029/2001WR000541, 2001.

Elliott, A. and Trowsdale, S.: A review of models for low impact urban stormwater drainage, Environmental Modelling & Software, 22, 394–405, https://doi.org/10.1016/j.envsoft.2005.12.005, http://linkinghub.elsevier.com/retrieve/pii/S1364815206000053, 2007.

Fletcher, T., Andrieu, H., and Hamel, P.: Understanding, management and modelling of urban hydrology and its consequences for receiving waters: A state of the art, Advances in Water Resources, 51, 261–279, https://doi.org/10.1016/j.advwatres.2012.09.001, http://linkinghub.elsevier.com/retrieve/pii/S0309170812002412, 2013.

Fuentes-Andino, D., Beven, K., Kauffeldt, A., Xu, C.-Y., Halldin, S., and Di Baldassarre, G.: Event and model dependent rainfall adjustments to improve discharge predictions, Hydrological Sciences Journal, 62, 232–245, https://doi.org/10.1080/02626667.2016.1183775, https://www.tandfonline.com/doi/full/10.1080/02626667.2016.1183775, 2017.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resources Research, 34, 751–763, https://doi.org/10.1029/97WR03495, http://doi.wiley.com/10.1029/97WR03495, 1998.

Heiner, B. J. and Vermeyen, T. B.: Laboratory Evaluation of Open Channel Area-Velocity Flow Meters:, Technical HL-2012-03, Denver, CO, USA, 2012.

Hernebring, C.: 10års-regnets återkomst – förr och nu: regndata för dimensioneringkontroll-beräkning av VA-system i tätorter. (Design storms in Sweden – then and now. Rain data for design and control of urban drainage systems), Tech. Rep. 2006-04, Svenskt Vatten AB, https://vattenbokhandeln.svensktvatten.se/produkt/10-ars-regnets-aterkomst-forr-och-nu-regndata-for-dimensionering-kontrollberakning-av-va-system-i-tatorter/, 2006.

Houska, T., Kraft, P., Chamorro-Chavez, A., and Breuer, L.: SPOTting Model Parameters Using a Ready-Made Python Package, PLOS ONE, 10, e0145 180, https://doi.org/10.1371/journal.pone.0145180, http://dx.plos.org/10.1371/journal.pone.0145180, 2015.

Kavetski, D., Kuczera, G., and Franks, S. W.: Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, Journal of Hydrology, 320, 173–186, https://doi.org/10.1016/j.jhydrol.2005.07.012, http://www.sciencedirect.com/science/article/pii/S0022169405003379, 2006.

Kleidorfer, M., Deletic, A., Fletcher, T. D., and Rauch, W.: Impact of input data uncertainties on urban stormwater model parameters, Water Science and Technology, 60, 1545–1554, https://doi.org/10.2166/wst.2009.493, https://iwaponline.com/wst/article/60/6/1545/15890/Impact-of-input-data-uncertainties-on-urban, 2009a.

Kleidorfer, M., Möderl, M., Fach, S., and Rauch, W.: Optimization of measurement campaigns for calibration of a conceptual sewer model, Water Science and Technology, 59, 1523–1530, https://doi.org/10.2166/wst.2009.154, https://iwaponline.com/wst/article/59/8/1523/12900/Optimization-of-measurement-campaigns-for, 2009b.

Krebs, G., Kokkonen, T., Valtanen, M., Koivusalo, H., and Setälä, H.: A high resolution application of a stormwater management model (SWMM) using genetic parameter optimization, Urban Water Journal, 10, 394–410, https://doi.org/10.1080/1573062X.2012.739631, http://www.tandfonline.com/doi/abs/10.1080/1573062X.2012.739631, 2013.

Krebs, G., Kokkonen, T., Valtanen, M., Setälä, H., and Koivusalo, H.: Spatial resolution considerations for urban hydrological modelling, Journal of Hydrology, 512, 482–497, https://doi.org/10.1016/j.jhydrol.2014.03.013, http://linkinghub.elsevier.com/retrieve/pii/S0022169414001875, 2014.

Krebs, G., Kokkonen, T., Setälä, H., and Koivusalo, H.: Parameterization of a Hydrological Model for a Large, Ungauged Urban Catchment, Water, 8, 443, https://doi.org/10.3390/w8100443, http://www.mdpi.com/2073-4441/8/10/443, 2016.

Lanza, L. G., Vuerich, E., and Gnecco, I.: Analysis of highly accurate rain intensity measurements from a field test site, Advances in Geosciences, 25, 37–44, https://doi.org/10.5194/adgeo-25-37-2010, https://www.adv-geosci.net/25/37/2010/, 2010.

