Identifying fit-for-purpose lumped surrogate models for large urban drainage systems using GLUE

Thrysøe, Cecilie; Arnbjerg-Nielsen, Karsten; Borup, Morten

Published in:
Journal of Hydrology

Link to article, DOI:
10.1016/j.jhydrol.2018.11.005

Publication date:
2019

Document Version
Peer reviewed version

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Identifying fit-for-purpose lumped surrogate models for large urban drainage systems using GLUE

Cecilie Thrysøe*, Karsten Arnbjerg-Nielsen, and Morten Borup

Urban Water Systems Section, Department of Environmental Engineering (DTU Environment), Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

*Corresponding author (C. Thrysøe) cethry@env.dtu.dk. Other email addresses: (M. Borup) morb@env.dtu.dk, (K. Arnbjerg-Nielsen) karn@env.dtu.dk

Authors declare no conflict of interest

Revised and resubmitted to Journal of Hydrology

Abstract

Distributed physically based models (DPMs) have become the standard tool for urban drainage modelling. However, high computational demands limit these in applications where fast or multiple simulations are needed. This paper presents simple fit-for-purpose cheaper-to-run surrogate models (SMs) for pipe network simulations which are validated against a DPM. The SMs are set up by lumping the DPM network into compartments in which the volume of water is governed by mass balances. Outgoing discharges to downstream compartment(s) and surcharging are computed from unambiguous volume-discharge curves. The SMs are applied on a 45 km² catchment, Elster Creek in Melbourne, Australia. The number of simulated states and simulation times are reduced by approximately 3 and 6 orders of magnitude, respectively. Different SM complexities are examined. The simplest SM using steady state training data performed well with NSE of 0.98 for volume in the most upstream compartment. When emulating the aggregated surcharge from that compartment, the SM captured all surcharge events correctly. NSE improved from 0.35 to 0.84 when subdividing the compartment into 17 subcompartments. Uncertainty of SM parameters was examined using the Generalized Likelihood Uncertainty Estimation (GLUE) methodology. Two different sampling methods were applied. Limits of acceptability for real-time control, warning and planning,
resulted in many accepted models upstream and few to none in downstream backwater-prone areas. All applications showed SM uncertainty bands within expected uncertainty bands for DPM, supporting the use of a simpler conceptual model in fit-for-purpose modelling in urban water systems when computational demands of DPMs are prohibitive.

Keywords: Hydraulic modelling, surrogate models, urban drainage, computational time, uncertainty, GLUE

1 Introduction

Projections of climate change show an increase in frequency and magnitude of extreme rainfall. Coupled with rapid urbanization, this will lead to increased loading of urban storm- and wastewater systems, thereby increasing the risk of pluvial flooding and combined sewer overflows (IPCC, 2012; Semadeni-Davies et al., 2008). To mitigate this risk, adaptation measures can be implemented by changing the infrastructure, in the form of, e.g. bigger pipes, installation of basins, including local infiltration measures etc. (Zhou et al., 2012). Designing adaptation measures requires models, typically in the form of distributed physically-based models (DPMs). Dynamic adaptation using Model Predictive Control to optimize the use of the existing systems storage capacity can also be applied (Lund et al., 2018). Real-time applications require models that in seconds or minutes can produce forecasts of selected hydraulic states several hours into the future, which is not feaseable with DPMs, and the usage of DPMs for scenario development and design optimization is also impeded due to the high computational time of the models (Henonin et al., 2013). As reported in (Leitão et al., 2010), the computational time for this model type is typically longer than real time, making them unfit for real-time modelling, questionable for warning purposes, and also prohibitive when exploring many planning scenarios, although a few examples are available (Löwe et al., 2017; Rauch et al., 2017). Hence many applications need a fast and numerically stable model with sufficient accuracy at key locations in the catchment, and even though the development in computers computational speed continues, this will not provide the orders of magnitude speed-up required in the foreseeable future. This
applies for planning purposes in at least two typical situations: 1) where city-scale planning purposes often focuses on certain parts of the city while other parts are less interesting and hence DPM results are never needed nor interpreted for many parts of the system, and 2) where calculation of pollution loads can be done with sufficient accuracy using simplified hydraulic models (Vaes and Berlamont 1998, Johansen et al., 1984). It also applies to forecast and control situations, where the uncertainty of the forecasts makes the use of DPM models obsolete because the improved calculation accuracy does not lead to improved results (Lund et al 2018).

