

Reviewer's comment: The authors have provided a thorough response to the first round review comments, and at parts revised the manuscript accordingly. Leaving out some of the content present in the original manuscript has improved the clarity of the revised manuscript. In my view there would still be room for a more concise presentation to highlight the main findings of the research.

Documentation of the two-stage calibration (impermeable vs. permeable parameters) deserves to be published to a wider audience, currently the results are in my understanding only available in the licentiate thesis of the first author and in a brief form in a conference extended abstract.

We would like to thank the reviewer for once again taking the time to review the paper and provide feedback. We accept the reviewer's point about the conciseness of the article and have responded by making the following adjustments:

- The title has been changed to explicitly mention the two-stage approach to calibration: "Event selection and two-stage approach for calibration of green urban drainage models"
- The perturbation of flow data in the calibration phase has been removed entirely, except for a short mention when discussing the rainfall multipliers (p 13 | 7-8).
- The question of model resolution per se has been made less prominent, especially in the introduction and the conclusions. However, we still think it is interesting to show some results from this exercise, since the two-stage calibration shows clear benefits in the validation phase of the low-resolution model.
- The subsection on "other sources of uncertainty" has been removed from the materials and methods section, since it is now mostly superfluous. The description of the low-resolution model has been included in Sect. 2.2 'Runoff model and calibration approach'.

Reviewer's comment: The authors have retained the analysis on +/- 40% perturbed discharge data and the level of model discretization. In my view, as already expressed in my first review, the results related to these analyses do not have much novelty value and are partly trivial. The authors argue the novelty by claiming that earlier studies have predominantly addressed impervious catchments instead of 'greener' urban catchments. I am not totally convinced this is the case. In page 2 lines 20-21 the authors refer to three earlier studies with regard to spatial model resolution assessments, and state below that 'these investigations used predominantly impervious catchments'. Strictly speaking, this is true, but I find it a bit misleading as out of six catchments addressed in the referenced studies two represent the same level of imperviousness as the study area in the current manuscript. Catchment 3 of Krebs et al., 2014, imperviousness 19 % (p. 483, Section 2 Study site and data) and Suceava catchment of Petrucci and Bonhomme, 2014, imperviousness ~20% (p. 1005, Figure 6b).

(Please note that the lines in question have been removed from the text in the revised version.) The reviewer is right in their clarification of the characteristics of the catchments used in previous studies.

Reviewer's comment: Also, the authors point out that the variability of the calibrated rainfall multiplier values is larger for low-resolution than for high resolution models (p. 17, lines 1-2) and attribute this to 'rainfall multipliers appearing to behave in a more physical way' (p. 17, line 9). This sounds like an obvious result as the low resolution models, unlike the high resolution models, have another parameter directly affecting the volume of runoff ('Percentage runoff routed from impervious to pervious', Table 3, p. 10). In page 14 lines 6-8 the authors state themselves that while

in Dotto et al. (2014) the mismatch between rainfall and discharge volumes was controlled by the calibrated value of percentage imperviousness in the current paper the same impact is obtained by manipulating rainfall multiplier values. Now in low-resolution models of the current manuscript, in essence, both of these controls are present (rainfall multipliers and 'Percentage runoff routed from impervious to pervious'), which leads to a larger variability in any of these two controls individually, which is not surprising.

We accept the reviewer's point about this issue. In response to other comments, some of the results relating to the LR model have been removed from the manuscript, including this part.

Reviewer's comment: I would like to see this manuscript published but I feel a bit frustrated to see the most valuable part of the research results being mixed with the rather inconclusive results related to other sources of uncertainty. The latter are often trivial or confirm earlier findings of other research. In my mind, the manuscript would certainly benefit from a tighter framing and focus. In its current state most of the Conclusions address the results related to performance differences between one-stage and two-stage calibrations, or alternatively the relatively obvious results about the relationships between rainfall multipliers and the scaling coefficient of flow series (or variation of rainfall multipliers between high- and low-resolution models). Hardly any conclusions are drawn on the selection of calibration events (P\_sum, PI\_mean, PI\_30m etc.), which according to the title of the manuscript should be in the focus. In my view this should be fixed by changing the title to emphasize the one-stage / two-stage calibrations and revising the contents accordingly. Should the authors wish to keep the original title, then the main results and conclusions should be aligned with it.

As explained above we have removed the parts of the manuscript focusing on checking/confirming earlier findings and focused more on the different calibration strategies and the two-stage approaches. The two-stage approach is now mentioned in the new title of the paper and covered more extensively in the Introduction. The conclusions have been revised to focus on various calibration strategies (although it is difficult to say which CS is the best, some CSs can be said to be worse than the others) and the effects of the two-stage calibrations.

Specific (technical) comments

Reviewer's comment: p. 6, line 4: Correct the reference Fuentes-Andino...

This has been corrected.

Reviewer's comment: p.7, line 18: "...values of certain event characteristic...", vague language, replace with e.g. 'given event characteristic'

This phrasing has been improved: "The single-stage CSs used the six events with the highest values for a given event characteristic..."

Reviewer's comment: p. 8, line 4: Not totally clear what 'Doing this' refers to. Consider rephrasing the sentence.

Rephrased to (now p8, line 25-26): “Adjusting which calibration stage the events are available for in the calibration procedure (in a manner that is consistent for all events) would require ...”

Reviewer’s comment: p. 11, line 23: Word ‘Figure’ missing from Figure 2.

This has been fixed.

Reviewer’s comment: p. 17, line 18: “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” Consider rephrasing, hard to follow a sentence.

This sentence has been rephrased (p15 line 30-31): “For the single-stage CSs the low-resolution model predicted up to five fewer events satisfactorily than the high resolution model”

Reviewer’s comment: p. 19, lines 10-11: “inter-CS variation for the same events varies from 0.15 to 1.25.” The difference of 1.25 is hard to see from Figure 6, as the y-scale is cut at zero.