Lepot, M., Momplot, A., Lipeme Kouyi, G., and Bertrand-Krajewski, J.-L.: Rhodamine WT tracer experiments to check flow measurements in sewers, Flow Measurement and Instrumentation, 40, 28–38, https://doi.org/10.1016/j.flowmeasinst.2014.08.010, http://linkinghub.elsevier.com/retrieve/pii/S0955598614000983, 2014.

Maheepala, U., Takyi, A., and Perera, B.: Hydrological data monitoring for urban stormwater drainage systems, Journal of Hydrology, 245, 32–47, https://doi.org/10.1016/S0022-1694(01)00342-0, http://linkinghub.elsevier.com/retrieve/pii/S0022169401003420, 2001.

Mancipe-Munoz, N. A., Buchberger, S. G., Suidan, M. T., and Lu, T.: Calibration of Rainfall-Runoff Model in Urban Watersheds for Stormwater Management Assessment, Journal of Water Resources Planning and Management, 140, 05014 001, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000382, http://ascelibrary.org/doi/10.1061/%28ASCE%29WR.1943-5452.0000382, 2014.

Mourad, M., Bertrand-Krajewski, J.-L., and Chebbo, G.: Stormwater quality models: sensitivity to calibration data, Water Science and Technology, 52, 61–68, https://doi.org/10.2166/wst.2005.0110, https://iwaponline.com/wst/article/52/5/61/12267/Stormwater-quality-models-sensitivity-to, 2005.

Muleta, M. K., McMillan, J., Amenu, G. G., and Burian, S. J.: Bayesian Approach for Uncertainty Analysis of an Urban Storm Water Model and Its Application to a Heavily Urbanized Watershed, Journal of Hydrologic Engineering, 18, 1360–1371, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000705, http://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000705, 2013.

Nord, G., Gallart, F., Gratiot, N., Soler, M., Reid, I., Vachtman, D., Latron, J., Martín-Vide, J. P., and Laronne, J. B.: Applicability of acoustic Doppler devices for flow velocity measurements and discharge estimation in flows with sediment transport, Journal of Hydrology, 509, 504–518, https://doi.org/10.1016/j.jhydrol.2013.11.020, http://linkinghub.elsevier.com/retrieve/pii/S0022169413008366, 2014.

Petrucci, G. and Bonhomme, C.: The dilemma of spatial representation for urban hydrology semi-distributed modelling: Trade-offs among complexity, calibration and geographical data, Journal of Hydrology, 517, 997–1007, https://doi.org/10.1016/j.jhydrol.2014.06.019, http://linkinghub.elsevier.com/retrieve/pii/S002216941400479X, 2014.

Rawls, W. J., Brakensiek, D. L., and Miller, N.: Green-ampt Infiltration Parameters from Soils Data, Journal of Hydraulic Engineering, 109, 62–70, https://doi.org/10.1061/(ASCE)0733-9429(1983)109:1(62), http://ascelibrary.org/doi/10.1061/%28ASCE%290733-9429%281983%29109%3A1%2862%29, 1983.

Rossman, L. A.: Storm Water Management Model Reference Manual. Volume I: hydrology (Revised), Tech. rep., U.S. Environmental Protection Agency, Cincinnati, 2016.

Rujner, H., Leonhardt, G., Marsalek, J., Perttu, A.-M., and Viklander, M.: The effects of initial soil moisture conditions on swale flow hydrographs, Hydrological Processes, 32, 644–654, https://doi.org/10.1002/hyp.11446, http://doi.wiley.com/10.1002/hyp.11446, 2018.

Schütze, M., Willems, P., and Vaes, G.: Integrated Simulation of Urban Wastewater Systems - How Many Rainfall Data Do We Need?, in: Global Solutions for Urban Drainage, pp. 1–11, American Society of Civil Engineers, Lloyd Center Doubletree Hotel, Portland, Oregon, United States, https://doi.org/10.1061/40644(2002)244, http://ascelibrary.org/doi/abs/10.1061/40644%282002%29244, 2002.

Sun, N., Hall, M., Hong, B., and Zhang, L.: Impact of SWMM Catchment Discretization: Case Study in Syracuse, New York, Journal of Hydrologic Engineering, 19, 223–234, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000777, http://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000777, 2014.

Teledyne ISCO: 2150 Area Velocity Flow Module and Sensor: Installation and Operation Guide, 2010.

