Hydrol. Earth Syst. Sci., 17, 4159–4176, 2013
www.hydrol-earth-syst-sci.net/17/4159/2013/
doi:10.5194/hess-17-4159-2013

Hydrology and Earth System Sciences
Open Access

Informal uncertainty analysis (GLUE) of continuous flow simulation in a hybrid sewer system with infiltration inflow – consistency of containment ratios in calibration and validation?

A. Breinholt[1,3], **M. Grum**[2], **H. Madsen**[3], **F. Örn Thordarson**[3], **and P. S. Mikkelsen**[1]

[1]Department of Environmental Engineering (DTU Ennvironment), Technical University of Denmark, Building 113, 2800 Lyngby, Denmark
[2]Krüger, Veolia Water Solutions and Technologies Denmark, Gladsaxevej 363, 2860 Søborg, Denmark
[3]Department of Informatics and Mathematical Modelling (DTU Informatics), Technical University of Denmark, Building 305, 2800 Lyngby, Denmark

Correspondence to: A. Breinholt (anbre@env.dtu.dk)

Received: 15 April 2012 – Published in Hydrol. Earth Syst. Sci. Discuss.: 12 July 2012
Revised: 29 August 2013 – Accepted: 4 September 2013 – Published: 24 October 2013

Abstract. Monitoring of flows in sewer systems is increasingly applied to calibrate urban drainage models used for long-term simulation. However, most often models are calibrated without considering the uncertainties. The generalized likelihood uncertainty estimation (GLUE) methodology is here applied to assess parameter and flow simulation uncertainty using a simplified lumped sewer model that accounts for three separate flow contributions: wastewater, fast runoff from paved areas, and slow infiltrating water from permeable areas. Recently GLUE methodology has been critisised for generating prediction limits without statistical coherence and consistency and for the subjectivity in the choice of a threshold value to distinguish "behavioural" from "non-behavioural" parameter sets. In this paper we examine how well the GLUE methodology performs when the behavioural parameter sets deduced from a calibration period are applied to generate prediction bounds in validation periods. By retaining an increasing number of parameter sets we aim at obtaining consistency between the GLUE generated 90 % prediction limits and the actual containment ratio (CR) in calibration. Due to the large uncertainties related to spatio-temporal rain variability during heavy convective rain events, flow measurement errors, possible model deficiencies as well as epistemic uncertainties, it was not possible to obtain an overall CR of more than 80 %. However, the GLUE generated prediction limits still proved rather consistent, since the overall CRs obtained in calibration corresponded well with the overall CRs obtained in validation periods for all proportions of retained parameter sets evaluated. When focusing on wet and dry weather periods separately, some inconsistencies were however found between calibration and validation and we address here some of the reasons why we should not expect the coverage of the prediction limits to be identical in calibration and validation periods in real-world applications. The large uncertainties result in wide posterior parameter limits, that cannot be used for interpretation of, for example, the relative size of paved area vs. the size of infiltrating area. We should therefore try to learn from the significant discrepancies between model and observations from this study, possibly by using some form of non-stationary error correction procedure, but it seems crucial to obtain more representative rain inputs and more accurate flow observations to reduce parameter and model simulation uncertainty.

1 Introduction

Simulation with deterministic urban drainage models is commonly used to assess the performance of sewer systems and to assess the efficacy of new upgrading or redesign proposals. Rarely are uncertainties addressed in these investigations, and decisions with large economic consequences are usually taken on a purely deterministic basis, as if model simulations were in full conformity with reality. Sometimes models are

Published by Copernicus Publications on behalf of the European Geosciences Union.

calibrated to level or flow data from a few places in a sewer system during some months. However, you need not to have much experience with calibration of urban drainage models before you arrive at the conclusion that different parameter sets are optimal for different rain events, even when applying state-of-the-art, physically distributed models in combination with high-resolution rain gauges located close to the catchment in question.

Different parameter sets, sometimes referred to as different models, will obviously have different consequences when applied in a long-term simulation setting typically used as a basis for evaluating upgrade proposals, a fact that is however mostly ignored in practice. There is thus an urgent need for uncertainty assessment tools that can be used when evaluating upgrade proposals as well as for associated needs such as flow meter checking and evaluating the magnitude of the unintended infiltration contribution to the sewer flow, which constitutes a major problem in many flat coastal urban catchment areas.

The generalized likelihood uncertainty estimation (GLUE) methodology (Beven and Binley, 1992; Beven and Freer, 2001) acknowledges that multiple parameter sets (models) may provide acceptable simulations of the response of the system of interest (Beven, 2006). GLUE has become an increasingly popular tool for model evaluation and uncertainty estimation of environmental models (Mitchell et al., 2009; Piñol et al., 2009; Juston et al., 2010; Staudt et al., 2010) and particularly within hydrological modelling from where the methodology originated (see e.g. Choi and Beven, 2007; Xiong and O'Connor, 2008; Blazkova and Beven, 2009a, b; Jin et al., 2010). Several GLUE applications have also been seen within urban drainage water quantity and quality modelling, (Aronica et al., 2005; Lindblom et al., 2007; Freni et al., 2008, 2009b, a; Mannina and Viviani, 2010; Lindblom et al., 2011), but GLUE, as well as Bayesian inverse methods (e.g. Dotto et al., 2009; Kleidorfer et al., 2009; Dotto et al., 2010; Freni and Mannina, 2010; Dotto et al., 2012), have so far mostly been applied to tailor-made models for relatively simple, well-defined urban drainage systems or in combination with high-quality data generated in research projects. Within flow modelling uncertainty is introduced from unreliable/inaccurate level or flow meters (Bertrand-Krajewski et al., 2003), inadequate rain gauge coverage (Willems, 2001; Vaes et al., 2005; Pedersen et al., 2010), and/or unreliable/inaccurate rain gauge measurements (input errors) (Barbera et al., 2002; Molini and Barbera, 2005; Shedekar et al., 2009).

In this paper we present an application of GLUE to a hybrid urban drainage system revealing the full complexity of reality in terms of flow variations (diurnal wastewater variations, fast rainfall runoff from paved areas and slow infiltration inflow from unknown sources), using flow data recorded by the responsible utility over three consecutive years. A state-of-the-art physically distributed model fed with comprehensive information about the system attributes is currently used by the local utility to interpret the measurements. We use a lumped, conceptual model to reduce the computational burden, but this model however represents the complex flow contributions mentioned above in a similar manner to the physically distributed model used in practice.

Recently the GLUE methodology was criticized for being statistically incorrect and for generating prediction limits without statistical coherence (Mantovan and Todini, 2006; Mantovan et al., 2007; Stedinger et al., 2008). This is due to the subjectivity in adopting a likelihood measure and in the choice of a threshold value to distinguish "behavioural" from "non-behavioural" parameter sets. In GLUE, modelling errors associated with each acceptable model are usually treated under the assumption that error series associated with a particular parameter set (such as over- or under-prediction of flow peaks) will be similar in prediction to those found in evaluation (Blazkova and Beven, 2009b) and hence GLUE is in many cases a welcomed alternative to traditional statistical inference that requires the error series to conform to a statistical known distribution often difficult to justify in real hydrological applications (Beven et al., 2008). It is in this context worth noting that the aforementioned papers that have criticised the GLUE approach all have used synthetic data to illustrate and consolidate their critique, and hence there seems to be a lack of research papers that clearly demonstrate that the statistical error assumptions conform to the specified likelihood function in real-world hydrological applications. In the synthetic case the benefits of classical statistical inference are evident: trust in the model is build in the model construction phase and confidence bounds can be generated and used for prediction. In Beven and Freer (2001); Beven et al. (2011) it is claimed that any effects of model nonlinearity, covariation of parameter values and errors in model structure, input data or observed variables, with which the simulations are compared, are handled implicitly within the GLUE procedure.

The scope of this paper is to examine the GLUE assumption that the error series associated with a particular parameter set will be similar in prediction to those found in evaluation. If true, we would expect that the performance of the GLUE derived uncertainty limits obtained in a calibration period should be similar in a validation period. Aiming at an overall coverage of 90 % of the observations, we investigate how well the GLUE generated 90 % prediction limits cover the observations in both dry and wet weather periods as the number of behavioural parameter sets increases, and we moreover check the coverage for different flow magnitudes using half a year for calibration. Validation periods are included to test the consistency of the generated prediction limits, that is, we test if the coverage obtained in validation periods corresponds to the coverage obtained in the calibration period. We also show how the limits of the posterior parameter space increases as more parameter sets are retained and use this information to draw conclusions on the physical interpretation of important model parameters, such as the

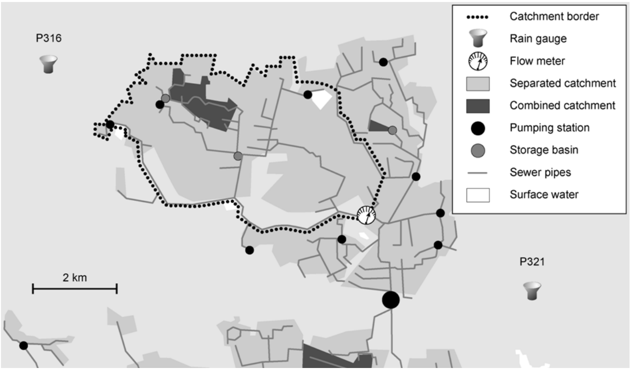

Fig. 1. The Ballerup catchment area.

Table 1. Catchment details.

Ballerup	Total area [ha]	Total area [%]	Imp.area [ha]	Imp.area [%]
Combined	92	7	33	77
Separated	1227	93	10	23
Total	1320	100	43	100

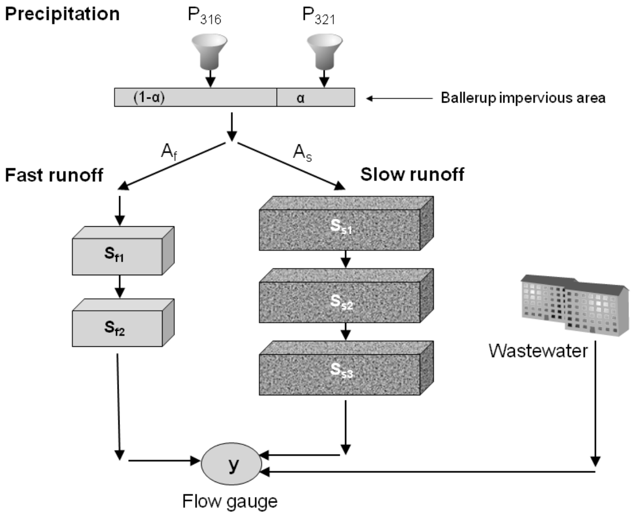

Fig. 2. The conceptual model.

size of contributing paved area versus the size of the area contributing with slow infiltration inflow. After this brief introduction, we first present the case study area, the calibration and validation data, and the model in Sect. 2.

