Event selection and two-stage approach for calibrating models of green urban drainage systems

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 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 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 NSE and the calibrated parameter values notably varied. The various calibration strategies satisfactorily predicted 7 to 13 out of 19 validation events. The two-stage strategies reproduced more validation events poorly (NSE < 0) 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 runoff volume and peak flow rates, 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 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

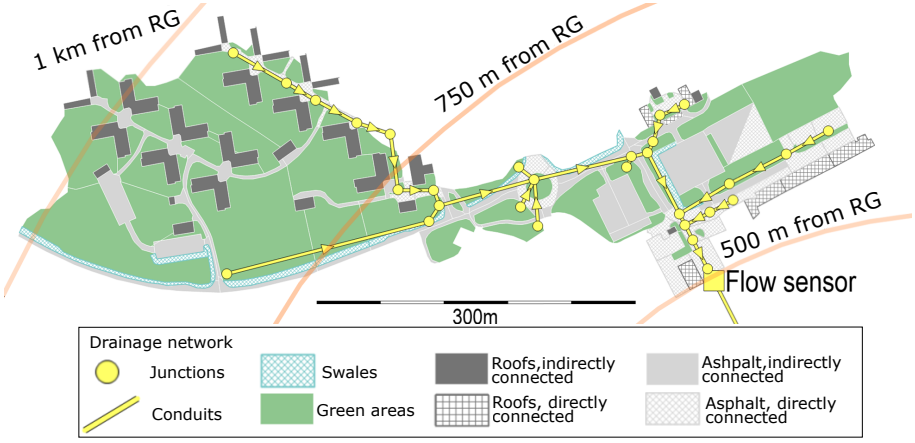
Calibration of generic urban drainage model codes is usually required to obtain a model representing an actual site with sufficient accuracy. In the calibration process, the information contained in records of relevant variables, such as rainfall and flow rates at the catchment outlet, is used for estimating 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). 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 (preliminary findings were published in Broekhuizen et al. (2019)), although split-stage calibration where different parameters affect different points or properties of the hydrograph have been investigated for natural catchments (see e.g. Fenicia et al. (2007); Gelleszun et al. (2017)).

Considering the above findings, the primary objective of this paper 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. Included in this is a proposal for a practical two-stage calibration strategy where parameters related to impervious and green areas are calibrated in two separate steps using different sets of events.



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.

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.

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 ±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 a minimum inter-event time of no precipitation of six hours. Events deemed suitable for use in calibration were selected using the following criteria:

1. A minimum total precipitation of 2 mm (Hernebring, 2006).
2. No or small gaps in rain and flow data , i.e. both have to be available for >90% of the event duration.
3. Sufficient in-pipe water depths for the flow sensor to work reliably: >10 mm during at least 50% of the event and >25 mm at least once in the event, based on recommendations from the manufacturer (Teledyne ISCO, 2010).
4. Peak flow >2 L s^{-1}, since relative measurement uncertainties are high below this point.
5. No snowfall or -melt, since these would introduce additional processes in the hydrological behaviour and model of the catchment.

Calibration and validation periods were separated by using the 19 observed events from 2016 for the validation period, and the 32 events from 2013-2015 and 2017 for the calibration period. In this way, all the calibration strategies (see Sect. 2.3) were tested (validated) against the same dataset and no calibration 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. Table 1 contains an overview of all events that were used in at least one calibration strategy as well as an initial estimate of the runoff from green areas.

2.2 Runoff model and calibration approach

The US EPA Storm Water Management Model (SWMM) was selected since it is a commonly used semi-distributed urban drainage model 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 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. Some spatial characteristics, such as the slope and the width of subcatchments, can 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).

Table 1. Characteristics of all rainfall events used in one or more calibration strategies.

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

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

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

[c] Calculated as c = a – b

[d] Calculated as d = c / P_sum

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 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 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.

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 hy-

Table 2. Calibration parameters and their ranges.

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

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

Parameter	Abbr.	Groups	Range	Reference
Subcatchment width [m]	width	5 individual subcatchments	20 – 200	Physical dimensions of subcatchments
Manning's coefficient [-]	n	Impervious surfaces (IMP)	0.005 – 0.015	(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.

drological 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 strategies (CS), with each CS using six rainfall events. The single-stage CSs used the six events with the highest values for a given event characteristic, and calibrated all parameters simultaneously. Two-stage calibration 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. 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. 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. 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 strategies, the six highest ranking events for each characteristic were selected. For the two-stage calibration strategies, the three highest ranking events with less than 12% runoff were selected for the first stage and the three highest ranking events with more than 12% runoff were selected for the second stage. Applying the calibrated rainfall multipliers in the calibration (Sect. 2.2) means that event properties relating to rainfall and percentage runoff will change, and the percentage runoff can change from <12% to >12% and vice versa. 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 strategy, (3) re-calculating which events are used in each CS, and (4) repeating the calibration for any CS that has had any of its events changed. Although this might improve the overall results of the proposed calibration procedure, it would also increase the complexity and raise several new issues, such as how to obtain a calibrated rainfall multiplier for the 10 events that were not used in any CS. We considered this to be beyond the paper's original scope of examining different strategies for calibration event selection and proposing a practically useable two-stage calibration procedure.