Although we accept the reviewers point, we still prefer to cut the y-axis in the figure. The reason for this is that it would make differences among the positive NS values more difficult to see, and it is (in our view) not so interesting to show the variation among different CSs, when they all have bad performance.

Reviewer’s comment: p. 21, line 7: “more less consistent than..” should probably read ‘less consistent result’

In response to other comments by the reviewer, this section of the text has been removed entirely.

Reviewer’s comment: p. 22, Table 8: not obvious what all column titles mean, what is ‘Clip mean NSE’?

Footnotes have been added to table 6 and 7 (numbers in new version of manuscript) to explain the column headers.

Finally, we realized while making the revisions that “calibration strategy” and “calibration scenario” were used interchangeably. We have now changed this to “calibration strategy” throughout.

# ~~Selection of calibration events~~ Event selection and two-stage approach for modelling calibration of green urban drainage models

Ico Broekhuizen<sup>1</sup>, Günther Leonhardt<sup>1</sup>, Jiri Marsalek<sup>1</sup>, and Maria Viklander<sup>1</sup>

<sup>1</sup>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 larger dataset in different ways. In this study, 14 single- and two-stage strategies for selecting the calibration events were tested in calibration of a high- and low-resolution SWMM model of a predominantly green urban area. The ~~event selection was considered in relation to runoff contributions from green pervious areas and such sources of uncertainty such as rainfall/runoff measurement uncertainties and catchment discretization~~ two-stage strategies used events with runoff only from impervious areas to calibrate the associated parameters, prior to using larger events to calibrate the parameters relating to green areas. Even though all 14 strategies resulted in successful model calibration (Nash-Sutcliffe Efficiency (NSE) > 0.5), the difference between the best and worst strategies reached 0.2 in ~~Nash-Sutcliffe Efficiency (NSE)~~ NSE and the calibrated parameter values notably varied. ~~Most, but not all, calibration strategies were robust to perturbations in calibration data and the use of a 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 reproduced more validation events poorly (NSE < 0)~~ performed better reproduced more validation events poorly (NSE < 0) than the single-stage strategies ~~when: (1) perturbing flow data in the calibration events by +40%; and (2) using a coarser catchment discretization, especially,~~ when: (1) perturbing flow data in the calibration events by +40%; and (2) using a coarser catchment discretization, especially, ~~but they also reproduced more events well (NSE > 0.5), and performed better than the single-stage strategies~~ but they also reproduced more events well (NSE > 0.5), and performed better than the single-stage strategies in terms of total flow-runoff volume and peak flow rates. ~~The two calibration strategies that performed the best in the validation phase were two-stage strategies. The findings in this paper show that, particularly when using a low spatial model resolution. The results show that,~~ various strategies for selecting calibration events may lead in some cases to different results in the validation phase, and ~~that~~ calibrating impervious and green area parameters in two separate steps in two-stage strategies may 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 individually. 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 (1) the five longest duration events and (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: (1) which randomly selected events were used, and (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, 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 (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~~ Thirdly, the possibility of using different (sets of) events to calibrate different (subsets of) parameters has not been investigated. One particular approach that might be useful in urban catchments is that small rainfall events will generate runoff only from impervious areas in the catchment, and could thus be used to calibrate only parameters concerning those areas, and events with more runoff where green areas also contribute could then be used to calibrate parameters concerning green areas. This two-stage calibration has not been investigated for urban drainage models, although split-stage calibration where different parameters affect different points or properties of the hydrograph have been investigated ~~before, for natural catchments (see 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 Bor~~

map catchment larger legend.pdf



**Figure 1.** Map of the studied catchment showing elements of the high-resolution rainfall-runoff model (see Sect. 2.2) 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.

~~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. Fenicia et al. (2007); Gelleszun et al. (2017)).~~

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 selecting calibration events and assessing the effects of such selections on the performance of a calibrated hydrodynamic model of a predominantly green urban catchment. ~~Part of~~ Included in 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 studies on a greener (less impervious) catchment, and (2) sensitivity of the earlier findings to the calibration data used~~ where parameters related to impervious and green areas are calibrated in two separate steps using different sets of events.

## 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). The catchment area comprises 63% of green areas, 12% of impervious areas connected directly to the storm sewer system, and 25% of impervious areas draining 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.

- 5 Flow rates in the storm sewer draining the catchment were 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  $\pm 19.0$  mm for the water depth measurements (the test range was 10-150 mm) and  $\pm 0.0985$  m/s for the velocity measurement (test range 0.1-0.6 m/s).
- 10 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.

15

The available precipitation record was divided into rainfall events with 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).
- 20 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  $> 2 \text{ L s}^{-1}$ , 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.

25

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 strategies \(see Sect. 2.3\)](#)) were tested (validated) against the same dataset and no calibration ~~scenarios~~ [strategies](#) 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.

30 Table 1 contains an overview of all events that were used in at least one calibration ~~scenario~~ [strategy](#) as well as an initial estimate of the runoff from green areas.

**Table 1.** Characteristics of all rainfall events used in one or more calibration [scenarios](#)[strategies](#).

Event #	Precipitation sum in preceding 72 hr	Precipitation sum (P_sum)	Precipitation duration (ID_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 <sup>-1</sup>	mm hr <sup>-1</sup>	mm	%	L s <sup>-1</sup>	L s <sup>-1</sup>	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

<sup>a</sup> 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).

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

<sup>c</sup> Calculated as  $c = a - b$

<sup>d</sup> Calculated as  $d = c / P\_sum$



## 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 that allows to route runoff from one sub-catchment to another. This routing feature was needed since it allows for a high-resolution (HR) model setup in which each subcatchment (146 were used in total) features a single land cover. The ~~high-resolution~~ high-resolution input data needed for this approach was available in the form of GIS data, aerial photographs, and observations from site visits. The advantage of these single 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 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). To test whether the different calibration strategies showed different sensitivity to the model discretization, a low-resolution model (LR) setup was also used. Here 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 being the same.