This is followed by an elaboration of the applied uncertainty analysis methodology in Sect. 3 in which the GLUE steps are outlined, the used combined likelihood measure is defined, and some performance indicators are presented. Finally, the results are presented and discussed in Sect. 4 and conclusions are drawn in Sect. 5, both with respect to the urban drainage engineering relevance and the method applicability.

2 Case study and model

2.1 Catchment and drainage system

The case study catchment with a total area of 1320 ha is situated in the western part of greater Copenhagen in the Ballerup Municipality, as shown in Fig. 1. Most of the area (93 %) is equipped with a separated sewer system, that is, a system with two parallel pipes for wastewater and stormwater, whereas only 7 % is equipped with a combined system where wastewater and stormwater flows into the same pipe (see Table 1). Such hybrid systems are quite common due to transition of the prevailing technological regime in urban drainage since the 1950s, from combined to separated systems.

In a recent calibration of a distributed hydrodynamic model with a rainfall dependent infiltration-inflow module (DHI, 2009) the effectively contributing impermeable area of the combined sewer system was however found to be larger than that of the separated area (see Table 1), probably because of infiltration inflow or unintended connections of drainage water to the wastewater system. A flow meter has been installed downstream from the catchment (Fig. 1) with the aim of detecting these contributions. The flow meter is a semi-mobile ultrasonic Doppler type and is placed in an intercepting concrete pipe ($d = 1.4$ m and slope 1.1 ‰ i.e. a potential gravity driven flow capacity of approx 2000 L s^{-1}), and logs every 5 min. There are roughly 50 000 inhabitants within the catchment area, which is one of several sub-catchments that divert water to the second largest wastewater treatment plant (WWTP) in Denmark, called Avedøre WWTP. There are a couple of small pumping stations and one larger storage basin within the catchment of approx 4000 m^3. The two closest rain gauges from the national Danish tipping bucket network (0.2 mm resolution; Jørgensen et al., 1998), P316 and P321 indicated on Fig. 1, are located outside the studied catchment area some 12 km apart.

2.2 Hydrological model

The primary scope with the paper is to test the usability of the GLUE methodology as a tool for uncertainty analysis and estimation in complex urban drainage modelling, that is, by evaluating how well the GLUE methodology performs when the behavioural parameter sets deduced from a calibration period are applied to generate prediction bounds in a validation period. For this test we decided to keep the model simple and compare observed and modeled flow at just one place downstream from the considered catchment. Hence we inferred that a state-of-the-art physically distributed model that calculates flow and levels in every pipe of the sewer

Table 2. Model equations.

Fast runoff:
$S_{f1,k+1} = \left(\alpha A_\mathrm{f} P_{316,k} + (1-\alpha) A_\mathrm{f} P_{321,k} - \frac{2}{K_\mathrm{f}} S_{f1,k}\right) \Delta t + S_{f1,k}$
$S_{f2,k+1} = \left(\frac{2}{K_\mathrm{f}} S_{f1,k} - \frac{2}{K_\mathrm{f}} S_{f2,k}\right) \Delta t + S_{f2,k}$
Slow runoff:
$S_{s1,k+1} = \left(\alpha A_\mathrm{s} P_{321,k} + (1-\alpha) A_\mathrm{s} P_{316,k} - \frac{3}{K_\mathrm{s}} S_\mathrm{s}1,k\right) \Delta t + S_{s1,k}$
$S_{s2,k+1} = \left(\frac{3}{K_\mathrm{s}} S_{s1,k} - \frac{3}{K_\mathrm{s}} S_{s2,k}\right) \Delta t + S_{s2,k}$
$S_{s3,k+1} = \left(\frac{3}{K_\mathrm{s}} S_{s2,k} - \frac{3}{K_\mathrm{s}} S_{s3,k}\right) \Delta t + S_{s3,k}$
Wastewater:
$D_k = a_0 + \sum_{i=1}^{2} \left(s_i \sin \frac{i2\pi k}{L} + c_i \cos \frac{i2\pi k}{L}\right)$
Observation equation:
$y_k = \frac{2}{K_\mathrm{f}} S_{f2,k} + \frac{3}{K_\mathrm{s}} S_{s3,k} + D_k$

system would be overly complex for the purpose considering both the computational requirements and the risk of overparameterization. Instead a simple modeling approach was chosen yet complex enough to describe the major flow components (diurnal wastewater variations, fast rainfall runoff from paved areas and slow infiltration inflow from unknown sources). The limitation of using such a simplistic modelling approach is that the model may be too simplistic, for example, in cases when system components, such as weirs, gates, pumping stations and storage tanks, play a significant impact on the observed flow or in cases with heavy backwater effects.

In a GLUE study of an urban drainage system Thorndahl et al. (2008) however applied a distributed hydrodynamic model and showed that the hydraulic parameters (Manning number and minor losses) played an insensitive role when extracting the behavioural parameters of the model, while the surface runoff part of the model (particularly the hydrological reduction factor and time of concentration) were very sensitive. Replacing a full hydrodynamic model normally used in practice with a lumped, conceptual hydrological model as depicted in Fig. 2 therefore seems adequate.

The model consists of two linear reservoirs for modelling the fast rainfall runoff relationship (representing the paved area of the system), and three linear reservoirs for modelling of the slow infiltration inflow to the sewer system. A double sinusoidal black box model was used for modelling the diurnal wastewater flow. Model equations are displayed in Table 2 while a nomenclature is provided in Table 3.

A time step of 15 min was used during both calibration and simulation, which is sufficient for a catchment this size where the concentration time is at least a few hours. The inputs to the model are measured precipitation from the two rain gauges, P_{316} and P_{321}, and α is a weighting factor governing the percentage of the total area that each rain gauge represent.

Table 3. Nomenclature.

Symbol	Description	Unit
Inputs:		
P_{316}	Rain gauge input	$\mathrm{m\,h^{-1}}$
P_{321}	Rain gauge input	$\mathrm{m\,h^{-1}}$
Rainfall-runoff parameters:		
A_f	Impermeable fast runoff area	ha
K_f	Retention time, fast runoff	h
α	Rain gauge weighting coefficient	–
A_s	Impermeable slow-runoff area	ha
K_s	Retention time, infiltration runoff	h
Wastewater flow parameters:		
a_0	Average wastewater flow	$\mathrm{m^3\,h^{-1}}$
s_1,s_2	Sine constants	–
c_1,c_2	Cosine constants	–
Model states:		
S_f1, S_f2:	Model states, fast runoff	$\mathrm{m^3}$
S_s1, S_s2:	Model states, infiltration runoff	$\mathrm{m^3}$
Outputs:		
y_k	Observed flow at time step k	$\mathrm{m^3\,h^{-1}}$
Time:		
k	Time step counter	–
Δt	Time step	0.25 h
Other:		
N	Number of observations	–
K	Number of retained parameter sets	–

2.3 Calibration and validation data

Data from half a year (April–October, 2007) was used for calibration. This period was selected because summer normally carries the heaviest rains. The length of the calibration period was chosen by considering a typical length of measuring campaigns used for calibration of urban drainage models; these campaigns usually last only 3–4 months. Two subsequent years (2008 and 2009) of the same season (April–October) were included for validation. There have been no significant changes of the sewer system since 2007, and a good basis for validating the GLUE generated prediction limits thus exists. Some flow data from the calibration period (10 %) and validation periods (1 % and 1.5 %) had to be discarded from the analysis as they were obviously erroneous; the rain data had already been subject to standardised quality control as described by Jørgensen et al. (1998).

The measured precipitation in the studied period was quite different from one year to the other and large spatial variation was observed. Figure 3 shows the accumulated precipitation measured by each rain gauge plotted against each other on a shifted log scale, for each of the years considered. Events plotted for $P_{321} = 0$ have only been recorded at P_{316}, whereas events plotted for $P_{316} = 0$ have only been recorded at P_{321}, that is, these are probably convective events with limited spatial extent. The rest are events that have been recorded at both gauges with less than 1 h time difference. In 2007, the total precipitation registered at the two rain gauges amounted to 574 mm (P_{316}) and 562 mm (P_{321}). The calibration period

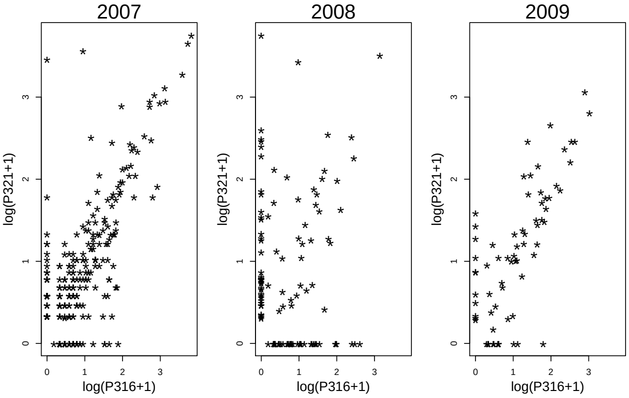

Fig. 3. Rain events measured at each rain gauge on a shifted log scale (1+acc.mm), April–October, 2007–2009.

was characterized by many heavy rain storms (4 events containing 35 mm or more). In the validation year 2008 the rain gauge P_{316} was clearly malfunctioning, recording consistently less precipitation than P_{321} and other rain gauges in the area. The total precipitation for the period amounted to 143 mm at P_{316} compared with 341 mm at P_{321}. The recordings from rain gauge P_{316} in August 2008 were classified with the term "suspicious values" by DMI (2009) but were nevertheless included in the study. The validation year 2008 thus serves as an example of how input errors propagate to model output and affect the model performance. The second validation year, 2009, offered one extreme rain event (>100 mm recorded at P_{316}; >70 mm recorded at P_{321}), and a few medium events (see Fig. 3). The total precipitation amounted to 322 mm (P_{316}) and 302 mm (P_{321}), which again was much less precipitation than during the calibration period in 2007.

