Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

Calibration event selection for green urban drainage modelling

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

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

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

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

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

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 event set) will influence the results of calibration (Tscheikner-Gratl et al., 2016).

Tscheikner-Gratl et al. (2016) studied such influence by calibrating water level in the outflow pipe of a catchment using ten different rain events. They found that two of them could not be reproduced in calibration and the others, while successful in calibration, could only predict up to six of the remaining events. When applying the calibrated models with design storms, they found that the calibrated models predicted different flooding volumes. In calibration of combined sewer overflow (CSO) volumes, Kleidorfer et al. (2009b) compared calibration results obtained for (i) the five longest duration events and (ii) the five highest peak flow events, finding that using the longest duration events reduced the number of measurement sites required for successful calibration. Schütze et al. (2002) demonstrated that calibration based on discrete events saved time compared to calibrating for a complete time series, but also that this introduced additional uncertainty. Mourad et al. (2005) showed that calibration of a stormwater quality model was sensitive to: (i) which randomly selected events were used, and (ii) 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, a limited number of generally available options for selecting calibration events has been considered; secondly, the modelling focused on traditional urban drainage systems where generation of runoff is dominated by impervious surfaces, but the current trend towards green urban drainage infrastructure creates the need to pay more attention to runoff processes on green areas Fletcher et al. (2013).

The primary objective of the paper that follows is to advance the knowledge of calibration processes for green urban areas by examining different strategies for calibration event selection and their effects on the performance of a calibrated hydrodynamic model of a predominantly green urban catchment. Since uncertainties in urban drainage modelling arise from other sources as well (Deletic et al., 2012), the calibration event selection is considered in relation to some of them.

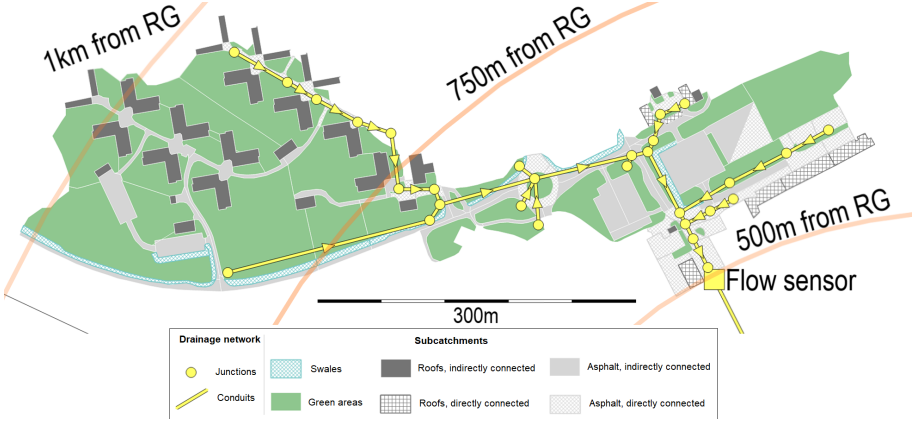
2 Materials and methods

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 draining directly to the storm sewer system, and 25% of impervious areas draining to 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

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019



1 map catchment.png

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

sensor to be a reliable instrument (Duchon, 2002; Lanza et al., 2010). Records were available for individual rain events in 2013-2015 and continuously for 2016 and 2017.

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

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

The available precipitation record was divided into rainfall events with at least six hours without precipitation between them. Events deemed suitable for use in calibration were selected using the following criteria:

1. A minimum total precipitation of 2 mm (Hernebring, 2006)
2. No or small gaps in rain and flow data , i.e. both have to be available for >90% of the event duration
3. Sufficient in-pipe water depths for the flow sensor to work reliably: >10 mm during at least 50% of the event and >25 mm at least once in the event, based on recommendations from the manufacturer (Teledyne ISCO, 2010).

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

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

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

Calibration and validation periods were separated by using the 19 observed events from 2016 for the validation period, and the 32 events from 2013-2015 and 2017 for the calibration period. In this way, all the calibration scenarios were tested (validated) against the same dataset and no calibration scenarios could benefit from including calibration events that also appeared in the validation set. The year 2016 was selected as the validation period for two reasons: it was the year with total precipitation closest to the annual mean, and the measured data records were continuous.

2.2 Runoff model and calibration approach

The US EPA Storm Water Management Model (SWMM) was selected since it is a commonly used semi-distributed urban drainage model and it allows to route runoff from one sub-catchment to another. This routing feature was needed since it allows for a high-resolution model setup in which each subcatchment (146 were used in total) features a single land cover. The high resolution input data needed for this approach was available in the form of GIS data, aerial photographs, and observations from site visits. The advantage of these single land-use subcatchments is that their parameter values maintain their physical meaning and can be calibrated (or appropriate values found in the literature) for each land use or cover. The traditional approach of using larger subcatchments with multiple land uses/covers usually necessitates calibration to estimate the values of parameters that then represent a weighted average value over multiple land uses/covers. Some spatial characteristics, such as the slope and the width of subcatchments, can also be estimated more easily for smaller, uniform subcatchments. This approach has been used successfully by e.g. Krebs et al. (2014, 2016), Petrucci and Bonhomme (2014) and Sun et al. (2014). Within SWMM the Green-Ampt infiltration method was selected since it can be calibrated with just two parameters (Rossman, 2016).

Whenever feasible, parameters for the different subcatchments were set directly from the available GIS data and site visits, i.e. the sizes and slopes of all subcatchments and sewer pipes, as well as the catchment widths of small and disconnected roofs. For other subcatchments the catchment width was calibrated together with the other model parameters. To reduce the scope of the calibration problem, parameters were grouped based on land cover, yielding a total of thirteen calibration parameters for the hydrodynamic model. Parameter values were limited based on values reported in literature (see Table 1). 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). The rainfall multipliers create a way of adjusting the rainfall volume in the calibration so that the simulated runoff volume can better match the observed runoff volume. It is, however, not possible to distinguish between deviations between rainfall at the gauge and the catchment-averaged rainfall, errors in the rainfall measurement, and errors in the runoff measurement. A more traditional approach would be to calibrate

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

Table 1. Calibration parameters and their ranges.