To avoid making the comparison too large in scope, a limited number of calibration 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 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), 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 strategy N_T6 consists of the six events that were selected most often in other calibration strategies with the goal of obtaining a set of events that score highly on a variety of characteristics.

2.4 Objective functions

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

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

Where O denotes observed values and S simulated values. The NSE measures the variance of the model errors (the numerator) as a fraction of the variance of the observations (the denominator). This fraction is then scaled so that it extends from –infinity (i.e., the worst possible fit) via 0 (the score that would be achieved by using the average of observations) to 1, for a perfect fit. The NSE is dimensionless, so it allows comparing runoff events of different magnitudes. However, when the variance of the observations is small (e.g. for small runoff events), it can become quite sensitive to small changes in the simulated hydrograph. The NSE was calculated for each individual event and the mean over all events 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:

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

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

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

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

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

3 Results and discussion

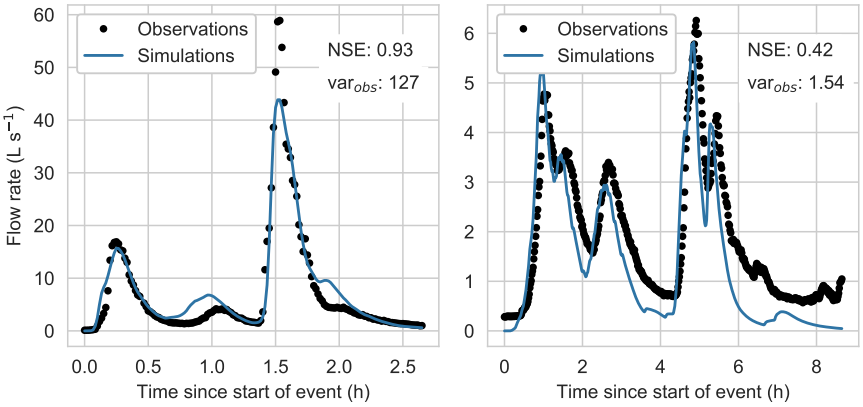
3.1 Calibration performance

The high-resolution (HR) model was successfully calibrated for all calibration strategies, with event mean NSE for all events ranging from 0.68 to 0.85 (see Table 4). The lowest event mean 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 right 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 strategies, the individual stages also produced successful calibrations (stage 1 event mean NSE 0.70 – 0.87, stage 2 event mean NSE 0.78-0.87), except for the second stage in T32S_QV_ppP for the reasons explained above. The NSE values for the individual calibration events in the different calibration strategies are similar to those reported by Krebs et al. (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 event mean NSE) than the single-stage calibration while for Q_max the single-stage calibration performed better (0.03 event mean 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 event mean NSE > 0.8 in both the HR and LR models) while the two scenarios based on percentage runoff performed worst.

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, as also reported by Tscheikner-Gratl et al. (2016)), the single stage calibrations had smaller volume errors. The changes in volume errors between the HR and LR model were similar to earlier findings by Krebs et al. (2016). Although the CSs based on peak flow rates (Q_max) performed well in terms of event mean 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.

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

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



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 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