3 Uncertainty assessment methodology

3.1 Implementation of GLUE

Prediction limits, or quantiles derived with the GLUE methodology are conditional on the choice of limits of acceptability, the choice of weighting function, the range of models (parameter sets) considered, the exploration of the model space (number of Monte Carlo runs and the method used for sampling the parameter space), the treatment of input and observation errors, and the assumption that the considered system remains unchanged within the validation period. The GLUE steps implemented in this investigation are detailed below.

1. Once a suitable model, M, and relevant input and observations has been selected for the purpose (Sects. 2.1 and 2.2) determine a reasonably broad prior domain for each model parameter θ_i based on the available background knowledge (for details see Sect. 3.2 below).

2. Select an estimation period, N. We used half a year of measurements, April–October, 2007. Carefully check and leave out faulty input data and observations from the estimation (Sect. 2.3). We omitted raw data which were obviously faulty, that is, when the measured velocity or level was zero. This happens occasionally when objects such as toilet paper, etc., clogs/attaches to the flow meter. Of course there might be cases when the gauge is only semi-clogged and hence unreliable measurements are sampled and included for the analysis but such data can be very hard to separate from good data.

3. Chose a likelihood measure $L[M(\boldsymbol{\Theta}|\boldsymbol{u},\boldsymbol{y})]$ to distinguish the behavioural parameter sets $\boldsymbol{\Theta}_B$ from all the parameter sets tried $\boldsymbol{\Theta}$, conditioned on input data $\boldsymbol{u}=(u_1..,u_k,u_{k+1},..,u_N)$ and observations $\boldsymbol{y}=(y_1..,y_k,y_{k+1},..,y_N)$. We used two different likelihood measures. The Nash–Sutcliffe model efficiency coefficient was applied to dry weather periods, L_{dw}, and an exponential likelihood measure, L_{ww}, was applied to wet weather periods, see Eq. (1). The Nash–Sutcliffe likelihood was chosen for the dry weather case because of the desire to fit the dry weather diurnal flow pattern well, whereas an exponential likelihood was chosen to fit the peaks of the hydrographs better (Freer et al., 1996; Beven and Freer, 2001; Thorndahl et al., 2008). The exponential likelihood accentuates the peaks, and weights them higher compared to local minima. The flow peaks are normally an important output in sewer flow modeling to assess surcharge and flooding.

   A flow threshold of 0.15 m^3 s^{-1} distinguishing dry and wet weather periods was determined from inspection of the flow observations. The likelihood measures are defined as

$$\begin{aligned} &L_{\mathrm{dw}} \propto -\frac{\sigma_\epsilon^2}{\sigma_{\mathrm{o}}^2}, \qquad \sigma_{\mathrm{o}}^2 > \sigma_\epsilon^2 \ \ \& \ \ y_k < 0.15 \\ &L_{\mathrm{ww}} \propto e^{-H\left(\frac{\sigma_\epsilon^2}{\sigma_{\mathrm{o}}^2}\right)}, \ y_k > 0.15, \end{aligned} \tag{1}$$

   where σ_ϵ^2 is the residual error variance assuming a zero mean bias, σ_{o}^2 is the observation variance and k is the time index. H is a shaping factor that in this application is fixed to 1. A combined likelihood measure inspired by Choi and Beven (2007) was calculated by multiplication of the dry and wet weather likelihoods:

$$\begin{aligned} &L[M(\boldsymbol{\Theta_i}|\boldsymbol{u},\boldsymbol{y})] \propto \\ &\varpi_1 L_{\mathrm{dw}}[M(\boldsymbol{\Theta}_i|\boldsymbol{u},\boldsymbol{y}_1)]\varpi_2 L_{\mathrm{ww}}[M(\boldsymbol{\Theta}_i|\boldsymbol{u},\boldsymbol{y}_2)], \end{aligned} \tag{2}$$

   where $\boldsymbol{y}_1$ denotes the dry weather observations, $\boldsymbol{y}_2$ denotes the wet weather observations, ϖ_1 and ϖ_2 are weighting coefficients both set to 1, and $\boldsymbol{\Theta}_i$ refers to each parameter set from the prior parameter domain. Equation (2) is effectively a Bayesian updating of likelihoods. The multiplicative form of the overall likelihood was chosen because we wanted to give equal weight to performance in dry and wet weather periods. If we had used one single performance measure for the whole calibration period we would have favored dry weather performance because dry weather periods constitute the majority of the considered calibration period. The more positive the likelihood values the better. Negative likelihood values are not considered because the observed mean in that case would be a better predictor than the model.

4. Select a method and a distribution to draw random parameter sets $\boldsymbol{\Theta}_i$ from. We consistently used uniform (non-informative) prior distributions and Latin Hypercube Monte Carlo sampling (LHS). The disadvantage with LHS is often argued to be the computational burden compared with a Markov chain Monte Carlo approach. A distributed hydrodynamic model would require extensive computational effort, but the lumped conceptual model presented here contains only 10 parameters, and thus the computational burden was not a challenge.

5. Dotty plots as described in Beven (2008) are used to (1) check where in the parameter space the higher likelihoods are located, to (2) check that prior parameter ranges have been chosen adequately broad, and to (3) evaluate parameter correlation. Sometimes it is necessary to adjust the prior domain and restart the Monte Carlo runs a couple of times. This could be necessary if the dotty plots show high likelihood values at the lower or upper end of any of the prior parameter ranges. However parameter ranges might also be constrained by physical considerations.

6. Decide how to extract the behavioural parameters, $\boldsymbol{\Theta}_B$. The procedure to derive the behavioural parameter sets has been either of four: (1) pre-define a likelihood threshold, (2) retain a pre-defined number of behavioural parameter sets, (3) retain a sufficient number of parameter sets to bracket a desired proportion of observations, or (4) use a limit of acceptability approach. In our case we chose the third procedure aiming at a coverage of 90 % of the observations with the 90 % prediction interval generated from a sufficient number of retained parameter sets. In our search for a sufficient number of parameter sets, we calculated prediction intervals for a gradually increasing number of retained parameter sets K based on L, that is

$$K = \dim\{\boldsymbol{\Theta}_B\} = \\ \{100; 500; 1,000; 3,000; 6,000; 10000\}. \tag{3}$$

Ideally, we are satisfied if 90 % of the observations fall inside the generated 90 % prediction interval.

7. The following steps are used to determine the prediction intervals, see also Beven and Freer (2001):

   a. At each time step k rank the ith simulated flow $y^k_{\mathrm{sim},i}$ produced by the retained parameter set $\boldsymbol{\Theta}_{B,i}$ and its associated likelihood $L[M\left(\boldsymbol{\Theta}_{B,i}|\boldsymbol{u},\boldsymbol{y}_{\mathrm{sim},i}\right)]$ value in descending order with respect to flow magnitude.

   b. Rescale the likelihoods to sum to unity $\sum_{i=1}^{K} L[M\left(\boldsymbol{\Theta}_{B,i}\right)] = 1$ where $M\left(\boldsymbol{\Theta}_{B,i}\right)$ denotes the ith behavioural Monte Carlo sample so that at any time step k, prediction quantiles can be formed using

   $$P\left(y^k_{\mathrm{sim},i} < y_{\max}\right) = \\ \sum_{i=1}^{K} L\left[M\left(\boldsymbol{\Theta}_{B,i}|y^k_{\mathrm{sim},i} < y_{\max}\right)\right] \tag{4}$$

   where $y_{\max}$ is some threshold flow.

   c. For the given certainty level β find two quantiles corresponding to $\frac{(1-\beta)}{2}\cdot 100\,\%$ and $\frac{(1+\beta)}{2}\cdot 100\,\%$. These two quantiles are called the lower, $\boldsymbol{y}_l$, and upper, $\boldsymbol{y}_u$, prediction limits. In this study we calculate prediction quantiles for $\beta = 0.90$.

3.2 Choice of prior parameter ranges

The fast runoff from the paved area is defined by the parameters A_f, K_f and α. The choice of a reasonable prior range for A_f was inspired by the calibrated physically distributed hydrodynamic model of the catchment. A_f represents the impermeable runoff area from both combined and separated catchment areas (the latter in case of illicit connections) which was calibrated to 43 ha (Table 1). To be on the safe side the prior of A_f was here allowed to range between 10 and 70 ha. To find a reasonable prior range for the fast runoff concentration time K_f of the system the distributed model was again used. A rain event with a duration of 1 h and with a constant intensity small enough not to exceed the pipe system's flow capacity was imposed on the system at different places in the catchment area, one place at a time, and the resulting hydrographs inspected. On this basis the prior range of K_f was set to 1–8 h. We expected rain gauge P_{316} to contribute most to the runoff because it is closer to the paved combined sewer area than P_{321} (see Fig. 1) but decided to test this assumption by allowing α to range between zero and one. The slow runoff contribution (infiltration inflow) is defined by the parameters A_s, K_s and α. By inspection of the observed hydrographs following rain events we decided a range for the slow runoff concentration time K_s of 8–80 h (0.33–3.33 days), that

Table 4. Choice of prior parameter ranges.

Para- meters	A_f [ha]	A_s [ha]	K_f [h]	K_s [h]	α [–]	a_0 [L s^{-1}]	s_1, s_2 [–]	c_1 [–]	c_2 [–]
	[10;70]	[0;80]	[1;8]	[8;80]	[0;1]	[60;90]	[−0.05;0.03]	[−0.04;0]	[−0.02;0.03]

is, K_s was differentiated from K_f. The area effectively contributing to infiltration inflow, A_s, was allowed to vary between 0 and 80 ha because a considerable amount of unintended water was believed to infiltrate the system. A lower limit of zero was chosen to allow for investigation of possible interactions between the runoff components of the model. A reasonable estimate of the average dry weather flow, a_0, could be derived by inspection of flow measurements in dry weather periods (60–90 L s^{-1}). The lack of physical interpretation of the other wastewater parameters s_1, s_2, c_1, c_2 made it difficult to decide prior ranges and therefore a trial and error approach was conducted before the final ranges displayed in Table 4 were selected.