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. 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

**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 <sup>a</sup>	0.5-5	(Krebs et al., 2016; Rossman, 2016)
		Impervious surfaces (IMP)	0.005 - 0.015	
Manning's number [-]	n	Grass areas (GR)	0.1 - 0.5	(Krebs et al., 2016; Rossman, 2016)
		Swales (SW)	0.1 - 0.5	
		Pipes	0.010 - 0.015	
Depression storage [mm]	s	Impervious surfaces (IMP)	0 - 2.5	(Rujner et al., 2018) <sup>d</sup>
		Grass areas (GR) <sup>b</sup>	0 - 20	
		Swales (SW) <sup>c</sup>	0 - 150	
Saturated hydraulic conductivity [mm hr <sup>-1</sup> ]	ksat	Grass areas (GR) <sup>e</sup>	1 - 200	(Rawls et al., 1983)
Initial moisture deficit [-]	imd	Grass areas (GR) <sup>e</sup>	0.10 - 0.35	

<sup>a</sup> 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.

<sup>b</sup> Includes vegetation and trees as well.

<sup>c</sup> 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.

<sup>d</sup> Field experiments on similar swales in the same city.

<sup>e</sup> Used for both grass areas and swales.

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

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

<sup>a</sup> 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.

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.

## 5 2.3 Event selection

This paper investigates single- and two-stage calibration ~~scenarios-strategies~~ (CS), with each CS using six rainfall events. The single-stage CSs used the six events with the highest values ~~of a certain~~ for a given event characteristic, and calibrated all parameters simultaneously. Two-stage calibration ~~scenarios-strategies~~ 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.

10 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

15 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.

5 Characteristics related to the rainfall, flow depths and flow rates were calculated for each event. For the single-stage calibration [scenariosstrategies](#), the six highest ranking events for each characteristic were selected. For the two-stage calibration [scenariosstrategies](#), 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  
10 the percentage runoff can change from <12% to >12% and vice versa. ~~Doing this consistently for all events~~ [Adjusting which calibration stage the events are available for](#) in the calibration procedure ([in a manner that is consistent for all events](#)) 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 [scenariostrategy](#), (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  
15 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~~ [strategies](#) (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  
20 characteristics and avoided multiple CSs with the exact same set-up of events. The names of the CSs ([see Table 4](#)) 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),  
25 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.

30 The calibration [scenario-strategy](#) N\_T6 consists of the six events that were selected most often in other calibration [scenariosstrategies](#) 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) 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, earlier findings regarding the other sources of uncertainties were based on predominantly impervious catchments and they should not be 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, 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 Bareo et al. (Bareo 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 is described in the remainder of this section.

*Rainfall input uncertainty.* 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 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 Paekman (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 study outflow measurement location. 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) also tested the effect of random errors; 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,

35 involves the presence of random errors in the calibration data sets used. The flow data from the validation period was not adjusted.

*Conceptualisation / model discretization.* 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 issue, other than using different models in parallel, which was  
5 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 has been investigated by several authors. Tschekner-Gratl et al. (2016) found that a lumped model was not able to reproduce the shapes of storm runoff hydrographs  
10 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  
15 them for a predominantly green urban catchment, two levels of discretization were compared: (1) the basic model set-up (the high-resolution model described in Sect. 2.2), and (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  
20 (including rainfall multipliers) being the same.

Calibration parameters and their ranges for the low-resolution model. Parameter Abbr. Groups Range Reference Subcatchment  
width mwidth 5 individual subcatchments 20–200 Physical dimensions of subcatchments n Impervious surfaces (IMP) 0.005  
–0.015 Pervious surfaces (GR) 0.1–0.5 Pipes 0.010–0.015 s Impervious surfaces (IMP) 0–2.5 Pervious surfaces (GR) 0–20  
25 Percentage runoff routed from impervious to pervious (%) See footnote <sup>a</sup> 1–99 ksat Grass areas (GR) 1–200 imd Grass areas  
(GR) 0.10–0.35

## 2.4 Objective functions

The objective function used for the calibrations was the Nash-Sutcliffe model efficiency:

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

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  
30 -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. 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:

$$\text{PFR} = \frac{\max S_i}{\max O_i} \quad (2)$$

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:

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

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<sup>-1</sup> in calculating these metrics.

### 3 Results and discussion

#### 3.1 Calibration performance

##### 3.1.1 Baseline calibration

~~The baseline calibration (i.e. using the high resolution model without flow data perturbations) was successful~~ The high-resolution (HR) model was successfully calibrated for all calibration ~~scenarios~~ strategies, 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 2right panel in Figure 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 errors. For the two-stage calibration ~~scenarios~~ strategies, 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

**Table 4.** Calibration results. Bold font indicates the best value in each column.

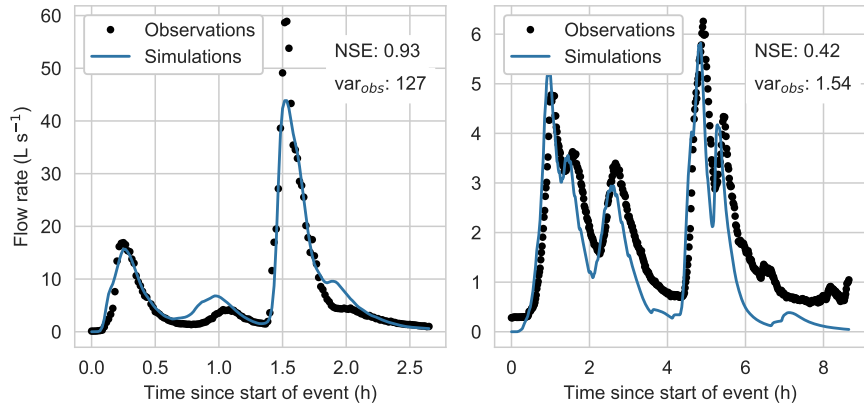
	High-resolution model			Low-resolution		
	<u>NSE</u>	<u>VE</u>	<u>PFR</u>	<u>NSE</u>	<u>VE</u>	<u>PFR</u>
N_T6	0.80	-0.07	0.93	<del>0.77</del> <del>0.76</del> 0.84	0.03	0.85 <del>0.78</del>
T6_P_sum	0.75	-0.11	0.96	<del>0.65</del> <del>0.65</del> 0.75	-0.07	0.90 <del>0.68</del>
T6_PI_mean	0.77	-0.04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