Because many applications of urban flood models require multiple fast simulations, several attempts have been made to reduce the computational time of the DPMs. One approach is to enhance computational performance by optimising the usage of computational resources through e.g. parallelisation (Leandro et al., 2014; Neal et al., 2010), but this requires that sufficient computational resources are available. Another approach is to reduce computational costs by means of simplification. Simplified models will, however, not be able to emulate the entire model space of the original models. Model errors are thus introduced when going from the original High-Fidelity (HiFi) model to a simpler model. Choosing the level of simplification is a trade-off between accuracy, acceptance of errors and computational time. This decision should be based on the type of process simulated and the objective of the problem to define the level of details needed in the required model output, the acceptable error margin and computational time allowed also known as fit-for-purpose modelling (Beven, 2015; Blazkova and Beven, 2009; Haasnoot et al., 2014; Jamali et al., 2018; Wright and Esward, 2013). Furthermore, the acceptable level of error introduced during a model simplification process should be seen in relation to the typically rather large uncertainties in the model input, such as rainfall data, and the errors between reality and the original model.

The methods of going from a HiFi model to a LoFi model can be divided into two main categories: LoFi DPMs and conceptual models. LoFi DPMs are also physically-based but less detailed compared to the original models (Razavi et al., 2012). Methods for going from a HiFi DPM to a LoFi DPM vary through
Some methods reduce the numerical accuracy, i.e. use a coarser spatial/temporal grid size (Fewtrell et al., 2011) or apply steady state (Jamali et al., 2018). Other reduce the number of physical processes in the model e.g. by reducing the Saint-Venant equations to either the diffusive or kinematic wave formulation (Fewtrell et al., 2011; Neal et al., 2010; Neelz, S. & Pender, 2013). Another approach is to simplify the system itself via pruning, trimming and merging rather than the model formulation. Davidsen et al. (2017) and Leitão et al. (2010) applied a combination of the above and conducted a 35 and 42 % reduction of the network elements resulting in a 70 and up to 61 % reduction in computation time, respectively. This method may, however, not be sufficient for a large number of applications, where a speed-up of several orders of magnitude is needed.

Another branch of model simplification is to use a cheaper-to-run conceptual model (Razavi et al., 2012; Ye et al., 2014). Results from the original model or measurements can be used to calibrate the conceptual model and in some cases validate it if no measured data are available. This is a data-driven approach and therefore the model does not need to consider any physical processes. These models are denoted surrogate models (SMs) (Razavi et al., 2012; Meirlaen et al., 2002). Most conceptual models apply the storage principle (Chow et al., 1988; Vaes, 1999; Vanrolleghem et al., 2009; Wolfs et al., 2013). The simplest model assumes that there is a linear and unambiguous relationship between the storage and the discharge, which is the dominating model component for flow routing for real-time control applications (e.g. Löwe et al., 2016; Joseph-Duran et al., 2014; Ocampo-Martinez, 2013). However, as pointed out in Chow et al. (1988), Vaes (1999) and Wolfs et al. (2013) the relationship between the storage and the discharge might be influenced by a lot of factors, such as attenuation and backwater, resulting in a non-linear ambiguous relationship. Coping with this has been handled by using different methods from the static-dynamic storage principle (Cunge, 1969; Vaes, 1999) relating the dynamic storage to the inflow, the combiner-splitter approach (Solvi et al., 2005), retention capacity and storage method in the SMUSI model (Muschalla et al., 2006), a combination of the two (Vanrolleghem et al., 2009) and multiple linear segments combined with transfer functions (Wolfs et al., 2013). The surrogate models in the current study are primarily based on the
use of non-linear unambiguous storage-discharge relationships as previously used by Vaes (1999) and Wolfs et al. (2013). This routing concept, as well as the model engine used in the current work, is already being used in real time applications for forecasting flows in the drainage system (Borup et al., 2017). As opposed to the previously mentioned studies, we expand the model to also include surcharging to the surface. For access to the code of the model engine see (Borup, 2018).

Hence the aim of this study is to test surrogate models of varying complexity for urban drainage networks (pipe, channels, etc.) hydraulic modelling with the objective of identifying suitable models for real time control, warning and planning purposes. First the surrogate model setup is examined and benchmarked against the original 1D HiFi network model. Then the uncertainty of the surrogate model parameters, i.e. the variability of the storage-discharge relationship, is examined by using the General Likelihood Uncertainty Estimation (GLUE) methodology and the results of the surrogate models are evaluated from a limit of acceptability approach for different application areas.