3.3 Performance measures

Ideally we would like to have narrow prediction limits with a high bracketing of observations. This indicates good model performance and provides confidence in the model when also applied to a validation set. To evaluate this we introduce some performance measures that have been applied in other GLUE studies (Jin et al., 2010; Li et al., 2010; Xiong et al., 2009). The containing ratio (CR) refers to the percentage of observations that fall inside the prediction limits and the average band width (ABW) is the average distance between the lower 5 % and upper 95 % prediction quantile:

$$\mathrm{ABW} = \frac{1}{N}\sum_{k=1}^{N}\left(y_\mathrm{u}^k - y_\mathrm{l}^k\right) \tag{5}$$

where N is the total number of time steps and y_u^k and y_l^k are, respectively, upper and lower prediction quantiles at any given time step, k. Finally the Average Relative Interval Length (ARIL) weights the band width with respect to the observed flow magnitude:

$$\mathrm{ARIL} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{y_\mathrm{u}^k - y_\mathrm{l}^k}{y_k}\right) \tag{6}$$

Note that when we refer to CR in the discussion of results we mean containment within the 90 % prediction limits and when referring to ABW and ARIL these are likewise calculated from 90 % upper and lower prediction limits.

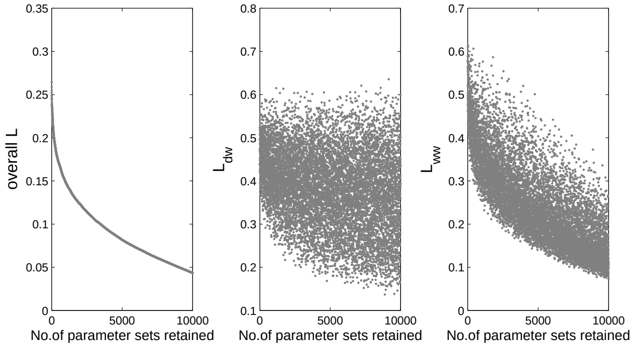

Fig. 4. Likelihood vs. number of retained parameter sets. Shown for overall likelihood (L), dry weather likelihood (L_dw) and wet weather likelihood (L_ww).

4 Results and discussion

4.1 Likelihood measure vs. number of retained parameter sets

Out of 200 000 sampled parameter sets, 18 720 returned positive likelihood values (as defined in Eq. 2). It is noted that we decided to limit the number of behavioural parameter sets to 10 000 although more parameter sets could have been included. Overall peak likelihood was found to 0.2644. Figure 4 shows how the overall likelihood, L, the dry weather likelihood, L_dw, and the wet weather likelihood, L_ww, generally decreases with increasing number of retained parameter sets. Note how both L_dw and L_ww are varying up and down, in the range of 0.2–0.6 for L_dw and 0.1–0.6 for L_ww, as more parameter sets are included, and that the decrease in overall likelihood primarily can be attributed to a decrease in L_ww.

4.2 Dotty plots, correlation structure and posterior parameter sets

Figure 5 shows Dotty plots of the wet weather parameters (upper part) and wastewater parameters (lower part), respectively. Dots are marked according to the number of parameter sets retained, but for clarity reasons we decided to limit the classification of the shown dots to $\dim\{\Theta_B\} = \{500; 3000; 10\,000\}$. Thus, the best 500 parameter sets (with the 500 highest likelihoods) have been coloured black, the best 501–3000 parameter sets dark-grey and the best 3001–10 000 parameter sets are light-grey. White areas reflect the parameter space where the likelihood measure is below that

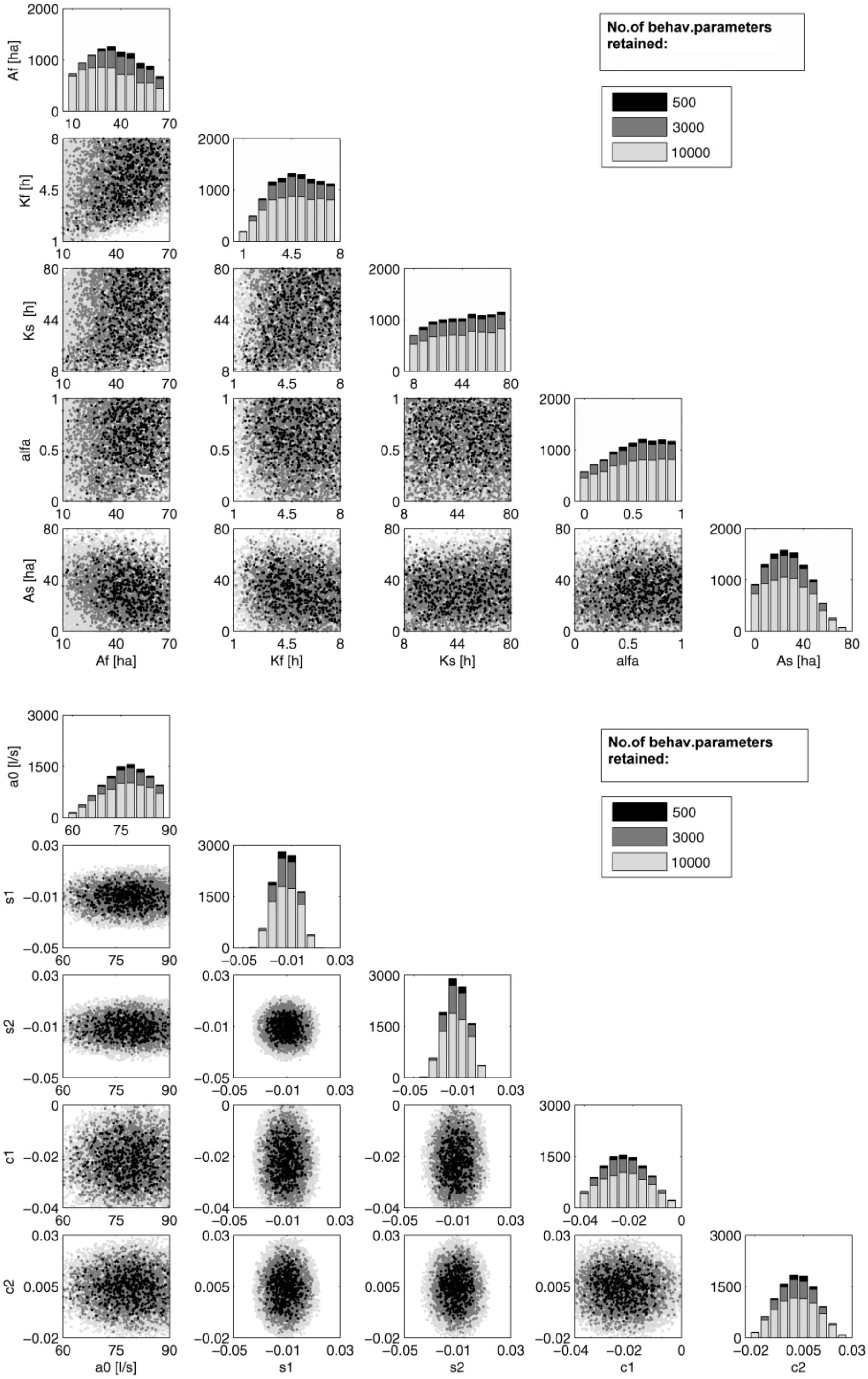

Fig. 5. Dotty plots of wet weather parameters (top) and dry weather parameters (bottom).

of the best 10 000 parameter sets. Histograms have been generated for each parameter and marked in accordance with the number of retained parameter sets.

The histograms for the dry weather model parameters (Fig. 5, bottom) are all quite peaky, showing well-defined posterior ranges and no parameter correlation. However, the histograms for the wet weather parameters (Fig. 5, top) are all more flat and the dotty plots are more scattered, showing less well-defined posterior ranges indicating these parameters are either insensitive to model performance or mutually correlated, or that the prior parameter ranges have been chosen too narrow. The latter is what we observe for K_s, where the prior

Table 5. Minimum and maximum of posterior wet weather parameter ranges for different numbers of retained parameter sets.

Parameter	A_f		K_f		α		A_s		K_s	
sets retained	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$
100	27.4	68.1	2.8	7.6	0.08	0.98	4.1	58.7	18.6	79.4
500	12.0	70.0	1.8	8.0	0.03	1.0	0.8	70.6	8.1	80
1000	12.0	70.0	1.7	8.0	0.00	1.0	0.5	70.6	8.1	80
3000	10.0	70.0	1.1	8.0	0.00	1.0	0.0	75.5	8.0	80
6000	10.0	70.0	1.0	8.0	0.00	1.0	0.0	79.4	8.0	80
10 000	10.0	70.0	1.0	8.0	0.00	1.0	0.0	79.4	8.0	80
Prior	10.0	70.0	1.0	8.0	0.00	1.0	0.0	80	8.0	80

Table 6. Minimum and maximum of posterior dry weather parameter ranges for different numbers of retained parameter sets.

Parameter	a_0		s_1		s_2		c_1		c_2	
sets retained	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$	$\hat{\theta}_\mathrm{min}$	$\hat{\theta}_\mathrm{max}$
100	64.5	86.7	−0.022	−0.002	−0.023	0.000	−0.034	−0.011	−0.007	0.017
500	62.2	89.6	−0.026	0.006	−0.025	0.002	−0.036	−0.006	−0.012	0.020
1000	60.6	89.9	−0.028	0.007	−0.029	0.007	−0.038	−0.004	−0.012	0.023
3000	60.1	90.0	−0.032	0.010	−0.032	0.010	−0.040	0.000	−0.018	0.025
6000	60.1	90.0	−0.034	0.013	−0.037	0.014	−0.040	0.000	−0.020	0.027
10000	60.0	90.0	−0.035	0.014	−0.037	0.014	−0.040	0.000	−0.020	0.029
Prior	60.0	90.0	−0.050	0.030	−0.05	0.030	−0.040	0.000	−0.020	0.030

range perhaps could have been chosen higher. For all the wet weather parameters good model performance (higher likelihood values) can be obtained over the entire prior parameter range with only 500 retained parameter sets, though parameters with higher likelihoods are more commonly found around the peaky areas of the histograms. The wet weather flow contribution seems to be almost equally well represented by either of the rain gauges (see histogram for α); however, the density of darker dots is higher between 0.4 and 1, which means that P_{321} unexpectedly explains most of the runoff despite the location farther away from the paved areas of the catchment that is served by a combined system.

Tables 5 and 6 show minimum and maximum of each posterior parameter range for all investigated numbers of retained parameter sets.