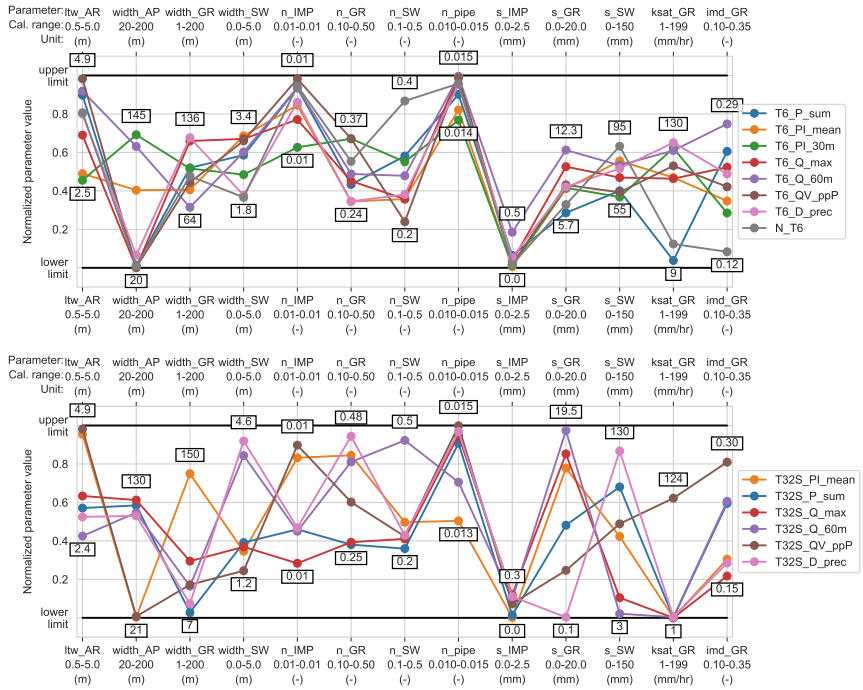
when the calibrated parameter values from the second stage were applied. This runoff was caused by impervious areas draining to green areas. The runoff from green areas was <5% of the total simulated runoff volume for 4 model runs, <10% for an additional 3 runs, and 11.6%, 11.7%, 21.7%, 22.9% and 25.7% respectively for 5 additional runs. Note that with 6 CSs with 3 first-stage events each, there were 18 model runs in total. The last mentioned 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 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.

3.2 Calibrated parameter values

3.2.1 Hydrologic model parameters

Figure 3 shows the calibrated parameter values (for the HR model), normalized with respect to their calibration ranges (see Table 2). There is considerable variation among the calibrated values obtained in different calibration 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 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).

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



param values.pdf

Figure 3. Normalized calibrated parameter values for the high-resolution model using different calibration strategies. The highest and lowest values found for each parameter are indicated.

parameter values obtained in different calibration 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

The values of rainfall multipliers found in the calibration process ranged from 0.48 to 2.92, indicating a mismatch between the observed rainfall and the rainfall that allows for the best fit of the simulated runoff to observed runoff. Several factors may contribute to this. First, underestimation of rainfall or underestimation of runoff by the respective sensors may lead to higher rainfall multipliers and vice versa. Errors in the size of (sub)catchments may also influence this. Second, the gauge rainfall may not match the catchment-averaged rainfall due to the spatial variability of the rainfall. Thirdly, some deficiencies in the model may be compensated for, to some extent, by adjusting the rainfall multiplier. Without further investigations it is not possible to distinguish between different factors influencing the values of rainfall multipliers. 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 strategies, the calibrated multipliers from different scenarios were close to each other (see Table 5), unlike for the hydrological model parameters (see Sect. 3.2.1). Secondly, decreasing or increasing all flow rates by 40% prior to calibration

Table 5. Calibrated rainfall multipliers (HR model) for all rainfall events that were used in at least 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[a]
211		0.70	0.70		0.70	0.70									0.70	6.8	15.8[a]
214						1.16									1.16	7.4	8.7
222			0.68		0.68					0.68					0.68	6.7	10.6
270		1.24	1.22				1.28	1.26							1.25	11.7	9.1
306				0.74						0.70	0.74				0.73	6.3	11.7
307	1.48		1.46	1.48	1.48	1.48			1.48	1.44	1.44	1.52	1.48		1.47	44.0	11.0[b]
310				1.06	1.06					1.06	1.06	1.14			1.08	9.2	13.0
530	1.14			1.10	1.10	1.12	1.04			1.08	1.08		1.14		1.10	7.4	10.2
939		0.60													0.60	4.2	9.5
962														0.98	0.98	8.3	25.4
971							1.08								1.08	2.8	10.4
978	1.38	1.38	1.34			1.34		1.40	1.42			1.36	1.38		1.38	34.4	13.9
982	1.22			1.20							1.26	1.22	1.26		1.23	6.9	12.8
984						2.02	1.94					2.12	2.00	1.90	2.00	4.8	29.6
995							2.92							2.88	2.90	6.1	9.9 [b]
997								1.24	1.26						1.25	30.8	16.6
1001	1.70	1.66	1.60		1.64			1.66	1.66	1.60		1.64	1.70	1.64	1.65	58.2	15.1
1004														0.78	0.78	3.3	32.3
1019	1.46	1.48						1.46	1.44						1.46	32.6	14.5
1028							1.30							1.30	1.30	3.7	33.4

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

[b] Vice versa.

changed the average rainfall multipliers by -37% and +33% respectively. The average value of the rainfall multipliers across all events was 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.