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

[a] In SWMM, the subcatchment width is an input, but in this group of subcatchments, the length (in the flow direction) showed more similarity among the subcatchments, so it was calibrated instead of the width.
[b] Includes vegetation and trees as well.
[c] The maximum value was intentionally set high since the swales' outlets are not always located exactly at the lowest points and the swales can be observed with larger ponds after heavy rain events.
[d] Field experiments on similar swales in the same city.
[e] Used for both grass areas and swales.

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.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

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

2.3 Event selection

This paper investigates single- and two-stage calibration scenarios (CS), with each CS using six rainfall events. The single-stage CSs used the six events with the highest values of a certain event characteristic, and calibrated all parameters simultaneously. Two-stage calibration scenarios calibrated first the parameters related to impervious areas, using a set of three rainfall events, followed by the pervious area parameters using another set of three rainfall events. Since only 12% of the total catchment surface is impervious and connected directly to storm sewers, it was assumed that the events, for which runoff volume was less than 12% of rainfall volume, produced runoff only from impervious areas. Therefore, these events were suitable for calibration of impervious area parameters in the first stage of the calibration process. Following this step, events with more than 12% runoff were assumed to also include runoff from green areas and were used to estimate pervious area parameters in the second stage of the calibration. When calibrating the green area parameters, the parameters related to impervious areas were kept fixed at their values from the first stage. This procedure splits the optimization problem into two smaller problems that have fewer parameters and shorter run times. The smaller number of parameters (reduced dimensionality) can ease the search for optimal parameter sets, while the shorter run time per iteration allows shortening the total time needed, increasing the number of iterations used, or including more events in the calibration.

Characteristics related to the rainfall, flow depths and flow rates were calculated for each event. For the single-stage calibration scenarios, the six highest ranking events for each characteristic were selected. For the two-stage calibration scenarios, the three highest ranking events with less than 12% runoff were selected for the first stage and the three highest ranking events with more than 12% runoff were selected for the second stage. To avoid making the comparison too large in scope, a limited number of calibration scenarios (eight single-stage and six two-stage) was selected for use in this study. This selection was made so that it included a range of different characteristics and avoided multiple CSs with the exact same set-up of events. The names of the CSs consist of two or three elements:

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

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

2.4 Other sources of uncertainty

Calibration data selection is not the only source of uncertainty in urban drainage modelling. Deletic et al. (2012) identify nine sources: (1) input data, (2) model parameters, (3) calibration data measurements, (4) calibration data selection, (5) calibration algorithm, (6) objective functions, (7) conceptualisation (e.g. discretization), (8) process equations and (9) numerical methods and boundaries. As described above, calibration data selection is the focus of this paper, however, it should not be viewed in isolation from the other sources listed above. Therefore, different strategies for selecting calibration events were considered in relation to the other sources of uncertainty as discussed below.

Rainfall input uncertainty. Since the rain gauge is located outside of the catchment and the maintenance of the gauge was carried out by different people, it is possible that there are structural errors in the rainfall measurements. This was investigated by examining the rainfall multipliers that were included for each event in the calibration (see Sect. 2.2).

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

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

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

Conceptualisation / model discretization. The model code (SWMM) employed in this study has been widely used for many years, with some improvements made to those parts of its conceptualisation that were deemed unsatisfactory. Therefore, it is safe to assume that the SWMM conceptualization (Rossman, 2016) is appropriate for urban drainage modelling and there was no need to consider this issue further. However, the choice of catchment discretization into the subcatchments in the model is done, somewhat subjectively, by the modellers for individual studies: therefore, two levels of discretization were compared: (i) the basic model set-up (the high-resolution model described in Sect. 2.2), and (ii) a simpler, more traditional set-up using five subcatchments. In the latter case, each subcatchment was created by aggregating multiple smaller subcatchments from

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

Table 2. Calibration parameters and their ranges for the low-resolution model.

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

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

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 2, with the total number of calibration parameters (including rainfall multipliers) being the same.

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

2.5 Objective functions

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

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

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 –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 vari-

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

ance of the observations is small (e.g. for small runoff events), it can become quite sensitive to small changes in the simulated hydrograph. To examine the impact of different objective functions, one calibration used Root Mean Square Error (RMSE):

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

RMSE has the same units as the observations (in this case L s^{-1} for the flow rate). 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:

$$\text{PFR} = \frac{\max S_i}{\max O_i} \tag{3}$$

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:

$$\text{VE} = \frac{\sum_{i=1}^{n}(S_i - O_i)}{\frac{1}{n}\sum_{i=1}^{n} S_i} \tag{4}$$

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

3 Results and discussion

3.1 Calibration performance

3.1.1 Baseline calibration

The baseline calibration (i.e. with NSE as objective function, using the high resolution model without flow data perturbations) was successful for all calibration scenarios, with average NSE for all events ranging from 0.68 to 0.85 (see Table 3). The lowest NSE corresponded to the two CSs based on the percentage runoff (T6_QV_ppP and T32S_QV_ppP). This result can be attributed to one event (see Figure 2), 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

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

Table 3. Calibration results.