As more parameter sets are included, the posterior parameter range of each parameter widens, and all posterior limits are close to the prior limits allready when 500 parameter sets are retained for the wet weather parameters (see Table 5). Except for a_0, the posterior parameter limits of the wastewater parameters needs more retained parameter sets to approach the prior limits and some of the parameters stays below the prior limits even with 10 000 parameter sets retained. Less peaked histograms and wide posterior parameter ranges are a clear sign of equifinality, that is, many parameter sets can be found that perform almost equally well. Table 7 shows the correlation between the parameters based on the 10 000 best parameters sets. The dry weather parameters are uncorrelated, confirming the pattern observed in Fig. 5 (bottom); however, the largest observed correlation is between a_0 and A_s (−0.43), indicating that a large average wastewater flow compensates for a small slow runoff area and vice versa. The negative correlation between A_f and A_s (−0.15), although rather small, indicates in the same manner that the fast and slow wet weather components of the model "compete" in representing the observed hydrographs, or in other words that the model/observations not clearly allow us to distinguish the fast from the slow runoff components.

4.3 Overall model performance in calibration and validation periods

Figure 6 (left) shows that the overall CR and ARIL increase with the number of retained parameter sets. The overall CR (Fig. 6, left top) increases from approx 58 % to 80 % going from 100 to 10 000 included parameter sets, and the curve flattens out and reaches a steady level below 90 %. It therefore seems unlikely that retainment of more parameter sets would increase the coverage further. Considering the overall CRs to the different number of retained parameter sets K and comparing the calibration year with the validation years only small deviations are observed. This indicates good consistency of the GLUE generated prediction limits between calibration and validation periods. The overall ARIL (Fig. 6, left bottom) increases in the calibration year from 0.38 to around 0.6 when K increases from 100 to 10 000. In the validation

Table 7. Correlation between parameters based on 10 000 retained parameter sets.

	A_f	A_s	K_f	K_s	α	a_0	s_1	s_2	c_1	c_2
A_f	1									
A_s	−0.15	1								
K_f	0.18	−0.09	1							
K_s	0.10	0.11	0.09	1						
α	0.06	0.10	−0.06	0.01	1					
a_0	−0.15	−0.43	−0.06	−0.2	−0.03	1				
s_1	0.00	−0.02	0.01	0.01	0.00	0.00	1			
s_2	−0.02	−0.04	0.00	−0.01	−0.01	0.01	−0.02	1		
c_1	−0.02	−0.04	0.00	−0.01	0.01	0.00	−0.01	0.01	1	
c_2	0.00	0.03	0.00	−0.01	0.00	−0.03	0.00	−0.01	−0.03	1

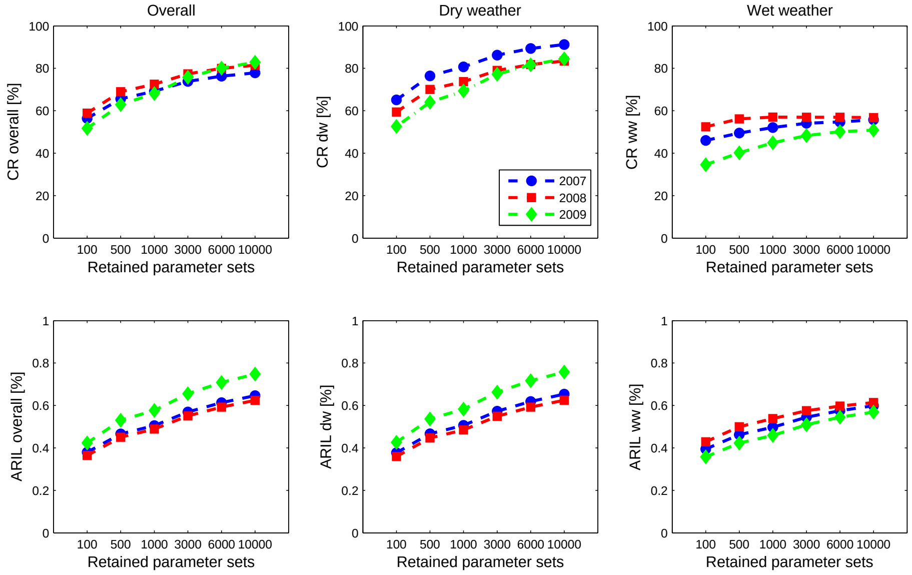

Fig. 6. CR (upper panels) and ARIL (lower panels) vs. the number of retained parameter sets in the calibration year (2007) and the two validation years (2008 and 2009) for the total 6 months period (left panels), the dry weather periods (middle panels) and the wet weather periods (right panels).

year, 2008, similar overall ARIL values are obtained while consistently higher values are found for the validation year 2009 to all K values, indicating that something may have changed in the system. When considering dry weather periods only (Fig. 6, middle top) it was shown possible to cover the desired 90 % (91.2 % exactly) of the observations during the calibration period by retaining 10 000 behavioural parameter sets. Note how the difference between the dry weather CR curves in the validation years decrease as the number of parameter sets approaches 10 000. However, the dry weather CRs of the validation years are consistently lower reaching a maximum of 80 % with 10 000 parameter sets retained. This inconsistency is unexpected because changes in the dry weather flow level or flow pattern normally occur due to changes in population size or in water consumption pattern, which could not be confirmed for the studied period. Other explanations could be changes in measurement conditions like calibration of the flow meter, flow meter placement in the pipe, or infiltration inflow occurring at a timescale larger than that can be accounted for with this model. Does the observed inconsistency suggest an inability of the GLUE methodology to fully describe the uncertainty of the system? We will take a closer look into this by considering selected hydrographs in Sect. 4.5 and conclude on this question in Sect. 4.8. The maximum dry weather ARIL (Fig. 6, middle bottom) was found to 0.65 for the calibration year 2007. A similar pattern was found for the validation year 2008 but not for 2009, which had consistently higher ARIL values and lower coverage, similar to what was found for the overall ARIL.

When considering the wet weather periods only, CR is generally lower than for the dry weather periods and for the simulated periods as a whole (Fig. 6, right top). The CR curves flatten out already after 1–3000 retained parameter sets at a level of just above 50 % for the calibration year,

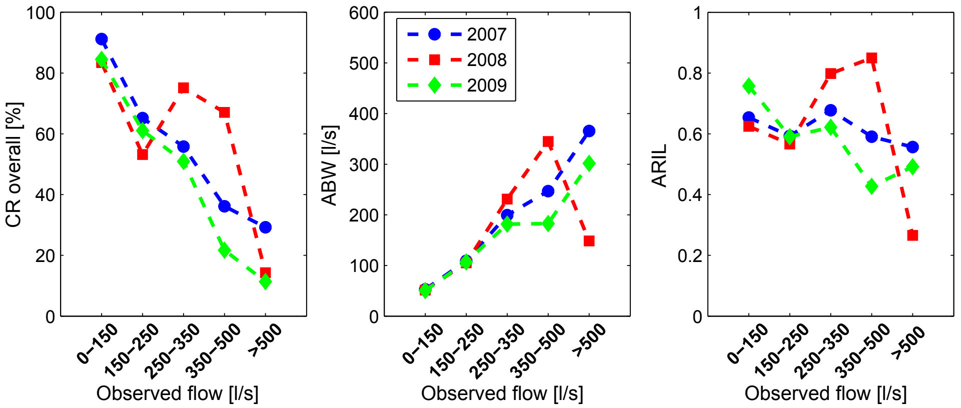

Fig. 7. CR, ABW and ARIL (calculated from 10 000 retained parameter sets) vs. observed flow magnitude for the calibration year (2007) and the validation years (2008 and 2009).

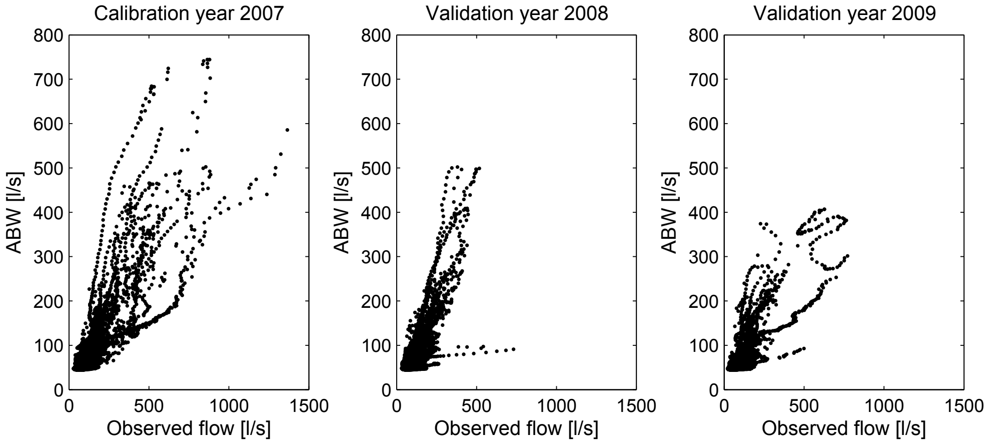

Fig. 8. Variation in band width versus observed flow magnitude. Band width is calculated from 10 000 behavioural parameter sets.

and 55 and 50 % for the validation years. This poor coverage may be caused by a misfit between the recorded rainfall and the measured runoff for heavy convective rain events with limited spatial extent, where the two rain gauges do not well represent the effective rainfall over the catchment due to their locations several kilometer away. Wider prior parameter ranges could perhaps have increased the coverage. Note also from this plot how the consistency between calibration and validation years increase as the number of retained parameter sets is increased. The ARIL (Fig. 6, right bottom) increases to almost 0.6 with 10 000 retained parameter sets in both calibration and validation periods, which is close to the value obtained overall and in dry weather periods alone. The wet weather ARIL values are quite similar between calibration and validation periods.