2 Methodology

2.1 Surrogate model formulation

The surrogate models are set up by lumping big parts of the distributed drainage system into compartments, as illustrated in Figure 1. Compartments are delineated areas in which the volume of water is modelled with a mass balance by the in- and outgoing discharges. Each compartment can receive water from multiple upstream compartments and discharge to multiple downstream compartments, outlets and surface locations. The division of compartments are subjective and the surrogate models can thus be set up to fit the given modelling purpose. The main guideline used for the compartment delineation in the current work is to place flow-limiting or altering structures, such as throttle pipes, weirs and nodes from which the water can flow in multiple directions along compartment boundaries, since this helps to ensure some
causation between the volume of water in the compartment and the discharges out of this. This minimum requirement can then be supplemented with states that have particular importance for the modeller.

Figure 1: Example of dividing distributed physically-based model into compartments.

Figure 2 shows two compartments in series. Here compartment $n$ can discharge water to the downstream compartment $n+1$. As opposed to (Borup et al., 2017) water can also surcharge to the surface, $Q_{\text{spill}}$. The model boundaries are set so that this surcharged water cannot re-enter the compartments but simply disappears out of the model. Therefore this surcharging is referred to as the spilling discharge as water spills out of the model (DHI, 2014a). Moreover, the SMs only represent the drainage network while the rainfall runoff entering the SMs ($Q_{\text{run}}$ on Figure 2) is obtained from the rainfall runoff simulations of the original HIFI model. This choice has been made to keep the focus on the simulation of the drainage network since this is the time consuming part of a 1D network DPM simulation. The discharges out of each compartment are only dependent on the storage volume in the compartment, which means that backwater from downstream compartments cannot be taken explicitly into account.

Figure 2: Conceptual setup of surrogate model compartments $n$ and $n+1$. The figure illustrates a setup where each compartment has exactly one discharge to the downstream compartment and one spilling discharge, but multiple connections are possible.

\[
\begin{align*}
Q_{\text{run},n} & \quad \text{Runoff discharge to } n\text{th compartment} \\
Q_{\text{drain},n} & \quad \text{Drainage discharge to } n\text{th compartment} \\
Q_{\text{out},n} & \quad \text{Discharge from } n\text{th compartment} \\
Q_{\text{spill},n} & \quad \text{Spilling discharge from } n\text{th compartment} \\
S_{\text{drain},n} & \quad \text{Storage volume in } n\text{th compartment}
\end{align*}
\]
The discharges out of a compartment are governed by volume-discharge curves which are generated from the SM parameters. A total of $2 \times L \times m$ parameters are used to describe one SM compartment, where $L$ is the number of volume-discharge parameter pairs and $m$ is the number of connections out of the compartment. Each volume-discharge parameter pair consists of a volume parameter and a discharge parameter. In-between the volume-discharge pairs the SM interpolates using piecewise linear functions, $f_i$, to create continuous volume-discharge curves for each connection. The term connection is used to describe outflows in the form of compartment-to-compartment, $Q_{\text{out},n}$ or compartment-to-surface, $Q_{\text{spill},n}$. $Q_{\text{spill},n}$ can both be in terms of spill to the surface and combined sewer overflows (CSOs). Spill to the surface, weirs and CSOs will have similar volume-discharge curves as they need to exceed a certain threshold of volume before the discharge begins. The piecewise linear approach has the advantage that non-linearities due to, e.g. overflows and flow constraints can be taken into account. In this study we derive the volume-discharge parameter pairs that are required to estimate the various $f_i$s from DPM result files. This process will be referred to as the training of the surrogate model. The expression for storage for compartment $n$, $S_n$, from Figure 2 is given as:

$$\frac{dS_n}{dt} = Q_{\text{run},n}(t) + Q_{\text{in},n}(t) - Q_{\text{out},n}(t) - Q_{\text{spill},n}(t)$$

where $t$ is the time. $Q_{\text{run},n}$ is the water routed to the sub-surface system, i.e. identical for the DPM and the surrogate model, and $Q_{\text{out},n}$ and $Q_{\text{spill},n}$ are calculated as function of the volume in compartment $n$, $S_n$, using their own individual piecewise linear function as illustrated in next the next paragraph.