	Baseline (objective function: NSE)		RMSE as obj. func.		Structural flow errors			
Criterion	NSE*	RMSE	NSE	RMSE*	NSE; flow -40% *	NSE; flow +40% *	NSE, low resolution model*	Rank[a]
N_T6	0.80	3.24	0.80	3.27	0.77	0.76	0.84	3
T6_D_prec	0.74	1.61	0.78	1.50	0.72	0.69	0.81	8
T6_P_sum	0.75	2.22	0.75	2.27	0.65	0.62	0.75	10
T6_PI_30m	0.74	4.23	0.73	4.22	0.72	0.72	0.74	13
T6_PI_mean	0.77	3.90	0.76	3.89	0.63	0.72	0.77	11
T6_Q_60m	0.79	4.12	0.78	4.12	0.77	0.77	0.81	7
T6_Q_max	0.85	3.55	0.85	3.54	0.82	0.84	0.86	1
T6_QV_ppP	0.68	2.78	0.65	2.52	-0.10	0.65	0.65	14
T32S_D_prec	0.76	1.54	0.77	1.62	0.82	0.72	0.84	2
T32S_P_sum	0.83	2.70	0.83	2.62	0.76	0.82	0.68	5
T32S_PI_mean	0.83	3.37	0.80	3.48	0.73	0.78	0.78	6
T32S_Q_60m	0.79	3.52	0.81	3.39	0.73	0.76	0.73	9
T32S_Q_max	0.82	3.68	0.82	3.40	0.78	0.83	0.80	3
T32S_QV_ppP	0.70	2.35	0.65	1.95	0.62	0.73	0.67	11

[a] CSs were ranked by each column marked with an asterisk *. The overall ranking is based on the sum of these per-column rankings.

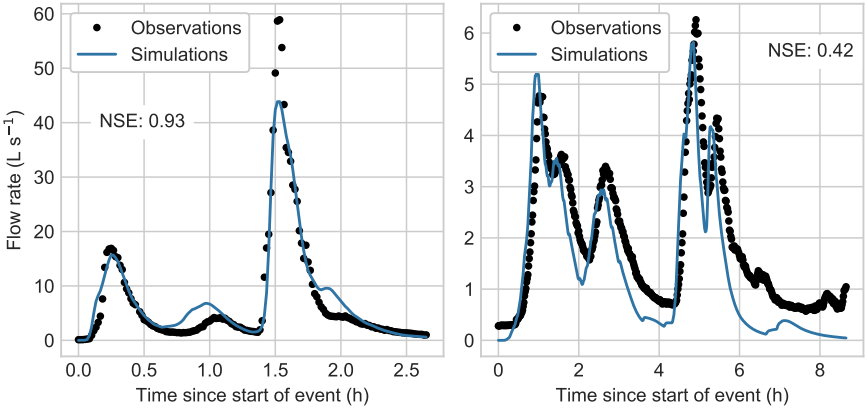
errors. For the two-stage calibration scenarios, the individual stages also produced successful calibrations (stage 1 NSE 0.70 – 0.87, stage 2 NSE 0.78-0.87), except for the second stage in T32S_QV_ppP for the reasons explained above. The NSE for the individual calibration events in the different calibration scenarios is similar to that reported by Krebs et al. (2013).

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

3.1.2 Sensitivity to objective functions

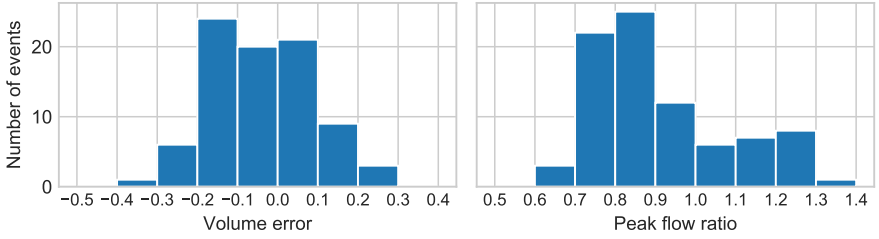
The differences between calibrations using NSE and RMSE as objective functions were small (see Table 3), with the largest differences being 0.05 (NSE) and 0.4 (RMSE) for T32S_QV_ppP. For three calibration scenarios the NSE calibration found a better RMSE than the RMSE calibration and for four CSs the RMSE calibration found a better NSE than the NSE calibration.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019



2 example hydrographs run130.pdf

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



3 VE PFR histograms.pdf

Figure 3. Histograms of peak flow ratios (left) and volume errors (right) for all individual events in all calibration scenarios.

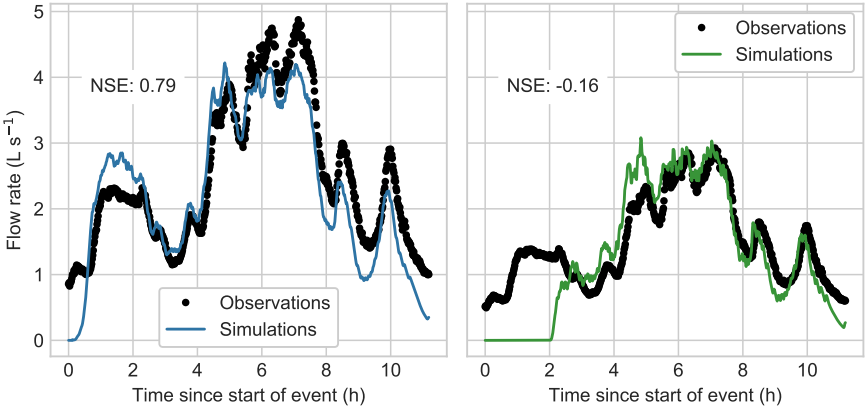
This indicates that the algorithm does in some cases find a local rather than a global optimum. However, the differences between them are small.

3.1.3 Sensitivity to model discretization

Calibration runs with a model setup consisting of five instead of 140 subcatchments showed NSE similar to that of the baseline run: the change in performance ranged from +0.08 (T32S_D_prec) to -0.06 (T32S_Q_60m), with only T32S_P_sum showing a larger loss of 0.15. The peak flows predicted by the low-resolution models were most often lower than in the high-resolution model and as a result, peak flow ratios were worse. Overall runoff volume was higher in the low-resolution models, which resulted in a smaller volume error. 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.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions



4 example hydrographs flow errors.pdf

Figure 4. Calibrated hydrographs for T6_QV_ppP in the baseline run (left) and after reducing all flow measurements by 40% (right).