4.4 Dependency of flow magnitude

Figure 7 shows how the performance measures CR, ABW and ARIL change with the flow magnitude using prediction limits generated from 10 000 parameter sets. Generally, the ABW (middle panel) increases proportionally with the flow, but the ability of the prediction limits to bracket the observations decreases with the flow magnitude (left panel). In the calibration year the CR drops from 90 % in dry weather to just 30 % for flows above 500 L s^{-1}, supporting the suggestion above about the influence of heavy convective rain events, and although the ABW (middle panel) increases from approx 50 L s^{-1} in dry weather to 380 L s^{-1} for flows higher than 500 L s^{-1} this is not enough to encompass the desired percentage of observations. Again a wider prior parameter space could probably increase CR but a likelihood measure

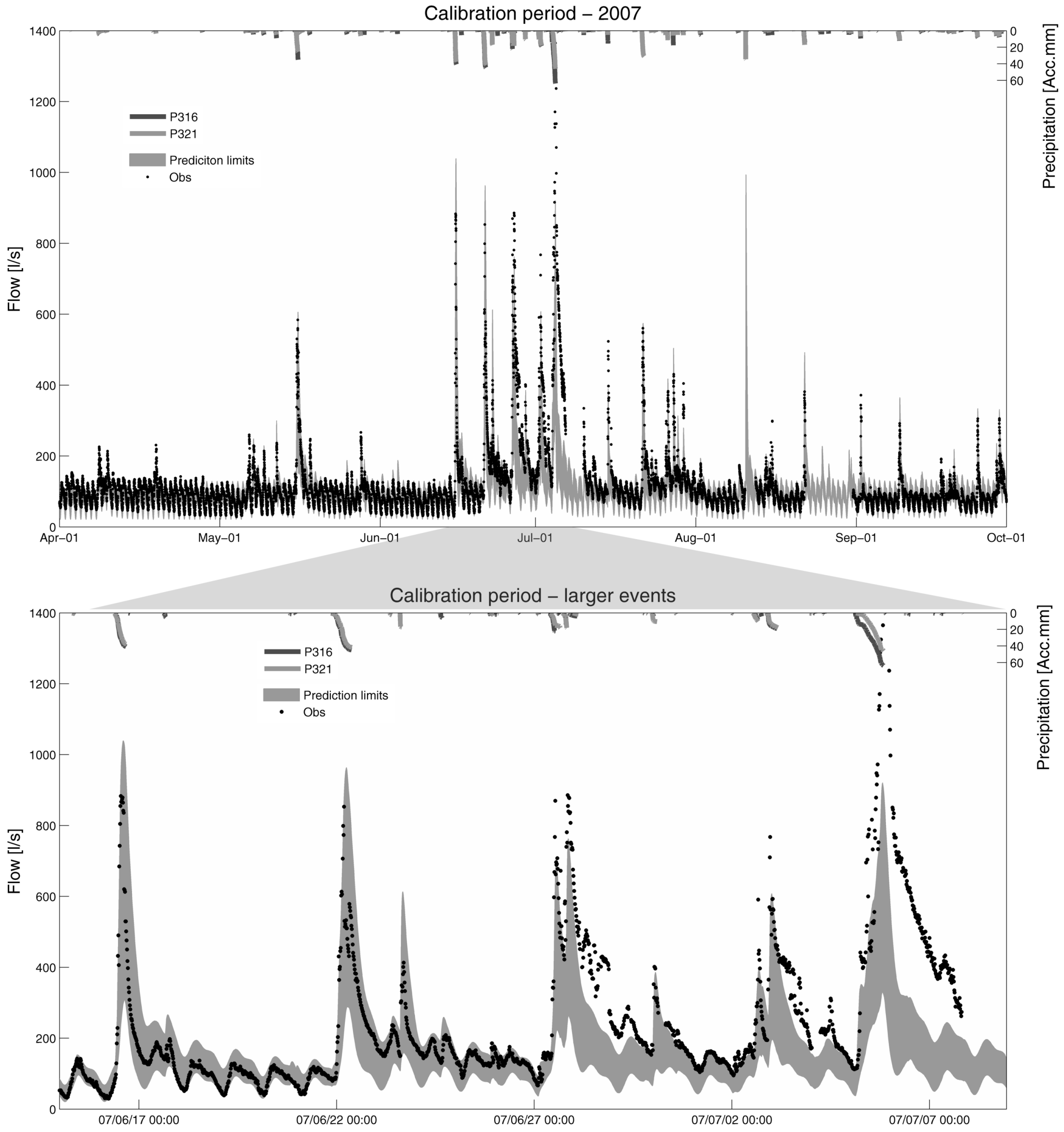

Fig. 9. Rainfall input, flow observations and 90 % flow prediction limits generated from 10 000 parameter sets. Whole calibration period (top) and enlargement for a period with wet weather flow conditions (bottom). Periods without flow observations were discarded from the analysis.

that favours enclosure of the largest events would also increase CR at higher flow rates.

Interestingly, the ARIL (right panel) is rather constant in the calibration year 2007, that is, the uncertainty of flow predictions with the model used here is almost proportional to the flow magnitude. The validation years show some deviations from the calibration year, which may be attributed to the small sample sizes used to compute the performance measures for especially the larger flow intervals, as well as differences in precipitation recorded at the two gauges and artifacts associated with individual rain events.

Whereas Fig. 7 (middle) shows the average band width (ABW) for different flow intervals, calculated as an average for all rain events in each year, Fig. 8 illustrates for each year how the band width evolves from time step to time step. It is seen that the average values actually cover up some large fluctuations in modelled band width. The “traces” of connected data illustrate how the band width evolves during individual rain events, how the band width generally increases with flow magnitude (corresponding to what is seen in Fig. 7, middle), and how less precipitation in 2008 and 2009 than in 2007 lead to smaller flows and band widths.

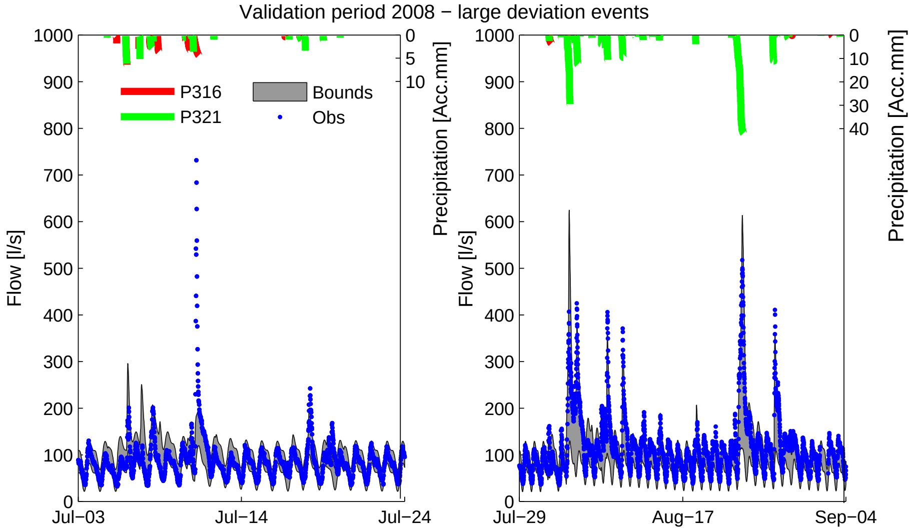

Fig. 10. Rainfall input, flow observations and 90 % flow prediction limits generated from 10 000 parameter sets for selected periods in the validation year 2008.

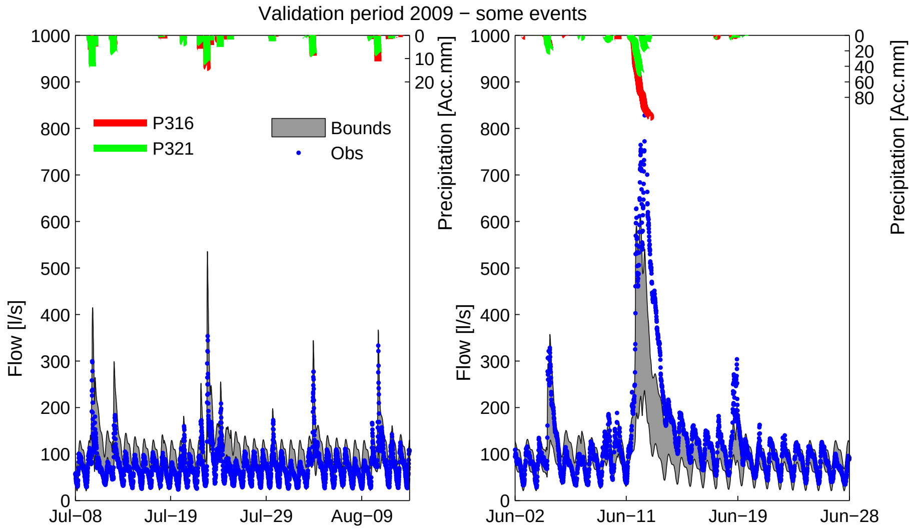

Fig. 11. Rainfall input, flow observations and 90 % flow prediction limits generated from 10 000 parameter sets for selected periods in the validation year 2009.

4.5 Analysis of hydrographs from calibration and validation periods

Figure 9 shows the rainfall input (accumulated rainfall per event for each rain gauge), flow observations and generated 90 % prediction limits for the whole calibration period (top panel) and an enlargement of a period with the largest events recorded (bottom panel).

The dry weather observations (flows of less than 150 L s^{-1}) are generally well covered by the prediction limits, which was also concluded from the performance measures (Figs. 6 and 7), but they seem to be close to the upper prediction limit in April and to the lower prediction limit in October, indicating that the mean dry weather flow declines gradually during the period. Accounting for this trend in the dry weather model could perhaps have resulted in smaller ABW and higher CR for dry weather periods.

The wet weather flows (flows higher than 150 L s^{-1}) are well covered for some events, for example, the events shown in the first half of the lower panel of Fig. 9 where 35–40 mm rainfall was recorded, but for the remaining events shown the observed peaks are higher than the upper prediction limit, the hydrograph tails are longer than the model suggests and the flow furthermore fluctuates in a way that cannot be described with the model used. The fast time constant K_f as well as the impermeable area A_f (or perhaps also A_s and K_s)

needs to be much larger for the prediction intervals to cover the last event shown (lower panel). This event as well as the other events shown explains why neither the Dotty plots nor the histograms in Fig. 5 (top) were able to clearly identify a higher likelihood area for these parameters. There is also the possibility of backwater effects in the system which are not dealt with in the model and this could perhaps explain the long tail of the last flow hydrograph seen in Fig. 9 (lower panel), but it cannot be excluded that the flow measurements are erroneous, or that the measured rainfall is non-representative (the two gauges measured about 50 and 65 mm rainfall, i.e. a convective rainfall pattern with large spatial variation is likely).