Tscheikner-Gratl, F., Zeisl, P., Kinzel, C., Leimgruber, J., Ertl, T., Rauch, W., and Kleidorfer, M.: Lost in calibration: why people still do not calibrate their models, and why they still should – a case study from urban drainage modelling, Water Science and Technology, 74, 2337–2348, https://doi.org/10.2166/wst.2016.395, https://iwaponline.com/wst/article/74/10/2337/19429/Lost-in-calibration-why-people-still-do-not, 2016.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson, B. A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resources Research, 44, https://doi.org/10.1029/2007WR006720, http://doi.wiley.com/10.1029/2007WR006720, 2008.

Warsta, L., Niemi, T. J., Taka, M., Krebs, G., Haahti, K., Koivusalo, H., and Kokkonen, T.: Development and application of an automated subcatchment generator for SWMM using open data, Urban Water Journal, 14, 954–963, https://doi.org/10.1080/1573062X.2017.1325496, https://www.tandfonline.com/doi/full/10.1080/1573062X.2017.1325496, 2017.

Table A1. Characteristics of all rainfall events used in the validation phase.

Event #	Precipitation sum in preceding 72 hr	Precipitation sum (P_sum)	Precipitation duration (D_prec)	Average precipitation intensity (PI_mean)	Highest 30-minute average precipitation intensity (PI_30m)	Runoff volume (QV)	Percentage runoff (QV_ppP)	Peak flow rate (Q_max)	Highest 60-minute average flow rate (Q_60m)	Runoff from green areas [a]	Of which originating from imperv. areas [b]	Originating from green areas [c]	Average percentage runoff from green areas [d]
	mm	mm	hr	mm hr^{-1}	mm hr^{-1}	mm	%	L s^{-1}	L s^{-1}	mm	mm	mm	%
745	0.01	10.8	26.3	0.41	3.1	1.39	12.9	10.1	5.81	0.09	0.03	0.07	0.6
748	0.58	3.24	11.3	0.29	2.29	0.36	11.2	28.6	6.88				
757	0.33	2.02	2.57	0.79	3.38	0.13	6.34	7.28	2.52				
761	1.06	28.2	61.00	0.46	5.78	4.07	14.4	29.9	21.9	0.69	0.19	0.49	1.7
767	0.08	2.51	5.77	0.44	1.5	0.3	11.8	4.6	3.24				
769	0.22	2.42	2.75	0.88	2.81	0.31	12.8	16.1	6.00	0.02	0.01	0.01	0.6
770	2.64	6.34	7.52	0.84	8.15	0.92	14.5	45.2	16.8	0.16	0.05	0.11	1.8
771	8.98	3.95	4.97	0.79	4.37	0.83	21.0	30.3	15.8	0.36	0.10	0.26	6.5
772	12.7	17.8	20.3	0.88	5.84	3.57	20.1	35.7	26.7	1.44	0.41	1.03	5.8
773	21.7	8.78	8.77	1.00	3.35	1.89	21.6	17.5	11.3	0.84	0.24	0.60	6.8
775	26.8	5.10	14.2	0.36	3.25	1.35	26.4	32.4	10.7	0.74	0.21	0.53	10.3
781	0.30	6.34	11.1	0.57	2.43	0.88	13.9	23.4	6.06	0.12	0.03	0.09	1.4
791	0.91	9.48	13.7	0.69	11.1	0.72	7.59	53.3	13.5				
793	0.01	4.97	7.08	0.70	1.86	0.32	6.37	5.60	2.70				
795	3.43	9.72	21.4	0.45	3.27	0.88	9.05	15.2	7.53				
798	9.83	2.05	5.72	0.36	1.64	0.15	7.41	4.58	2.44				
799	2.13	11.4	15.9	0.72	2.55	1.20	10.6	11.1	6.24				
820	0.26	10.9	14.6	0.74	2.44	1.19	11.0	12.3	8.76				
822	11.2	20.3	17.4	1.17	6.24	3.41	16.8	51.3	28.6	0.97	0.28	0.70	3.4

[a] Calculated assuming 100% runoff from impervious areas: a = QV - 0.12 P_sum, where 0.12 is the percentage of directly connected impervious area. (Some of this runoff originated from impervious areas that drained to green areas).

[b] Calculated as b = a (25 / (25+63)), where 25 and 63 are the percentages of indirectly connected impervious surfaces and green surfaces respectively.

[c] Calculated as c = a - b

[d] Calculated as d = c / P_sum