3.1.4 Sensitivity to structural flow measurement errors

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

3.2 Calibrated parameter values

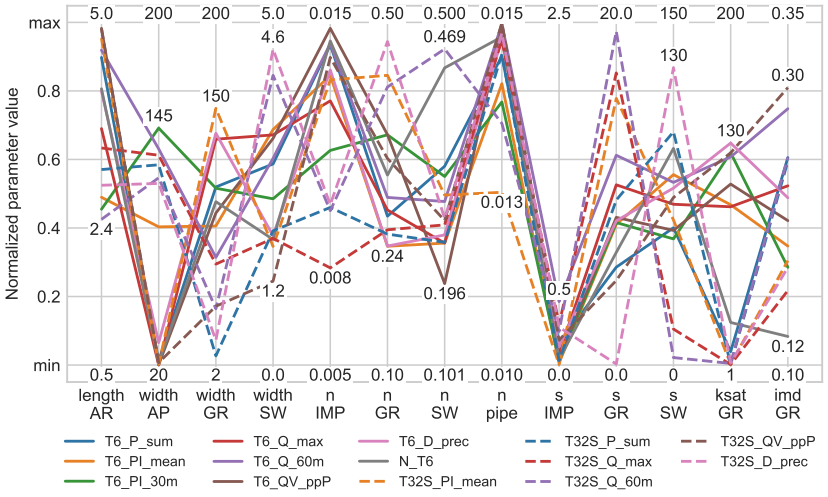
3.2.1 Hydrologic model parameters

Figure 5 shows the calibrated parameter values (for the baseline run), normalized with respect to their calibration ranges (see Table 1). There is considerable variation among the calibrated values obtained in different calibration scenarios, demonstrating that even for parameters with a clear physical interpretation, identification of the best (ideal) value is not straightforward. Gupta et al. (1998) also found considerable variation in the parameter values obtained when using different years as calibration periods for a natural catchment model. Nonetheless, the span of parameter values is considerably reduced compared to the range imposed during calibration, showing that the boundaries were not set too tightly and that the calibration procedure does offer benefits over estimating parameter values directly.

Calibrated parameter values are always uncertain estimates. This uncertainty has been investigated for urban drainage models and shown to be dependent on parameter type, study catchments, model structures, catchment discretization and measurement

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions
Open Access
EGU



5 calibrated param values.pdf

Figure 5. Normalized calibrated parameter values for different calibration scenarios and the baseline run. The highest and lowest values found for each parameter are indicated.

errors (Dotto et al., 2009, 2011, 2014; Kleidorfer et al., 2009a; Sun et al., 2014). The variation found here among the optimum parameter values obtained in different calibration scenarios suggests that the selection of calibration events could also affect the uncertainty of parameter estimates and this influence should be investigated further.

3.2.2 Rainfall multipliers

The values of rainfall multipliers found in the calibration process ranged from 0.48 to 2.92, showing that there could be significant measurement errors (in precipitation and/or flow) and/or differences between the gauge rainfall and the catchment average rainfall fitting best with the observed flow rates. For rainfall events that were included in multiple calibration scenarios, the calibrated multipliers from different scenarios were close to each other (see Table 4). This variation is much smaller than that for the hydrological model parameters (see Sect. 3.2.1). This indicates that the rainfall multipliers compensate for discrepancies between the observed and best-fitting rainfall, rather than for other aspects of catchment runoff modelling. The average value of the rainfall multipliers across all events is 1.2.

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

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

Table 4. Baseline run calibrated rainfall multipliers for events that were used in at least three CSs.

↓ Criterion Event→	1	2	3	4	5	6	7	8	9	10	11	12
T6_P_sum					1.4			1.3			1.7	1.4
T6_PI_mean				0.7	1.5	1.0	1.1		1.3			
T6_PI_30m		0.7		0.7	1.5	1.1	1.1				1.6	
T6_Q_max					1.5		1.1	1.4	1.3	2.0	1.7	
T6_Q_60m					1.5	1.1		1.3	1.2	2.1	1.6	
T6_QV_ppP										1.9	1.6	
T6_D_prec			1.3					1.3			1.7	1.5
N_T6					1.5		1.1	1.4	1.2		1.7	1.5
T32S_PI_mean				0.8	1.5	1.0	1.1		1.2			
T32S_P_sum	0.7	0.7	1.3		1.5			1.3			1.6	
T32S_Q_max	0.7				1.5		1.1	1.3		2.0		
T32S_Q_60m	0.7	0.7			1.5	1.0	1.1				1.6	
T32S_QV_ppP			1.3				1.2			1.9		
T32S_D_prec	0.7		1.2					1.4			1.7	1.5

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

With the low-resolution model, in contrast to the high-resolution model, there was considerable variation in the values of the rainfall multipliers for each event found by the different calibration scenarios, see Figure 6. The values obtained were 25% lower to 50% higher (for the same event in the same calibration scenario) than in 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, T32S_QV_ppP had three higher and three lower multipliers and other CSs had all higher multipliers. This behaviour indicates that (despite similar resulting performance) 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.