T32S\_QV\_ppP for the reasons explained above. The NSE values for the individual calibration events in the different calibration ~~scenarios is similar to that~~ strategies are similar to those reported by Krebs et al. (2013). ~~Overall, the two scenarios based on peak flow performed~~ (2013). Using the HR model, there were four event characteristics (P\_sum, PI\_mean, QV\_ppP, D\_prec) for which the two-stage calibration performed better (up to 0.08 NSE) than the single-stage calibration while for Q\_max the single-stage calibration performed better (0.03 NSE). However, when using the low-resolution (LR) model, three event characteristics (P\_sum, Q\_max, Q\_60m) had better performance with the single-stage than with the two-stage approach. Overall, N\_T6, T6\_Q\_max and T32S\_Q\_max performed best (being the only CSs with ~~mean~~-NSE > 0.8 in both the HR and LR models) while the two scenarios based on percentage runoff performed worst (~~only CSs with mean NSE < 0.7~~).

10 Considering the errors in total runoff volume, the two-stage CSs performed better for the HR model. However, for the LR model (where runoff volumes were higher in general), the single stage calibrations had smaller volume errors. These findings are similar to earlier findings by Tscheikner-Gratl et al. (2016) and Krebs et al. (2016). Although the CSs based on peak flow rates (Q\_max) performed well in terms of NSE, they are actually among the worst performers in terms of peak flow ratio in both the HR and LR model. This may be attributed to the possibility for models to obtain high NSE values despite underestimating high peak flows (see left panel in Figure 2). In general the LR model resulted in lower peak flow ratios (as also shown by Tscheikner-Gratl et al. (2016)), and this effect was stronger for the two-stage CSs.

15





example hydrographs run130.pdf

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

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/strategies](#), 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. [\(Note that with 6 CSs with 3 first-stage events each, there were 18 model runs in total.\)](#) 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/strategies](#), 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.

### 3.1.1 Low-resolution-model

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\_pree) 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.

### 3.1.1 Sensitivity to structural flow measurement errors

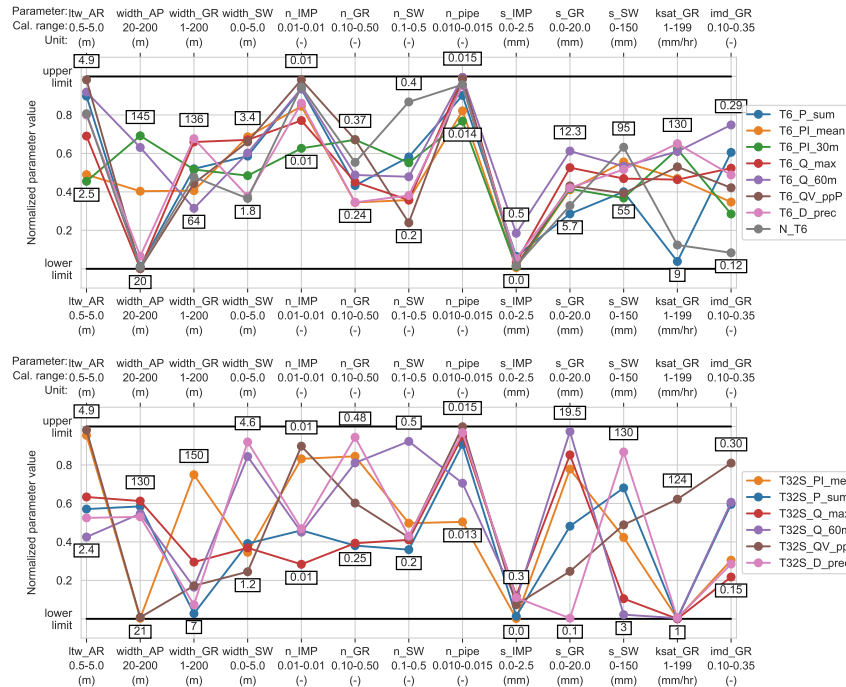
Calibration results (NSE) are shown in Table 4 for the cases of structural flow data errors of -40% and +40%. For most calibration scenarios there was a small loss in NSE, except for T6\_QV\_ppP, which failed to calibrate with an NSE of -0.1 when the flow data was reduced by 40%. Three of the events 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 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 ?? 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  $1 \text{ L s}^{-1}$  threshold. Except for the events described above, the flow errors could be compensated for in calibration, 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.

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.

## 3.2 Calibrated parameter values

### 3.2.1 Hydrologic model parameters

Figure 3 shows the calibrated parameter values (for the [baseline run HR model](#)), normalized with respect to their calibration ranges (see Table 2). There is considerable variation among the calibrated values obtained in different calibration [scenarios/strategies](#), 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 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. [The variation among the two-stage CSs was larger than that among the single-stage CSs for most parameters, which may be caused by the dataset used to estimate each parameter being](#)



param values.pdf

**Figure 3.** Normalized calibrated parameter values for [the high-resolution model using different calibration scenarios and the baseline run strategies](#). The highest and lowest values found for each parameter are indicated.