Figure 10 shows the rainfall input, flow observations and generated 90 % prediction limits for selected periods in the validation year 2008, where rain gauge P_{316} was malfunctioning for a longer period. The smallest ABW and ARIL for 2008 (Fig. 7, middle and right) occurs for the highest observed flow category ($> 500\,\mathrm{L\,s^{-1}}$), which is due to the high flow observations on 11th July where only 5 mm rainfall was recorded at the two gauges (Fig. 10, left), which is also visible as the isolated, flat "trace" on Fig. 8 (middle). In this case a large convective rainfall event with limited spatial extent may have passed over the catchment without significantly affecting the rain recordings, or the flow observations are erroneous. Figure 10 (right) shows several significant flow events in August 2008 where gauge P_{316} did not record any rainfall at all, probably due to technical malfunctioning, and this causes the flow predictions to be underestimated (the flow observations are consistently close to, or above the upper prediction limit for all the illustrated rain events).

Figure 11 shows the rainfall input, flow observations and generated 90 % prediction limits in the second validation year 2009 for a selected period where both the dry and wet weather flows were well covered by the prediction limits (left) and for a period where the largest event in 2009 occurred (right). In this latter case the gauges recorded very different rainfall amounts (50 and 100 mm), and the model underestimated the peak, the timing and the tailing of the observed hydrograph, which explains the S-shaped "trace" visible in Fig. 8 (right). Note also from the left figure that the flow observations in dry weather are very low and close to the lower bound which is general for 2009. The lower dry weather flow in 2009 explains the higher ARIL values obtained in dry weather periods of 2009 that were observed in Fig. 6.

4.6 Interpretation of posterior parameter ranges

In Sect. 4.2 we saw that posterior ranges approached the priors for many of the wet weather parameters retaining just 500 parameter sets. With the large uncertainties that originate from inadequate rain inputs (spatial heterogeneity not represented by two rain gauges), as well as flow measurement errors and possible model structure inadequacies discussed above, it is hardly surprising that posterior parameter ranges become so wide and dotty plots look so scattered. It is important to recognize that each parameter set carries along with it an implicit (non-stationary) error series and that models (parameter sets) that underpredict in calibration are expected to underpredict in similar circumstances in prediction etc. The uncertainties are not being transferred to the model parameters, but the error series are (implicitly) weighted along with the simulated outputs from each model. The posterior parameters lack physical interpretation because of parameter compensation and thus cannot be used, for example for inference about the relative size of infiltration area versus size of paved area, which otherwise would be desired knowledge. Such parameter compensations will be apparent in any calibration exercise unless prior knowledge about what is acceptable or not acceptable for parameters and their interactions can be specified.

4.7 Consistency in model-failure?

Due to the rather consistent coverage of the observations between different periods for different behavioural thresholds it might be worth investigating if some form of non-stationary error correction can improve the results. This could be done for example by applying a transform or bias correction to the results. One way to check the model for non-stationary errors is to subdivide the whole range of observed runoff into an appropriate number of smaller intervals (windows), for example, 0–50 $\mathrm{L\,s^{-1}}$, 50–75 $\mathrm{L\,s^{-1}}$, etc., and then calculate the residuals between simulated and observed runoffs in each window. Figure 12 shows such a residual histogram with 10 bins. Clearly the residuals are not randomly distributed around zero for any of the windows. In each window the need for bias correction is different which is understood from the appearance of the histograms and the mean of the residuals μ. A negative μ means that the model underestimates the flows and vice versa. For dry weather flows, that is, flows less than $150\,\mathrm{L\,s^{-1}}$, the model performs reasonably well, but for larger flows the model tends to underestimate the runoff. This suggest that some form of non-stationary error correction might be possible, for example, as outlined in Xiong and O'Connor (2008). Such a bias correction will probably improve the calibration results but generally there are more bins with positive outliers than negative which indicates that a bias correction is not trivial. A bias correction implementation is however beyond the scope of this paper.

4.8 Epistemic uncertainties

The evidence from the changing nature of the errors in this study between and within periods (epistemic uncertainties) suggests that it is very difficult to test GLUEs ability to provide uncertainty bounds that bracket observations in both calibration and validation, which would also be the case if a formal likelihood approach had been applied.

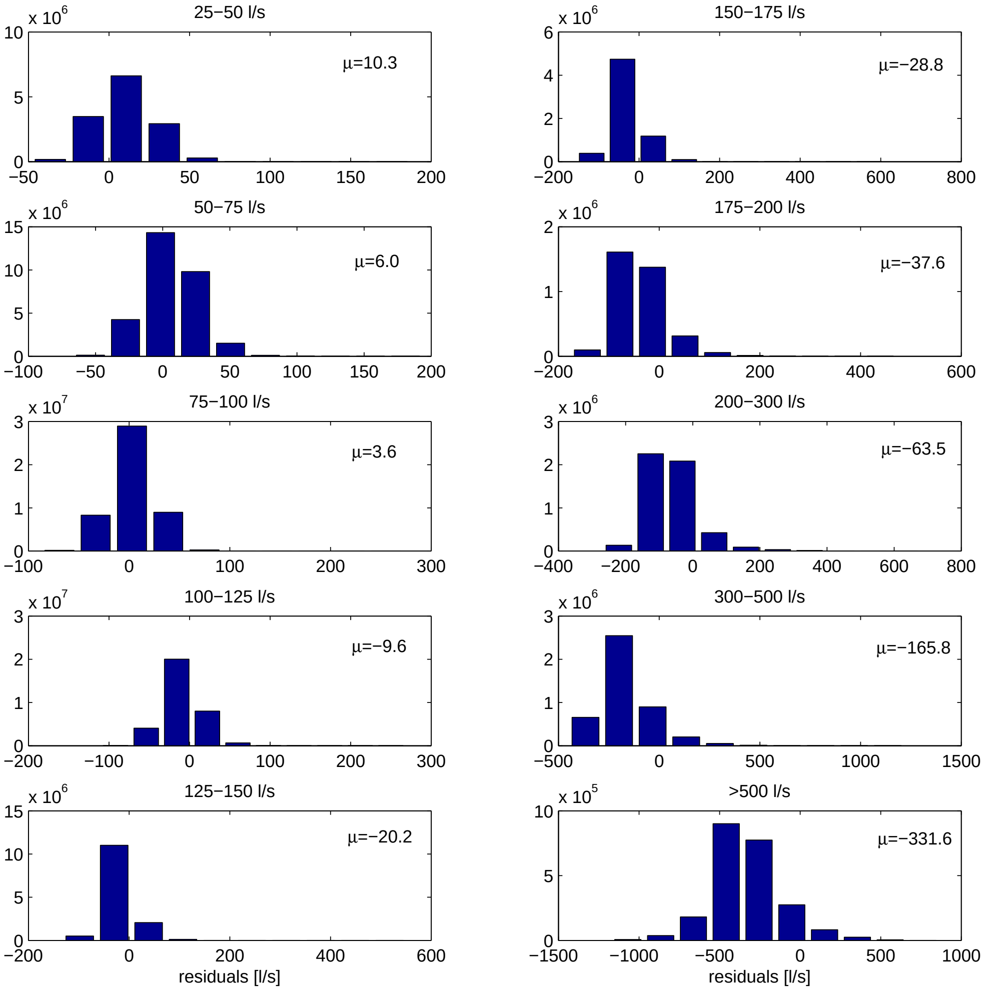

Fig. 12. Histogram and mean of residuals μ for each window. The observed runoffs were subdivided into 10 intervals (windows) and the residuals calculated for all 10 000 simulations.

The experiences from this investigation suggest that calibration of much more complex models (physically distributed, hydrodynamic) used in practical urban drainage engineering in catchments with insufficient rain gauge coverage to questionable flow measurements from shorter measuring campaigns is problematic not least because a calibrated model normally implies a reduction in the safety factor used in modelling of urban drainage systems in Denmark (Hansen et al., 2005).

5 Conclusions

In this study a simple conceptual hydrological model has been applied to simulate flow in a sewer system, that receives water from both combined and separated catchments. The GLUE methodology was applied to assess the uncertainty on flow simulation and parameter estimation. To be able to derive the behavioural parameters, a combined likelihood measure was formulated. For the dry weather flow periods the Nash–Sutcliffe model efficiency coefficient was used, whereas an exponential likelihood measure, that has the property of fitting the peaks better, was used for the wet weather periods. Instead of preselecting the number of behavioural parameter sets, it was decided to retain an increasing proportion of parameter sets (100; 500; 1000; 3000; 6000; 10 000), ideally until the GLUE generated 90 % prediction limits encompassed 90 % of the observations. However, as the overall CR curve was shown to be flattening out at 10 000 retained parameter sets, this number was decided a sufficient maximum number to include. The overall CR increased from approx 58 % to 80 %, as the proportion of behavioural parameter sets included increased from 100 to 10 000 and hence it was not possible to obtain the desired coverage. Considering dry weather periods separately, the prediction limits generated from 10,000 parameter sets enclosed a little more than 90%, while in wet weather periods on average only around 55 % was enclosed. Furthermore, the proportion of observations enclosed decreased with increasing flow magnitude, despite that the prediction limits expanded proportionally with the flow.

Two subsequent half-year summer periods were included for validation to check the consistency of the GLUE generated prediction limits. It was concluded that overall the obtained CRs in the validation periods were similar to that obtained in calibration for all the considered retained proportions of parameter sets, and thus good consistency was found. However, when looking separately at dry weather and wet weather periods, as well as at different flow levels, several inconsistencies were observed between calibration and validation periods. These inconsistencies could in dry weather presumably be attributed to changes in measurement conditions, and in wet weather attributed to inadequate rain input coverage, unreliable flow meter measurements, and/or model deficiencies (e.g. backwater effects not accounted for), etc. Retaining just 500 parameter sets meant that the wet weather posterior parameter ranges approached those of the priors, which is a clear sign of equifinality. Hence the obtained posterior parameter ranges cannot be used for interpretation of, for example, the size of contributing paved area vs. size of slow infiltration-inflow area. The posterior wastewater parameter limits were generally more well determined.