3.3 Validation performance

3.3.1 Individual events

The successful calibrations predicted 7-13 out of the 19 validation events satisfactorily (NSE >0.5), see Table 5. The two-stage calibration scenarios were less sensitive to perturbations of the flow data in the calibration period and to switching from the high resolution to the low-resolution model. T32S_P_sum, T32S_Q_max, and T32S_QV_ppP actually predicted a higher number of events satisfactorily with the low-resolution model than with the calibrated high resolution model.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

6 rainfall multis.png

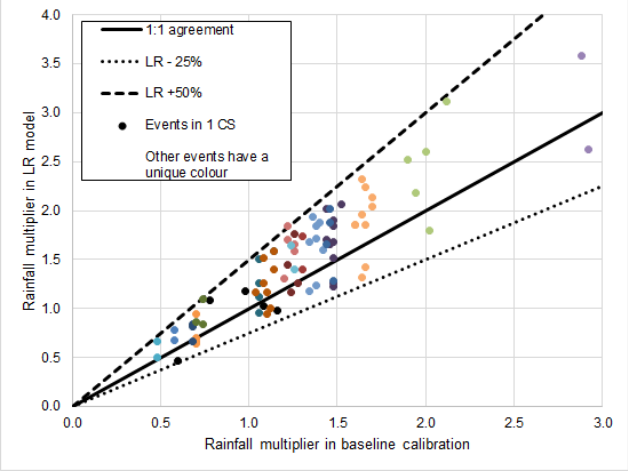


Figure 6. Rainfall multipliers in baseline calibration (horizontal axis) compared to the LR-model calibration (vertical axis). Each dot is a rainfall multiplier calibrated by one calibration scenario for one event. Identical events appearing in multiple calibration scenarios share the same colour.

Table 5. Number of validation events with NSE >0.5 out of 19 total events.

	Baseline	RMSE as obj. func.	Cal. flow -40%	Cal. flow +40%	Low-res. model	Total
N_T6	12	12	10	8	7	49
T6_D_prec	11	12	9	9	6	47
T6_P_sum	11	11	9	9	8	48
T6_PI_30m	9	9	9	9	9	45
T6_PI_mean	10	10	6	12	6	44
T6_Q_60m	8	8	9	9	6	40
T6_Q_max	12	11	9	11	10	53
T6_QV_ppP	12	11	7[a]	9	10	42
T32S_D_prec	12	12	12	12	10	58
T32S_P_sum	10	12	9	10	13	54
T32S_PI_mean	13	12	12	12	13	62
T32S_Q_60m	10	10	9	9	10	48
T32S_Q_max	11	12	8	10	12	53
T32S_QV_ppP	11	11	12	10	12	56

[a] Run was unsuccessful in calibration

The events that most often caused failure in validation were four events with peak flow rates of 10 L s^{-1} or less, and therefore, such failures may be attributed to: (i) relatively high measurement uncertainties, and (ii) high sensitivity of the NSE to even small changes in the hydrographs. However, it should be noted that the two smallest events (both with a peak flow rate

of 4.6 L s^{-1}) were predicted with NSE>0.5 by some calibration scenarios. For the other CSs, examination of the hydrographs showed that they predict well the magnitude of events, but produce wrong timing.

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

The volume errors were similar for all high-resolution calibrated models and showed a general tendency to underestimate flow volumes by 25%. When using the low-resolution model, the single-stage CSs underestimated runoff volume by around 40%, while two-stage scenarios underestimated it by a maximum of 27%. Across all CSs, two-stage versions had similar or better performance in terms of total runoff volume. Peak flow ratios were <1 for most events, but for the events that generally did poorly in validation (see above) peak flows (as well as flow volumes) were over predicted instead. The results for both total volumes and peak flows indicate that for most events flows were underestimated, which may be (at least partially) attributed to the discrepancies between observed rainfall and flow found in the calibration phase (see Sect. 3.2.2).

The peak flow ratios obtained for the 19 validation events using the calibrated models from the baseline are shown in the upper panel of Figure 7. Underestimation of peak flows was most frequent, but the largest errors occured when the flow was overestimated. The variation among CSs was generally larger when the prediction error was larger. The corresponding figure for volume errors is shown in the middle panel of Figure 7. Again, underestimation was more common, but overestimation did occur for a limited number of events. For both peak flows and total volumes, the variation among events was generally larger than the variation among different calibration scenarios, showing that selecting a limited number of validation events may also influence the results of the model evaluation.

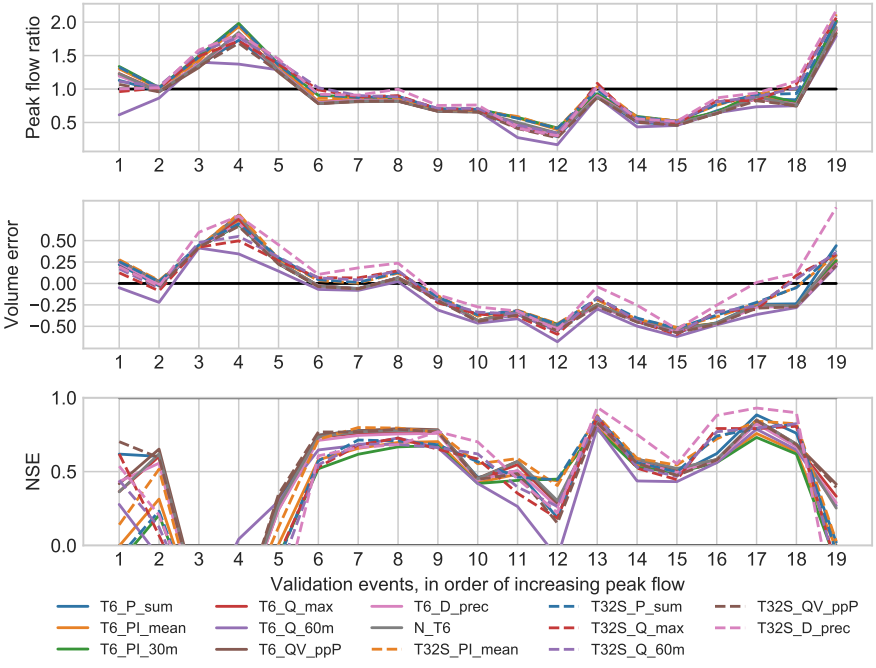
When examining the NSE of the validation events (see the bottom panel of Figure 7), more variation among the different CSs became visible, although the amount of variation was still event-dependent: inter-CS variation for the same events varies from 0.15 to 1.25. This shows that some events can have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

3.3.2 Overall performance

To assess the overall performance of different calibration scenarios 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 problems with this metric: First, since NSE ranges from negative infinity to plus one, one poorly fitting event can offset multiple well-fitting events. Second, two simulated hydrographs of equally poor fit can have rather different (negative) NSE values, producing different impacts on the overall results, which is not justified by a visual comparison. Therefore, this mean metric is not considered a reliable metric for comparisons, when poorly fitting events are present.