[smaller \(3 events instead of 6\). The depression storage in green areas and swales might be compensating for each other in the two-stage CSs. The depression storage for impervious areas shows little variation \(0 - 0.3 mm\) between the different CSs, with only T6\\_Q\\_60m resulting in a slightly higher value \(0.5 mm\).](#)

5 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-strategies](#) 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

10 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 matching best the observed flow rates. [For rainfall-Two arguments support that the rainfall multipliers do indeed fulfil the role of compensating for this mismatch. Firstly, for rainfall events that were included in multiple calibration](#)

~~scenarios/strategies, the calibrated multipliers from different scenarios were close to each other (see Table 5). This variation was much smaller than that, unlike for the hydrological model parameters (see Sect. 3.2.1). The average value of the rainfall multipliers across all events was 1.2.~~

5 ~~When all flow data was decreased Secondly, decreasing or increasing all flow rates 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) changed the average rainfall multipliers by -37% and +33% respectively. The average value of the rainfall multipliers across all events~~  
10 ~~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 work as intended, i.e. compensating for discrepancies between the observed and best-fitting rainfall, rather than for other aspects of catchment runoff modelling. 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 1.2, which suggests that there was some structural disagreement between the observed rainfall and flows. The close agreement between the different CSs shows that, unlike the hydrological model parameters, the rainfall multipliers are not sensitive to differences between the CSs.~~

~~In runs with the low-resolution model, 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, as shown in Figure ???. The multiplier values obtained ranged from 25% lower to 50% higher, for the same event in the same calibration scenario, compared to the baseline calibration. Three of the low-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, 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).~~

30 ~~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.~~

**Table 5.** ~~Baseline-run-calibrated~~ Calibrated rainfall multipliers (HR model) for all rainfall events that were used in at least ~~three CSs~~ one CS.

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 <sup>a</sup>
211		0.70	0.70		0.70	0.70									0.70	6.8	15.8 <sup>a</sup>
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 <sup>b</sup>
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 <sup>b</sup>
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

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

<sup>b</sup> Vice versa.

### 3.3 Validation performance

#### 3.3.1 Individual events

The ~~successful-calibrations-predicted-8-13-out-of-the-19-validation-events-satisfactorily-(NSE>0.5),-see-Table-??.~~ 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

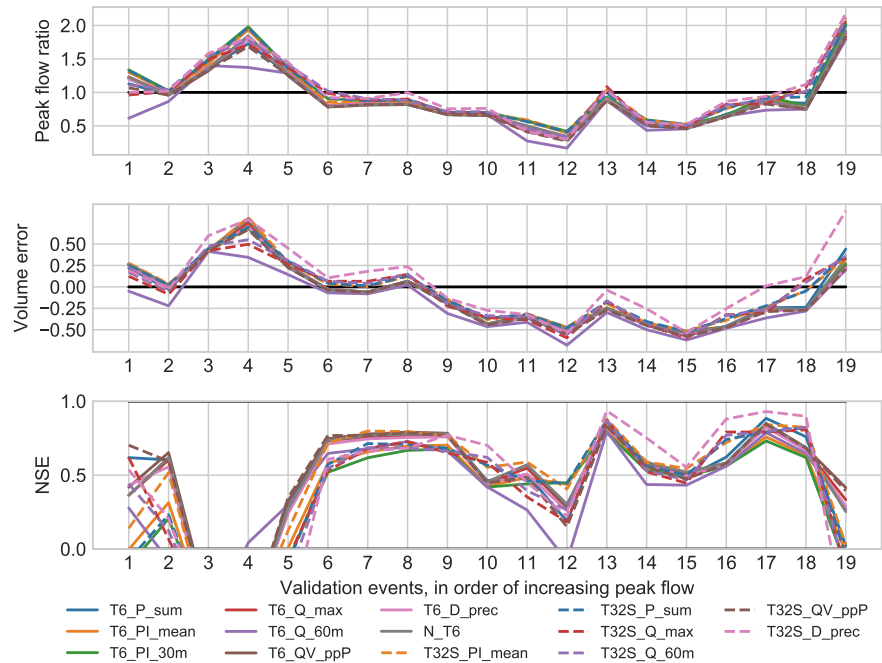
scenarios were less sensitive to perturbations of the flow data in the calibration period, i.e. they predicted more validation events satisfactorily than their single-stage counterparts. When switching from the high resolution to the low-resolution model the single-stage CSs were no longer able to predict 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 events satisfactorily. Over all four calibration runs, the two-stage calibrations were able to predict more events satisfactorily than their single-stage counterparts.

Number of validation events with NSE >0.5 out of 19 total events. Bold font indicates the best value in each column.

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	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 <sup>a</sup>	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
----------	----------------	----------------	----------------	---------	----	----	----	---	---	----	-----------	----	---	---	---	----	----------	----	---	---	---	----	-----------	---	---	---	----	------------	----	---	----	---	----	----------	---	---	---	---	----	----------	----	---	----	----	----	-----------	----	----------------	---	----	----	-------------	----	----	----	----	----	------------	----	---	----	----	----	--------------	----	----	----	----	----	------------	----	---	---	----	----	------------	----	---	----	----	----	-------------	----	----	----	----	----

validation performance for individual events is visualized in Figure 4 for peak flow ratio, volume error and NSE. The events that most often caused failure in validation were four events with peak flow rates of 10 L s<sup>-1</sup> or less (i.e. events 1-4 in the figure), and therefore, such failures may be attributed to: (1) relatively high measurement uncertainties, and (2) low variance of the observations leading to high sensitivity of the NSE to even small changes in the hydrographs differences between observed and predicted hydrographs (see section 2.4 and Figure 2). However, it should be noted that the two smallest events (both with a peak flow rate of 4.6 L s<sup>-1</sup>) were predicted with NSE >> 0.5 by some calibration scenarios T6\_P\_sum and T32S\_QV\_ppP. For the other CSs, examination of the hydrographs showed that they predict predicted well the magnitude of events, but produce wrong timing produced wrong timing, which was also reflected in the peak flow and volume errors. Another event that failed in validation for all CSs was that with the highest peak flow rate (53 L s<sup>-1</sup>, see Table A1 event 19 in Figure 4, 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<sup>-1</sup>), so it could have suffered from high spatial variation of the rainfall.