The observed inconsistencies between calibration and validation periods indicated by CR and ARIL would most likely also have been observed in the case a formal approach had been chosen, simply because events such as a sudden lower dry weather flow or malfunctioning rain gauges in a validation period are unexpected events (epistemic events) and cannot be predicted from a set of calibration data. Hence we cannot reject the GLUE methodology as a tool for uncertainty analysis on the basis of this study. The evidence from the changing nature of the errors in this study between and within periods suggests that it might be very difficult to find a valid error model for use in a formal likelihood approach, and that we should therefore try to learn from the significant discrepancies between model and observations. One way to do so could be to use some non-stationary error correction procedure to improve the predictive capability. This could even be applied to areas of wet and dry periods. So if the model was consistently under predicting for “types” of periods and events then this could be accounted for and then reasons why this type of correction improves predictions could be analysed and discussed. This was however beyond the scope of this paper.

In practical urban drainage engineering applications, it is not uncommon that large hydrodynamic models with many more parameters are calibrated to flow data, collected from measuring campaigns of shorter duration than used here, with equally poor rain input representation. Bearing in mind that these models are indispensable tools in redesign and upgrade proposals, and sometimes used for flow forecasting, it seems crucial from this study to (1) obtain more representative rain inputs (perhaps by radars), (2) use more reliable flow meters and (3) replace measuring campaigns with online monitoring to secure a higher coherence between model simulations and observations.

Acknowledgements. We wish to thank Jacob Nørremark, Spildevandscenter Avedøere I/S for support with background information and flow meter data, and Lisbeth Brusendorff, DTU Environment for help with the graphical layout. We also thank Jim Freer for fruitful discussions leading to the improvement of the analysis conducted in the manuscript. This research was supported by a PhD fellowship co-founded by Krüger A/S, DTU Environment and the Ministry of Science, Technology and Innovation through the Urban Water Technology graduate school (www.urbanwatertech.dk), and by the Danish Council for Strategic Research (Storm and Wastewater Informatics Project).

Edited by: J. Freer

References

Aronica, G., Freni, G., and Oliveri, E.: Uncertainty analysis of the influence of rainfall time resolution in the modelling of urban drainage systems, Hydrol. Process., 19, 1055–1071, doi:10.1002/hyp.5645, 2005.

Barbera, P. L., Lanza, L., and Stagi, L.: Tipping bucket mechanical errors and their influence on rainfall statistics and extremes, Water Sci. Technol., 45, 1–9, 2002.

Bertrand-Krajewski, J.-L., Bardin, J.-P., Mourad, M., and Béranger, Y.: Accounting for sensor calibration, data validation, measurement and sampling uncertainties in monitoring urban drainage systems, Water Sci. Technol., 47, 95–102, 2003.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18–36, 2006.

Beven, K.: Environmental Modelling: An Uncertain Future?, Routledge, London, UK, available at: http://www.uncertain-future.org.uk/ (last access: 20 June 2010), 2008.

Beven, K. and Binley, A.: The future of distributed models – model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298, 1992.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11–29, 2001.

Beven, K. J., Smith, P. J., and Freer, J. E.: So just why would a modeller choose to be incoherent?, J. Hydrol., 354, 15–32, 2008.

Beven, K., Smith, P. J., and Wood, A.: On the colour and spin of epistemic error (and what we might do about it), Hydrol. Earth Syst. Sci., 15, 3123–3133, doi:10.5194/hess-15-3123-2011, 2011.

Blazkova, S. and Beven, K.: Uncertainty in flood estimation (vol 5, pg 325, 2009), Struct. Infrastruct. E., 5, 437–437, doi:10.1080/15732470903064614, 2009a.

Blazkova, S. and Beven, K.: A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, Water Resour. Res., 45, W00B16, doi:10.1029/2007WR006726, 2009b.

Choi, H. T. and Beven, K.: Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, J. Hydrol., 332, 316–336, 2007.

DHI: MOUSE RDII Reference Manual, Tech. rep., Danish Hydraulic Institute, Hørsholm, Denmark., 2009.

DMI: Operation of the rain gauge system on behalf of The Water Pollution Committee of The Society of Danish Engineers (in Danish), technical report of 2008, Danish Meteorological Institute, 2009.

Dotto, C. B. S., Deletic, A., and Fletcher, T. D.: Analysis of parameter uncertainty of a flow and quality stormwater model, Water Sci. Technol., 60, 717–725, doi:10.2166/wst.2009.434, 2009.

Dotto, C. B. S., Kleidorfer, M., Deletic, A., Fletcher, T. D., McCarthy, D. T., and Rauch, W.: Stormwater quality models: performance and sensitivity analysis, Water Sci. Technol., 62, 837–843, doi:10.2166/wst.2010.325, 2010.

Dotto, C. B., 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 Res., 46, 2545–2558, doi:10.1016/j.watres.2012.02.009, 2012.

Freer, J., Beven, K., and Ambroise, B.: Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, Water Resour. Res., 32, 2161–2173, 1996.

Freni, G. and Mannina, G.: Bayesian approach for uncertainty quantification in water quality modelling: The influence of prior distribution, J. Hydrol., 392, 31–39, doi:10.1016/j.jhydrol.2010.07.043, 2010.

Freni, G., Mannina, G., and Viviani, G.: Uncertainty in urban stormwater quality modelling: The effect of acceptability threshold in the GLUE methodology, Water Res., 42, 2061–2072, 2008.

Freni, G., Mannina, G., and Viviani, G.: Urban runoff modelling uncertainty: Comparison among Bayesian and pseudo-Bayesian methods, Environ. Modell. Softw., 24, 1100–1111, 2009a.

Freni, G., Mannina, G., and Viviani, G.: Uncertainty in urban stormwater quality modelling: The influence of likelihood measure formulation in the GLUE methodology, Sci. Total Environ., 408, 138–145, 2009b.

Hansen, A., Liu, L., Linde, J. J., Mark, O., and Mikkelsen, P. S.: Accounting for uncertainty in urban drainage system performance assessment using safety factors applied to runoff, 10th International Conference on Urban Drainage, 21–26 August 2005, Copenhagen/Denmark, 10ICUD, 2005.

Jin, X., Xu, C.-Y., Zhang, Q., and Singh, V.: Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model, J. Hydrol. – Amsterdam, 383, 147–155, 2010.

Jørgensen, H. K., Rosenørn, S., Madsen, H., and Mikkelsen, P. S.: Quality control of rain data used for urban runoff systems, Water Sci. Technol., 37, 113–120, 1998.

Juston, J., Andrén, O., Kätterer, T., and Jansson, P.-E.: Uncertainty analyses for calibrating a soil carbon balance model to agricultural field trial data in Sweden and Kenya, Ecol. Modell., 221, 1880–1888, 2010.

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

Li, L., Xia, J., Xu, C.-Y., and Singh, V.: Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models, J. Hydrol. – Amsterdam, 390, 210, 2010.

Lindblom, E., Madsen, H., and Mikkelsen, P. S.: Comparative uncertainty analysis of copper loads in stormwater systems using GLUE and grey-box modeling, Water Sci. Technol., 56, 11–18, 2007.

Lindblom, E., Ahlman, S., and Mikkelsen, P.: Uncertainty-based calibration and prediction with a stormwater surface accumulation-washoff model based on coverage of sampled Zn, Cu, Pb and Cd field data, Water Res., 45, 3823–3835, doi:10.1016/j.watres.2011.04.033, 2011.

Mannina, G. and Viviani, G.: An urban drainage stormwater quality model: Model development and uncertainty quantification, J. Hydrol. – Amsterdam, 381, 248–265, 2010.

Mantovan, P. and Todini, E.: Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, J. Hydrol., 330, 368–381, 2006.

Mantovan, P., Todini, E., and Martina, M. L. V.: Reply to comment by Keith Beven, Paul Smith and Jim Freer on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology", J. Hydrol. – Amsterdam, 338, 319, 2007.

Mitchell, S., Beven, K., and Freer, J.: Multiple sources of predictive uncertainty in modeled estimates of net ecosystem CO_2 exchange, Ecol. Modell., 220, 3259–3270, 2009.

Molini, A. and Barbera, L. G. L. P. L.: The impact of tipping-bucket raingauge measurement errors on design rainfall for urban-scale applications, Hydrol. Process., 19, 1073–1088, doi:10.1002/hyp.5646, 2005.

Pedersen, L., Jensen, N. E., Christensen, L. E., and Madsen, H.: Quantification of the spatial variability of rainfall based on a dense network of rain gauges, Atmos. Res., 95, 441–454, 2010.

Piñol, J., Espadaler, X., Pérez, N., and Beven, K.: Testing a new model of aphid abundance with sedentary and non-sedentary predators, Ecol. Modell., 220, 2469–2480, 2009.

Shedekar, V. S., King, K. W., Brown, L. C., Fausey, N. R., Heckel, M., and Harmel, R. D.: Measurement Errors in Tipping Bucket Rain Gauges under Different Rainfall Intensities and their implication to Hydrologic Models, Conf.paper, ASABE Annual International Meeting, 21–24 June, 1–9, 2009.

Staudt, K., Falge, E., Pyles, R. D., Paw U, K. T., and Foken, T.: Sensitivity and predictive uncertainty of the ACASA model at a spruce forest site, Biogeosciences, 7, 3685–3705, doi:10.5194/bg-7-3685-2010, 2010.

Stedinger, J. R., Vogel, R. M., Lee, S. U., and Batchelder, R.: Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, Water Resour. Res., 44, 1–17, 2008.

Thorndahl, S., Beven, K., Jensen, J., and Schaarup-Jensen, K.: Event based uncertainty assessment in urban drainage modelling, applying the GLUE methodology, J. Hydrol., 357, 421–437, 2008.

Vaes, G., Willems, P., and Berlamont, J.: Areal rainfall correction coefficients for small urban catchments, Atmos. Res., 77, 48–59, doi:10.1016/j.atmosres.2004.10.015, 2005.

Willems, P.: Stochastic description of the rainfall input errors in lumped hydrological models, Stoch. Environm. Res. Risk As., 15, 132–152, doi:10.1007/s004770000063, 2001.

Xiong, L. and O'Connor, K. M.: An empirical method to improve the prediction limits of the GLUE methodology in rainfall-runoff modeling, J. Hydrol., 349, 115–124, doi:10.1016/j.jhydrol.2007.10.029, 2008.

Xiong, L., Wan, M., Wei, X., and O'Connor, K. M.: Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation, Hydrol. Sci. J., 54, 852–871, 2009.