The exclusion of low flow (<10 L s^{-1} peak) events avoids this issue, but does not reward calibration scenarios that do manage to predict these events satisfactorily. Another option is to set all NSE values <-1 to -1 before calculating the mean,

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019



7 error stats validation events b.pdf

Figure 7. Error statistics for individual validation events for all calibration scenarios in the baseline runs.

which results in NSE ranging from 0.29 to 0.47. Adoption of the median NSEs (insensitive to outliers) lead to a higher range of 0.43 to 0.61, showing that the average or overall validation performance depends more on the outlier events than on typical events.

A more commonly used approach is to combine all the events into a single time series prior to calculating the NSE on the joint time series. This procedure indicated satisfactory performance for all CSs (NSE 0.57 – 0.70). The discussion of various metrics shows that caution is needed when averaging performance over multiple events, as metrics may not reflect the fact that a significant number of events is poorly predicted in all CSs (see Table 5).

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

To obtain an overall ranking of the different CSs in the baseline run, they were ranked by five characteristics (see Table 6) and then the sum of the individual ranks was taken. This shows that the two-stage CSs performed better in the validation period than the single-stage CSs.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

Table 6. Summarized performance for all 19 validation events for the baseline run (no flow data errors, NSE used as the objective function in calibration, high resolution model).

Criterion	NSE mean	NSE mean1	NSE median	NSE joint	# NSE <0	# NSE >0.5	RMSE joint	VE joint	PFR mean
N_T6	0.33	0.45	0.58	0.65	2	12	3.54	-0.24	0.91
T6_D_prec	0.34	0.43	0.56	0.64	2	11	3.58	-0.25	0.91
T6_P_sum	0.39	0.45	0.60	0.66	2	12	3.50	-0.23	0.91
T6_PI_30m	0.13	0.29	0.49	0.57	5	9	3.90	-0.24	0.98
T6_PI_mean	0.18	0.33	0.51	0.59	4	10	3.81	-0.24	0.96
T6_Q_60m	0.37	0.37	0.43	0.60	3	8	3.76	-0.29	0.81
T6_Q_max	0.34	0.44	0.58	0.65	2	12	3.53	-0.24	0.92
T6_QV_ppP	0.36	0.47	0.58	0.67	2	12	3.46	-0.24	0.90
T32S_D_prec	0.22	0.34	0.61	0.70	4	12	3.27	-0.02	1.00
T32S_P_sum	0.19	0.34	0.56	0.68	5	10	3.36	-0.15	0.99
T32S_PI_mean	0.26	0.44	0.59	0.70	2	13	3.27	-0.16	1.00
T32S_Q_60m	0.26	0.33	0.53	0.68	4	10	3.41	-0.13	0.99
T32S_Q_max	0.31	0.34	0.53	0.67	4	11	3.41	-0.13	0.96
T32S_QV_ppP	0.42	0.46	0.58	0.65	2	11	3.52	-0.26	0.87

3.3.3 Sensitivity to the objective function

For most calibration scenarios, the models that were calibrated with different objective functions (NSE in the baseline run, RMSE in the alternative) retained a similar performance in the validation phase. However, there are differences for some of the two-stage CSs, see Table 7 for a description and Figure 8 for an example.

3.3.4 Low-resolution model

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

The volume errors were twelve to nineteen percent points higher for the single-stage calibration scenarios. The two-stage scenarios showed both worsened performance (T32S_P_sum, T32S_PI_mean) and improved performance (T32S_Q_60m and

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Table 7. Effects of calibration with RMSE as the objective function instead of NSE.

Objective function	NSE		RMSE		
Criterion	NSE Mean[a]	# NSE >0.5 [b]	NSE Mean[a]	# NSE >0.5 [b]	Visual hydrograph comparison
N_T6	0.45	12	0.45	12	About the same for both objective functions
T6_D_prec	0.43	11	0.47	12	The same
T6_P_sum	0.45	12	0.45	11	The same
T6_PI_30m	0.29	9	0.31	9	The same
T6_PI_mean	0.33	10	0.34	10	The same
T6_Q_60m	0.37	8	0.35	8	The same
T6_Q_max	0.44	12	0.43	11	The same
T6_QV_ppP	0.47	12	0.40	11	The same
T32S_D_prec	0.34	12	0.44	12	RMSE leads to lower flow rates.
T32S_P_sum	0.34	10	0.36	12	The same
T32S_PI_mean	0.44	13	0.40	12	The same
T32S_Q_60m	0.33	10	0.37	10	RMSE leads to higher flow rates initially, but then the same
T32S_Q_max	0.34	11	0.40	12	RMSE leads to flow rates that are higher initially, but the same or lower later in the event.
T32S_QV_ppP	0.46	11	0.35	11	The same

[a] calculated after setting individual event values <-1 to -1.
[b] number of validation events (19 total) with NSE >0.5