The peak flow ratios obtained for the 19 validation events using the calibrated models from the baseline high-resolution models are shown in the upper panel of Figure 4. 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 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 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 strategies, showing that selecting a limited number of validation events may also influence the results of the model evaluation. T32S\_D\_prec stood out by predicting higher runoff volumes and peaks, and therefore better performance, for the events labelled 13-18 in Figure 4. Across all CSs, two-stage versions had similar or better performance in terms of total runoff volume. Peak flow ratios were <1 was underestimated 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



error stats validation events.pdf

**Figure 4.** Error statistics for individual validation events for all calibration scenarios-strategies in the baseline-runsHR model.

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. For several events (10, 16, 18) the two-stage CSs (except T32S\_QV\_ppP) showed better performance than the single-stage CSs, but there were no events where all the single-stage CSs performed better.

### 3.3.2 Overall performance of the high-resolution model

The successful CSs predicted 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. For the single-stage CSs the low-resolution model predicted up to five fewer events satisfactorily than the high-resolution model, while from the two-stage CSs only T32S\_D\_prec satisfactorily predicted fewer events, and T32S\_P\_sum, T32S\_Q\_max, and T32S\_QV\_ppP actually predicted a higher number of events satisfactorily.

To assess the overall performance of different calibration ~~scenarios-strategies~~ for the validation period, several ways of combining the individual events were considered (see Table ??6). 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 ( $<10 \text{ L s}^{-1}$  peak) events ~~avoids-would avoid~~ this issue, but ~~does-would~~ not reward calibration ~~scenarios-strategies~~ 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 mean NSE ranging from 0.29 (T6\_PI\_30m) 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 (T6\_QV\_ppP). The two-stage CSs had worse performance than the single-stage CSs (except for PI\_mean).~~ 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 with NSE ranging from 0.57 (T6\_PI\_30m) - 0.70 )-(T32S\_PI\_mean and T32S\_D\_prec). This last metric also showed better performance for two-stage CSs than their single-stage counterparts (except for QV\_ppP), i.e. the opposite of what was found for the mean NSE.~~ 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 ??). It depended on the chosen criterion which CSs performed best, but T6\_PI\_mean, T6\_PI\_30m and T6\_Q\_60m were always near the bottom in the NSE-based metrics and would therefore not be recommended. Of the two-stage CSs, T32S\_PI\_mean showed the best performance in the NSE based metrics.~~

The considerations in the previous paragraph concern the NSE and are not necessarily applicable to other statistics in the same way. 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. The volume errors were similar for all high-resolution single-stage calibrated models and showed a general tendency to underestimate flow volumes by 25%. For the two-stage calibrated models volume errors were smaller with underestimation of around 15% (except for T32S\_QV\_ppP).

### 3.3.3 Overall performance of the low-resolution model

~~The effect of the low-resolution model depended on the calibration scenario considered, see Table 7. Some scenarios scored better in terms of NSE (gains of up to 0.17 and 3 events predicted with  $\text{NSE} > 0.5$ ), while others lost performance by the same metrics (up to 0.24 and 5 events). This is a more 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, and T32S\_D\_prec showed a volume error of only -2%. The average peak flow ratio over all events indicated better performance for the two-stage scenarios predicted more events satisfactorily with the low-resolution model than with the high-resolution model.~~



**Table 6.** Summarized validation performance ~~for all (over 19 validation events)~~ for the high-resolution model. Bold font indicates the best value in each column.

	<del>Mean NSE</del> <u>Mean NSE</u>	<del>Clip-mean NSE</del> <u>Clip mean NSE</u> <sup>a</sup>	<del>Median Joint NSE</del> <sup>b</sup>	<del>Joint NSE</del> # neg <u>NSE</u> <sup>c</sup>	<del># good NSE</del>
N_T6	0.33	0.45	<del>0.58</del> 0.65	<del>2</del> 2	12
T6_P_sum	0.39	0.45	<del>0.60</del> 0.66	<del>2</del> 2	12
T6_PI_mean	0.18	0.33	<del>0.51</del> 0.59	4	10
T6_PI_30m	0.13	0.29	<del>0.49</del> 0.57	<del>5</del> 5	9
T6_Q_max	0.34	0.44	<del>0.58</del> 0.65	<del>2</del> 2	12
T6_Q_60m	0.37	0.37	<del>0.43</del> 0.60	3	8
T6_QV_ppP	0.36	<del>0.47</del> <b>0.47</b>	<del>0.58</del> 0.67	<del>2</del> 2	12
T6_D_prec	0.34	0.43	<del>0.56</del> 0.64	<del>2</del> 2	11
T32S_P_sum	0.19	0.34	<del>0.56</del> 0.68	<del>5</del> 5	10
T32S_PI_mean	0.26	0.44	<del>0.59</del> <b>0.70</b>	<del>0.70</del> 2	<del>13</del> <b>13</b>
T32S_Q_max	0.31	0.34	<del>0.53</del> 0.67	4	11
T32S_Q_60m	0.26	0.33	<del>0.53</del> 0.68	4	10
T32S_QV_ppP	<del>0.42</del> <b>0.42</b>	0.46	<del>0.58</del> 0.65	<del>2</del> 2	11
T32S_D_prec	0.22	0.34	<del>0.61</del> <b>0.70</b>	<del>0.70</del> 4	12

<sup>a</sup> calculated after setting individual event values <-1 to -1.

<sup>b</sup> calculated after merging all event time series into a single series.

<sup>c</sup> Number of events with NSE < 0

<sup>d</sup> Number of events with NSE > 0.5

20 For the CSs than for the single-stage calibration scenarios, the volume errors in the LR were twelve to nineteen percent points higher. The two-stage scenarios showed both worsened performance (T32S\_P\_sum, T32S\_PI\_mean) and improved performance (~~T32S~~CSs. The CSs based on rainfall intensity (PI) showed the best performance in terms of peak flows. T6\_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 60m had the worst performance for total volume and peak flow (despite being calibrated to events that score highly on both characteristics), and would therefore not be recommended.