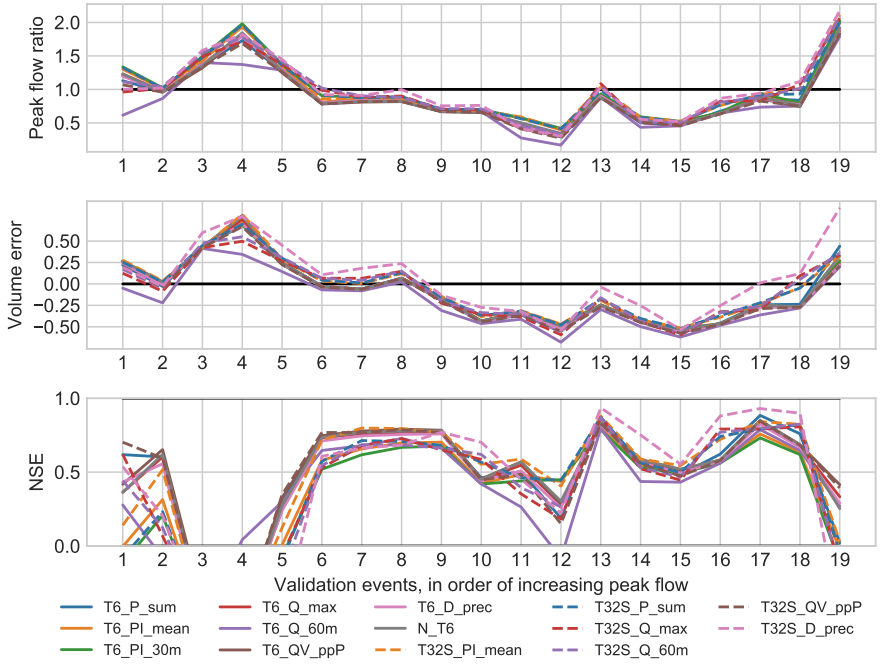
3.3 Validation performance

3.3.1 Individual events

The 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^{-1} 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 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^{-1}) were predicted with NSE > 0.5 by T6_P_sum and T32S_QV_ppP. For the other CSs, examination of the hydrographs showed that they predicted well the peak flow and the total runoff volume of the events, but produced wrong timing compared to the observed hydrograph. Another event that failed in validation for all CSs was that with the highest peak flow rate (53 L s^{-1}, 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^{-1}), 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 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 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 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 volumes and peak flows indicate that for most events flows were underestimated, which may be (at least partially) attributed to the need to multiply the observed rainfall by (on average) a factor of 1.2 to best match the observed flow during the calibration phase (see Sect. 3.2.2); such adjustment was not applied in the validation phase.

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: the difference (in NSE) between the best and worst CS for the same events varied 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



error stats validation events.pdf

Figure 4. Error statistics for individual validation events for all calibration strategies in the HR model.

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

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 with the LR model than with the HR model, and T32S_P_sum, T32S_Q_max, and T32S_QV_ppP actually predicted more events satisfactorily with the HR model.

To assess the overall performance of different calibration 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 L s^{-1} peak) events would avoid this issue, but would not reward calibration strategies that do manage to predict these events satisfactorily. Another option is to set all NSE values <-1

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

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

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

[b] calculated after merging all event time series into a single series.

[c] Number of events with NSE < 0

[d] Number of events with NSE > 0.5

to -1 before calculating the mean, which results in mean NSE ranging from 0.29 (T6_PI_30m) to 0.47 (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 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 downside of this metric is that it can hide the fact that poorly-fitting events are present, e.g. T32S_P_sum has the (shared) 3rd highest joint NSE, despite having five events with negative 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. 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

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

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

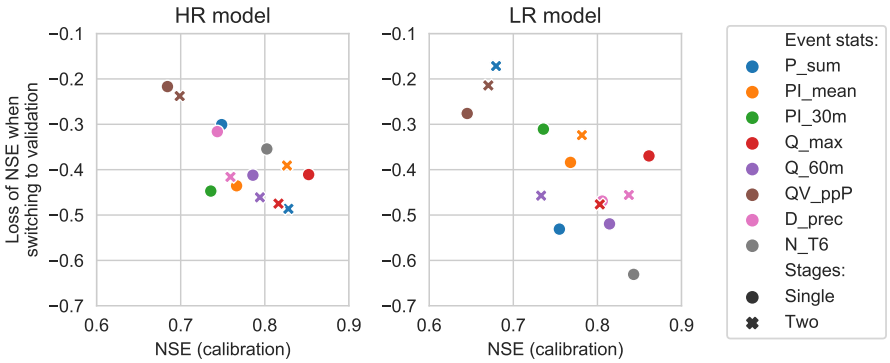
[b] calculated after merging all event time series into a single series.