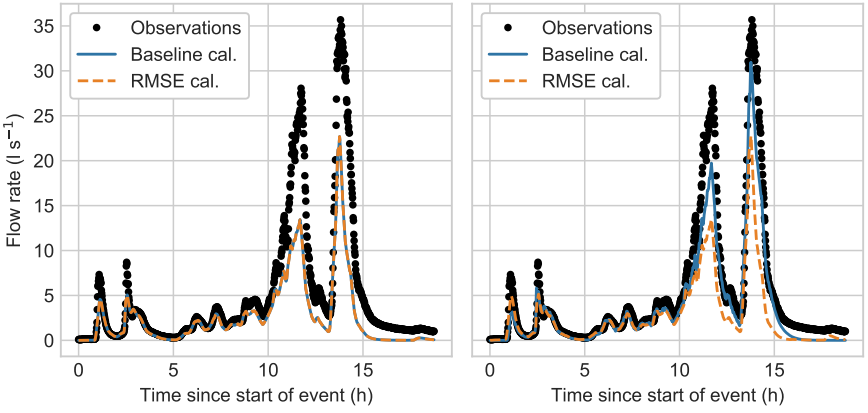
T32S_Q_max, T32S_QV_ppP). When comparing the hydrographs from the two different model discretizations per event, the high-resolution model usually performed better. However, for the last three CSs mentioned, the low-resolution performed better compared to the other CSs. For T32S_Q_60m and T32S_Q_max, the low-resolution model predicted the observed hydrographs better for most validation events. These three calibration scenarios were also the only ones where the low-resolution model resulted in lower values for the calibrated rainfall multipliers.

3.3.5 Sensitivity to structural flow errors

The introduction of structural flow measurement errors in the calibration data had little effect on performance in the validation phase. Although there were some changes 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%.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions



8 example hydrographs objective func.pdf

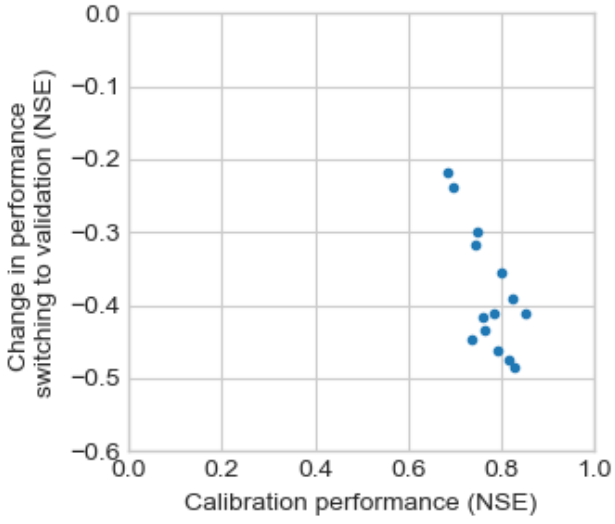
Figure 8. Examples of hydrographs showing typical (left panel, N_T6) and differing (right panel, T32S_D_prec) behaviour when calibrated for different objective functions.

Table 8. Comparison of the high resolution (HR) and low resolution (LR) models during the validation period. Bold font indicates the best value.

	NSE mean [a]		# events NSE >0.5		Volume error		Visual comparison per event	
Criterion	HR	LR	HR	LR	HR	LR	HR better	LR better
N_T6	0.45	0.21	12	7	-0.24	-0.43	17	2
T6_D_prec	0.43	0.34	11	6	-0.25	-0.44	15	4
T6_P_sum	0.45	0.22	12	8	-0.23	-0.38	16	3
T6_PI_30m	0.29	0.43	9	9	-0.24	-0.34	14	5
T6_PI_mean	0.33	0.38	10	6	-0.24	-0.43	15	4
T6_Q_60m	0.37	0.29	8	6	-0.29	-0.46	16	3
T6_Q_max	0.44	0.49	12	10	-0.24	-0.36	14	5
T6_QV_ppP	0.47	0.37	12	10	-0.24	-0.40	15	4
T32S_D_prec	0.34	0.38	12	10	-0.02	-0.05	15	4
T32S_P_sum	0.34	0.51	10	13	-0.15	-0.27	15	4
T32S_PI_mean	0.44	0.46	13	13	-0.16	-0.22	14	5
T32S_Q_60m	0.33	0.28	10	10	-0.13	-0.04	8	11
T32S_Q_max	0.34	0.33	11	12	-0.13	-0.07	7	12
T32S_QV_ppP	0.46	0.46	11	12	-0.26	-0.18	12	7

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

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019



9 loss of val performance.png

Figure 9. Loss of performance (NSE) when switching from calibration to validation

This resulted in lower flows (and therefore better fits) in validation events for the five events that caused problems for most other CSs (i.e. the four lowest and the single highest peak flow rate(s), see Sect. 3.3.1).

3.3.6 Overall ranking for validation

For an overall ranking of the different calibration scenarios in the validation period the baseline runs were ranked by each of the following statistics: mean NSE (limited to -1), number of events with NSE >0.5, RMSE (calculated over the joint time series of all events), volume error (see RMSE), and mean peak flow ratio. The ranks for each characteristic were then summed to obtain an overall ranking, see Table 9. T32S_PI_mean and T32S_D_prec performed best, with T6_PI_30m and T6_Q_60m bringing up the rear.

3.4 Degradation of performance from calibration to validation

In calibration, the NSE for the different calibration scenarios ranged from 0.68 to 0.85, while in validation it ranged from 0.29 to 0.47 Table 9. The CSs that did better in calibration lost more performance when switching to the validation period, see figure 9. Considering the change in overall rank from calibration to validation, the two-stage scenarios showed smaller changes then the single-stage scenarios. Several scenarios showed large gains (+10 for T6_QV_ppP, +7 for T6_P_sum, +5 for T32S_PI_mean) while the largest losses were smaller (-7 for T6_Q_60m, -6 for T6_Q_max). The findings in this Sect. demonstrate that good calibration performance is not necessarily indicative of good validation performance and vice versa, and therefore validation should be performed, if at all possible.