25 Most of the LR, two-stage calibrations had higher mean NSE than their single-stage counterparts (except for Q\_max and Q\_60m, see Table 7), and visual comparison of the hydrographs showed that for most events the HR model performed better. However, for the last three CSs mentioned, the low-resolution performed better compared to the other CSs. For the two-stage calibrations performed significantly better than their single-stage counterparts in terms of volume error and peak flow (see Table 7), and the two-stage CSs that were based on observed flow rates (T32S\_Q\_60m\_max 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. 60m) outperformed the HR model in the visual comparison of hydrographs.

30

**Table 7.** Summarized validation performance (over 19 events) for the low-resolution models. Bold font indicates the best value in each column.

The columns marked with \* are not discussed in the text, but shown here for completeness and comparability with Table 6.

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

<sup>a</sup> calculated after setting individual event values <-1 to -1.

<sup>b</sup> calculated after merging all event time series into a single series.

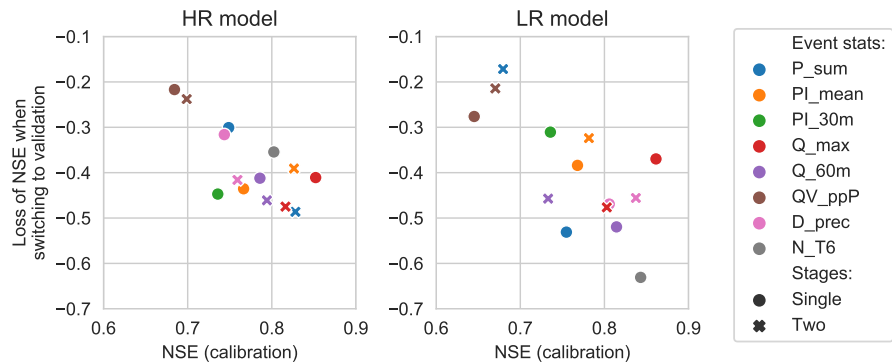
<sup>c</sup> Number of events with NSE < 0

<sup>d</sup> Number of events with NSE > 0.5

### 3.3.3 Sensitivity to structural flow errors

The introduction of structural flow measurement errors 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))

### 3.4 Degradation of performance from calibration to validation



v20190710 loss val performance.pdf

**Figure 5.** Loss of performance (NSE) when switching from calibration to validation.

In calibration, the NSE for the different calibration strategies ranged (for the HR model) from 0.68 to 0.85, while in validation this was lowered to 0.29 to 0.47 (NSE values <-1 were set to -1 prior to taking the mean, see Sect. 3.3.1).

### 10 3.5 Degradation of performance from calibration to validation

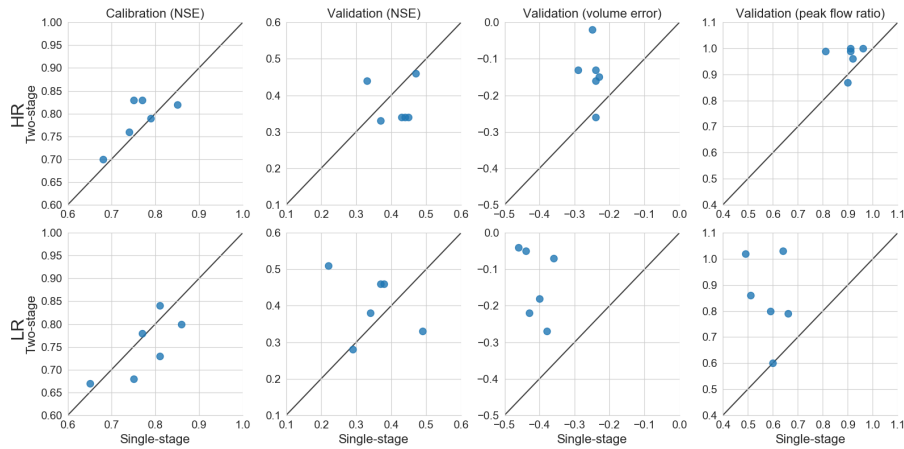
In calibration, the NSE for the different calibration scenarios ranged from 0.68 to 0.85, while in validation it ranged from 0.29 to 0.47.2). For the LR model the variation between different CSs was slightly larger, ranging from 0.65 to 0.86 in calibration and from 0.21 to 0.51 in validation. The CSs that did better in calibration lost more performance (measured by NSE) when switching to the validation phase (see Figure 5). The range of performance loss for the different calibration scenarios was larger

5 for the low-resolution model than for the high-resolution model5). In particular, the CSs based on percentage runoff (QV\_ppP) had the worst calibration performance (in both HR and LR models), but lost the least when switching to the validation phase. For the high-resolution-high-resolution model all but one of the two-stage calibrations lost more performance when switching to the validation phase than their single-stage counterparts, whereas. By contrast, for the low-resolution model all but one of the two stage calibrations had a smaller performance loss. The findings in this section demonstrate that good calibration performance

10 is not necessarily indicative of good validation performance and vice versa, and therefore, whenever feasible, validation should be performed. 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. (20142014), Krebs et al. (20142014)), 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.

### 15 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 can be compared directly (see Figure 6). For the high-resolution model, calibration performance of the two-stage CSs was somewhat better than for the single-stage CSs. By contrast, in the validation phase the NSE was better for the single-stage



**Figure 6.** Comparison of single-stage and two-stage calibration strategies.

CSs. However, the volume error and peak flow ratio were better for the two-stage 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 most often had higher NSE. In addition, the two-stage calibrations resulted in much better performance in terms of volume error and peak flows than their single-stage counterparts.