### 2.1.1 Training of the surrogate model

The training data are produced by applying rain data to a DPM and then extracting the resulting volumes and discharges from the model results. This process is illustrated in Figure 3(a-c) when using a constructed staircase rain as rainfall data, and Figure 3(d-f) when using a historical rain series instead.
Figure 3: Process of extracting steady state training data (a-c) and dynamic training data (d-f): (a) Input data to the DPM: a staircase rain, (b) DPM resulting volume over time with volume steady state points, (c) steady state volume-discharge points produced by the DPM and equivalent volume-discharge parameter pairs with linear interpolation performed by the SM. (d) Input data to the DPM: a series of historical rain events, (e) DPM resulting volume over time, (f) Cloud of volume-discharge points produced by the DPM and equivalent 3 volume-discharge parameter pairs fitted to represent the shape of the cloud with linear interpolation performed by the SM.

Using a constructed staircase rain for producing the training data allows for the system to reach close to steady-state conditions at the end of each step of the staircase. This allows for extracting un-ambiguous volume-discharge relationships with just a single simulation. The drawback of using these steady state data points is that the steady state conditions are only reached when backwater reaches its maximum, which
implies that the discharge out of a compartment might be underestimated for rain events that do not cause backwater in the system. If, on the other hand, historical rain series are used, the dynamic storage and discharge responses are represented in the training data. But in this case much more data needs to be produced and these training data will show a volume-discharge relationship that is far from un-ambiguous, as illustrated on Figure 3(f). Since we are only dealing with model-to-model simulations and not real observations, any ambiguity caused by uncertainties of the input data e.g. distribution of rainfall is not included.

The number of volume-discharge parameter pairs, \( L \), can be varied. Since the steady-state data points in Figure 3(c) describe an un-ambiguous volume-discharge curve, these steady state volume-discharge points can be used directly as parameter pairs for the SM. Often the shape of the curve can, however, be described sufficiently accurate with much fewer points. In this study this reduction is made using a generic optimizer, \( \text{fmincon} \), from (MATLAB and Statistics Toolbox Release 2015b, 2015) to search for the \( L \) volume-discharge pairs that leads to the lowest squared difference between the full set of training data points and the linear interpolation between the reduced number of points.

### 2.1.2 Surrogate model setups

When setting up a SM decisions need to be made on the coarseness of compartment resolution, the number of volume-discharge parameter pairs for each connection out of the compartments, what input data to use for the DPM to compute the training data and if any additional functions or parameters should be included to describe, e.g. attenuation and other hydraulic constraints. The importance of these decisions is explored in the following using five different SM setups:

- \( \text{SM}_{\text{steady},20} \) uses \( L=20 \) and steady-state training data.
- \( \text{SM}_{\text{steady},4} \) uses \( L=4 \) and steady-state training data.
- \( \text{SM}_{\text{steady},4,\text{div}} \) uses \( L=4 \) and steady-state training data. One compartment is further divided into 17 subcompartments to study the influence of compartment resolution.
- $S_{\text{steady,4+UH}}$ uses $L=4$ and steady-state training data and additionally, to capture the dynamic behaviour seen in Figure 3, we also apply a unit hydrograph (UH) as a post-processing step. The concentration time is fitted manually to each compartment by applying a pulse to the DPM and estimating the travel time of the peak.

- $S_{\text{dynamic,4}}$ uses $L=4$ and historical events as training data. This consists of multiple volume-discharge points as illustrated in Figure 3(f). The parameters are found by using the same volume parameters as for $S_{\text{steady,4}}$ and then calculating the corresponding discharge parameters as the mean of the discharge points of the dynamic training data with approximately the same volume.

### 2.2 MIKE URBAN model formulation

The HiFi DPM used in this study is a MIKE URBAN model (DHI, 2014a). This software from DHI simulates unsteady flows in a pipe network with altering free surface and pressured flow conditions. The hydrodynamic computation is based on an implicit, finite difference numerical solution of the 1D Saint-Venant equations. The computational grid in MIKE URBAN consists of alternating $Q$- and $h$- points in each pipe, i.e. points where the discharge, $Q$, and water level, $h$, are computed at each time step. There is a minimum of three computational points in each pipe and $h$ is furthermore computed for each node. The model automatically generates the computational grid based on the velocity condition (DHI, 2014a). To get an overview of the amount of surcharged water from the drainage network, all nodes in the models are set to spilling nodes, i.e. water can spill to the surface from the drainage network, but it cannot re-enter the drainage system. MIKE URBAN rainfall runoff computation is a separate simulation from where the hydrologic load is used as input to the pipe network. The rainfall runoff simulations are several orders of magnitude faster than the network simulations. The rainfall runoff model applied in this study is model type B – kinematic wave (DHI, 2014b).
2.3 Uncertainty analysis