Table 9. degradation of performance when switching from the calibration to the validation period

Criterion	Calibration NSE	Validation NSE mean	Change	Rank in calibration	Rank in validation	Change
N_T6	0.80	0.45	-0.35	3	7	-4
T6_D_prec	0.74	0.43	-0.31	8	11	-3
T6_P_sum	0.75	0.45	-0.30	10	3	+7
T6_PI_30m	0.74	0.29	-0.45	13	13	0
T6_PI_mean	0.77	0.33	-0.44	11	12	-1
T6_Q_60m	0.79	0.37	-0.42	7	14	-7
T6_Q_max	0.85	0.44	-0.41	1	7	-6
T6_QV_ppP	0.68	0.47	-0.21	14	4	+10
T32S_D_prec	0.76	0.34	-0.42	2	2	0
T32S_P_sum	0.83	0.34	-0.49	5	6	-1
T32S_PI_mean	0.83	0.44	-0.39	6	1	+5
T32S_Q_60m	0.79	0.33	-0.46	9	7	+2
T32S_Q_max	0.82	0.34	-0.48	3	4	-1
T32S_QV_ppP	0.70	0.46	-0.24	11	10	+1

3.5 Single-stage vs. two-stage calibrations

For those selection criteria, for which both single and two-stage calibrations were performed, the results of the two options were compared directly (see Table 10). In terms of NSE and volume error, the two-stage calibrations performed better than the single-stage calibrations, except for Q_max. In terms of peak flow ratio the results were mixed. For D_prec and PI_mean the two-stage variant outperformed the single-stage across all metrics, for Q_max the single-stage variant performed better and for other CSs the results depended on the metric used. In validation the differences between single and two-stage calibration were less pronounced, see Table 11. In terms of NSE, the single-stage calibrations performed better, but they had the same number of satisfactorily predicted events as the two-stage calibrations. In terms of RMSE, VE and PFR the two-stage calibrations performed better, except for QV_ppP. This is also the only criterion where all metrics indicated the same, i.e. that the single-stage calibration had better results in the validation period.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Open Access
Discussions

Table 10. Comparison of single and two-stage calibration strategies in the calibration phase. The highest score for each selection criterion is highlighted.

Criterion	NSE single-stage	NSE two-stage	VE single-stage	VE two-stage	PFR single-stage	PFR two-stage
D_prec	0.74	0.76	-0.09	-0.02	0.95	0.97
P_sum	0.75	0.83	-0.07	-0.03	1.07	0.90
PI_mean	0.77	0.83	-0.04	-0.03	0.90	0.96
Q_60m	0.79	0.79	-0.09	-0.04	0.91	0.98
Q_max	0.85	0.82	-0.03	-0.06	0.89	0.86
QV_ppP	0.68	0.70	-0.11	-0.06	0.89	0.85

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-67
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 8 March 2019

Hydrology and Earth System Sciences
Discussions

Table 11. Comparison of single and two-stage strategies in the validation phase. The highest score for each selection criterion is highlighted.

Criterion	NSE[a] single-stage	NSE[a] two stage	# NSE >0.5 single-stage	# NSE >0.5 two-stage	RMSE joint single-stage	RMSE joint two stage	VE joint single-stage	VE joint two stage	PFR mean single-stage	PFR mean two stage
D_prec	0.41	0.34	11	12	3.62	3.27	-0.25	-0.02	0.92	1.00
P_sum	0.44	0.34	12	10	3.54	3.36	-0.23	-0.15	0.92	0.99
PI_mean	0.33	0.44	10	13	3.81	3.27	-0.24	-0.16	0.96	1.00
Q_60m	0.37	0.33	8	10	3.76	3.41	-0.29	-0.13	0.81	0.99
Q_max	0.44	0.34	12	11	3.53	3.41	-0.24	-0.13	0.92	0.96
QV_ppP	0.47	0.46	12	11	3.46	4.52	-0.24	-0.26	0.90	0.87

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

4 Conclusions

The objective of this study was to compare different strategies for the selection of calibration events for a hydrodynamic model of a predominantly green urban area. Calibration strategies consisted of single- and two stage calibrations and considered a number of different metrics by which calibration events can be selected from a larger group of candidate events. Calibration strategies were tested with two different objective functions, on data sets with structural flow data errors, and with high and low spatial resolution models.

In the baseline run (high resolution model, Nash-Sutcliffe as objective function, no structural flow data errors), all calibration scenarios produced successful calibrations, albeit with varying performance: NSE values ranged from 0.68 to 0.85. For the two-stage calibrations, both stages gave satisfactory results (NSE 0.70-0.87). The two-stage calibrations performed better than their single-stage counterparts in terms of NSE and runoff volume error. The choice of NSE or RMSE as the objective function had only a small impact on the results.

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

Switching from a high-resolution to a low-resolution model discretization has only a small impact on calibration performance metrics. However, the values of the rainfall multipliers for each event show much more variation than with the high-resolution models. Most high-resolution calibration models find higher values for the multipliers, but three two-stage CSs find lower values instead.

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

In the validation phase, there were again (as in the calibration) only small differences between the two considered objective functions. Concerning model discretization, the low-resolution single-stage calibration scenarios show significantly larger volume errors than their high-resolution counterparts, while most two-stage calibration scenarios show either the same or even improved volume errors. Two 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.

An overall ranking of the different scenarios across the different influential factors (objective function, flow data errors, model discretization) showed that T6_Q_max, T32S_D_prec and N_T6 performed the best in calibration. However, in the validation phase this order was changed considerably with T32S_PI_mean, T32S_D_prec and T6_P_sum forming the top three. The ranking of the two-stage scenarios was more consistent between calibration and validation than that of the single-stage scenarios.

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

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

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

References

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

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

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

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

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

Del Giudice, D., Albert, C., Rieckermann, J., and Reichert, P.: Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation: IMPROVING INPUT UNCERTAINTY QUANTIFICATION, 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Tscheikner-Gratl, F., Zeisl, P., Kinzel, C., Leimgruber, J., Ertl, T., Rauch, W., and Kleidorfer, M.: Lost in calibration: why people still do not calibrate their models, and why they still should – a case study from urban drainage modelling, Water

Science and Technology, 74, 2337–2348, https://doi.org/10.2166/wst.2016.395, https://iwaponline.com/wst/article/74/10/2337/19429/Lost-in-calibration-why-people-still-do-not, 2016.

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