#### 4 Conclusions

The primary objective of this study was to compare different strategies for the selection of calibration events for a **hydrodynamic** **combined hydrologic-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 **based on observed precipitation and catchment outflow** by which calibration events can be selected from a larger group of candidate events. **Calibration strategies were tested with** **The single-stage calibrations used six events to calibrate all model parameters simultaneously, while the two-stage calibrations used three events (with less runoff than the percentage of directly connected impervious area in the catchment) to calibrate impervious area parameters, followed by using three events (with more runoff) to calibrate green-area parameters. The results of different calibration strategies for high and low spatial resolution models and on data sets with structural flow data errors. The conclusions drawn below are strictly valid for the specific data and catchment characteristics used in this study** are summarized below. **It should be noted that the precise performance values presented in this paper may vary for different catchments and datasets.**

**In the baseline run (high resolution model, no structural flow data errors), all calibration scenarios** **For the high-resolution model, all calibration strategies** 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~~ generally performed better in the calibration phase than their single-stage counterparts in terms  
20 of NSE and runoff volume error. The two-stage calibrations also were faster since they reduced the dimensionality (number  
of simultaneously calibrated parameters) of the calibration problem ~~and the number of model runs at each iteration.~~ The CSs  
N\_T6, T6\_Q\_max and T32S\_Q\_max performed best in calibration, while CSs based on percentage runoff performed worst.  
Although the obtained values of the SWMM model parameters varied between the different CSs (and this variation was greater  
for two-stage 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 calibration performance metrics. However,  
5 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  
10 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. 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  
15 findings were insensitive to calibration data selection except in the case of T6\_QV\_ppP.~~

~~The calibrated scenarios were validated against~~ For the model validation phase an independent set of 19 validation events  
was used. All calibrated scenarios predicted ~~7-8~~ to 13 of these events satisfactorily (NSE >0.5). ~~A group of four events with  
peak flow rates of less than  $10 \text{ L s}^{-1}$  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  
20 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~~ Although the question of which CS performed best depended on the performance  
metric considered, it can be said that T6\_PI\_mean and T6\_PI\_30m performed poorly in NSE-based metrics, and T6\_Q\_60m  
performed poorly according to all metrics. Variation among the different CSs was larger for the LR model than for the HR  
25 model. For the HR model the two-stage CSs had more events with negative NSE, but higher NSE when the events were  
combined into a single time series. For the LR model the two-stage CSs had both more events with negative NSE and with  
NSE > 0.5, resulting in better mean NSE for the two-stage CSs. For volume error and peak flow error the two-stage CSs  
performed better. ~~Better performance in regards to flow volumes and peak flows bears more~~ , especially with the LR model,  
which bears significance for engineering design. The two-stage CSs based on flow rates (Q\_max and Q\_60m) were the only  
30 two CSs where the LR version outperformed the HR version when visually comparing the hydrographs.

~~Concerning model discretization, the low-resolution single-stage calibration scenarios showed significantly larger volume  
errors than their high-resolution counterparts, while most~~ To summarize, there was clearly variation between the different CSs

in both the calibration and the validation phase, and although it is difficult to say which CS performs best (since this depends on the performance metric used), some CSs perform poorly throughout. Although the two-stage calibration scenarios 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 model setup. Two-stage calibrations also predicted more validation events satisfactorily when the calibration flow data was perturbed. Earlier studies found that high-resolution models lost less performance when switching to CS had more problematic validation events with the HR model, they also had more satisfactorily predicted validation events. Finally, the two-stage CSs clearly performed better in terms of total runoff volume and peak flow in the validation phase (Krebs et al., 2014; Sun et al., 2014), but, in the current paper, this depended on the set of calibration data that was selected, and this effect was particularly strong for the LR model.

*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~~ We gratefully acknowledge the financial support provided 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~~ and the VINNOVA (Swedish Governmental Agency for Innovation Systems) DRIZZLE - Centre for Stormwater Management (Grant no. 2017-04390). We are also gratefully acknowledge technical expertise provided by the Stormwater&Sewers network and would particularly like to thank Helen Galfi, Ralf Rentz and Karolina Berggren for their work in setting up and maintaining the field measurements. The authors would like to thank CHI/HydroPraxis for providing a license for PCSWMM.

## 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\(466\)](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., Vezzano, 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](https://doi.org/10.1016/0022-1694(94)90057-4), <http://linkinghub.elsevier.com/retrieve/pii/S0022169494900574>, 1994.
- 10 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>,  
15 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.
- Fenicia, F., Savenije, H. H. G., Matgen, P., and Pfister, L.: A comparison of alternative multiobjective calibration strategies for hydrological modeling, *Water Resources Research*, 43, <https://doi.org/10.1029/2006WR005098>, <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006WR005098>, 2007.
- 20 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.
- 25 Gelleszun, M., Kreye, P., and Meon, G.: Representative parameter estimation for hydrological models using a lexicographic calibration strategy, *Journal of Hydrology*, 553, 722–734, <https://doi.org/10.1016/j.jhydrol.2017.08.015>, <http://www.sciencedirect.com/science/article/pii/S0022169417305413>, 2017.
- 30 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.
- 35 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.svenskvatten.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, e0145180, <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.
- 5



- 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.
- 10 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.
- 15 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.
- 25 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](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](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000382), <http://ascelibrary.org/doi/10.1061/%28ASCE%29WR.1943-5452.0000382>, 2014.
- 30 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.
- 35 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](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\(62\)](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.
- 10 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: 15 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](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](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000777), <http://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000777>, 2014.
- 20 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, 745 <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 sub-catchment 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 <sup>-1</sup>	mm hr <sup>-1</sup>	mm	%	L s <sup>-1</sup>	L s <sup>-1</sup>	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

<sup>a</sup> 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).

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

<sup>c</sup> Calculated as  $c = a - b$

<sup>d</sup> Calculated as  $d = c / P\_sum$