[c] Number of events with NSE < 0

[d] Number of events with NSE > 0.5

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), 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 CSs than for the single-stage CSs. The CSs based on rainfall intensity (PI) showed the best performance in terms of peak flows. T6_Q_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.

Most of the LR, two-stage calibrations had higher event 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, 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_max and T32S_Q_60m) outperformed the HR model in the visual comparison of hydrographs.



v20190710 loss val performance.pdf

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

3.4 Degradation of performance from calibration to validation

In calibration, the event mean 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. 3.3.2). 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 event mean NSE) when switching to the validation phase (see Figure 5). 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 model all but one of the two-stage calibrations lost more performance when switching to the validation phase than their single-stage counterparts. By contrast, for the low-resolution model all but one of the two stage calibrations had a smaller performance loss. Previous studies found that high-resolution models lead to more transferable parameter estimates (e.g. less loss of performance when switching to validation, Sun et al. (2014), Krebs et al. (2014)), but in the current study this seems dependent on the calibration data set used.

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 event mean NSE was better for the single-stage 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 event mean 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.

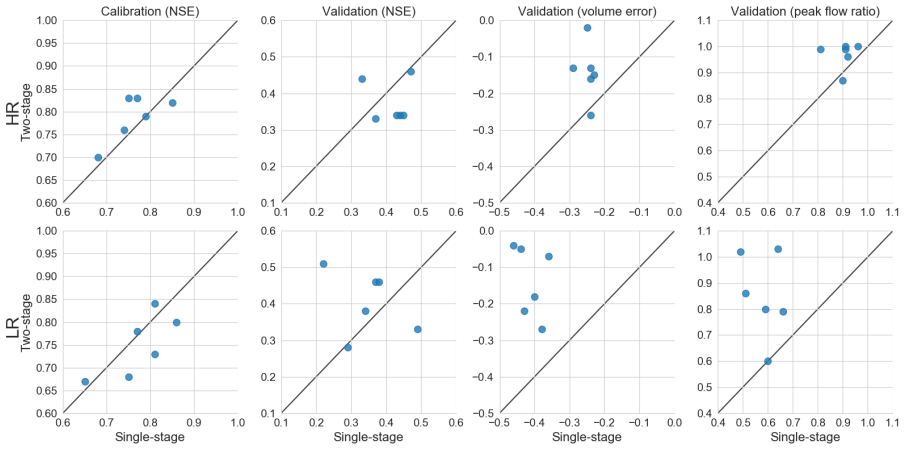


Figure 6. Comparison of single-stage and two-stage calibration strategies.

4 Conclusions

The primary objective of this study was to compare different strategies for the selection of calibration events for a combined hydrologic-hydrodynamic model of a predominantly green urban area. 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. 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 are summarized below. It should be noted that the precise performance values presented in this paper may vary for different catchments and datasets.

For the high-resolution model, all calibration strategies produced successful calibrations (i.e., NSE > 0.5), albeit with varying performance: event mean NSE values ranged from 0.68 to 0.85. For the two-stage calibrations, both stages gave satisfactory results (event mean NSE 0.70-0.87). The two-stage calibrations generally performed better in the calibration phase than their single-stage counterparts in terms of event mean 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.

For the model validation phase an independent set of 19 validation events was used. All calibrated scenarios predicted 8 to 13 of these events satisfactorily (NSE >0.5). 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 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 event mean NSE for the two-stage CSs. For volume error and peak flow error the two-stage CSs performed better, 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 two CSs where the LR version outperformed the HR version when visually comparing the hydrographs.

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 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, and this effect was particularly strong for the LR model.

Code and data availability. Rainfall and flow data are available from the first author upon request. The model is available in a non-georeferenced form since it is based on proprietary data. The calibrations were carried out with the SPOTPY library (https://github.com/thouska/spotpy).

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. We gratefully acknowledge the financial support provided by the Swedish Research Council Formas (grant number 2015-121) 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

Broekhuizen, I., Leonhardt, G., Marsalek, J., and Viklander, M.: Selection of Calibration Events for Modelling Green Urban Drainage, in: New Trends in Urban Drainage Modelling, edited by Mannina, G., pp. 608–613, Springer International Publishing, Cham, 2019.

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.

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., 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.

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.

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.

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.

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.

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., 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.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[c] Calculated as c = a – b

[d] Calculated as d = c / P_sum