Every time we apply a model and go from reality to a model space we introduce uncertainties (Refsgaard et al., 2006). In this project there are uncertainties both related to the DPM and the SM. The three major sources of uncertainties in the surrogate models are the model formulation, structure and parameters. Especially the model structure as the SM is highly simplified compared to the DPM. This is both in terms of the equations used to describe the movement of water in the network and also the delineation of the compartments. The training data used can also be a large source of uncertainty as this determines the volume-discharge parameters in the SM and sometimes the structure of the SM. To assess the possible errors introduced by these uncertainties the Generalized Uncertainty Estimation (GLUE) methodology is applied (Beven and Binley, 1992).

GLUE is a methodology to identify models that are acceptable or behavioural model setups from a set of all possible models based on user defined limits of acceptability. It thus deals with the problem of multiple acceptable models (the equifinality thesis). In GLUE the model errors are handled implicitly by assuming similar characteristics of errors in calibration and in prediction. According to Beven (2009) the decisions required to apply the GLUE framework involve the following: (i) Define model structure(s) (ii) Decide which model parameters should be varied (iii) Define prior distribution of model parameters (iv) Decide method for sampling the model space (v) Define likelihood measure (vi) Decide limit of acceptability. After this a large sample of model realizations can be generated with Monte-Carlo simulations.

The most common application of GLUE is to condition models on measured observations (Aronica et al., 2005; Beven and Binley, 2013; Thorndahl et al., 2008). However, in this study we will condition the surrogate models on the original HiFi DPM instead of observations as also done by e.g. Meirlaen et al., (2002), Vaes (1999) and Wolfs et al. (2013).
2.3.1 Selection of model structure and parameters for GLUE

The DPM is run for a series of historical rain events resulting in a swarm of scattered volume-discharge points due to the ambiguous volume-discharge relationships of a dynamical simulation, as shown in Figure 3(f). Note that for the highest values of volume and discharge there is a lower sample size due to the rareress of these extreme events. In the GLUE analysis the discharge parameters are varied from the distribution of discharge points for fixed volume parameters. The number of volume-discharge parameter pairs will be L=4 for each connection. The volume-discharge pairs for surcharging and the volume parameters will be fixed as the ones identified for SM_{steady,4}.

2.3.2 Identifying prior parameter distributions

To identify the discharge parameter distributions, the volume axis is divided into ten equally sized bins. For each of the four volume parameters in each connection, the suitable volume bin is identified. Hereafter a vertical histogram showing the density of the discharge point variation in that bin is computed. These histograms are used to estimate a triangular distribution for each discharge parameter by letting the mode equal the median of the data and min and max values equal the corresponding smallest and largest value, respectively.

2.3.3 Sampling methods

The histograms for each of the L=4 discharge parameters are used to estimate triangular distributions, from where L × 500 discharge parameters are sampled for each connection in each compartment. We apply two different sampling types: sampling type 1 and sampling type 2.

*Sampling method 1* allows for all L=4 discharge parameters to vary independently of each other. This entails that the slope and shape of the volume-discharge curves will vary freely within the span of the training data. For this sampling method it might occur that the n\textsuperscript{th} discharge parameter will be higher than the lowest value for the (n+1)\textsuperscript{th} discharge parameter, i.e. a negative slope in the volume-discharge curve. Such a parameter constellation will imply that the discharge will decrease as a result of an increasing volume,
which conceptually is wrong for this type of model. Such a parameter constellation is thus rejected and not included in the sample.

Attempting to explore the entire span of possible outcomes caused by variation of parameters, we introduce sampling method 2. Sampling method 2 applies the same discharge parameters as sampling method 1, but sorts them from low to high. This means that for all compartments and compartment connections, e.g. the smallest set of discharge parameters will be applied together for the same simulation. By correlating all the discharge parameters in the system we aim to cover the largest possible variation of discharge in the system for with a relatively small sample size. The two types of parameter sets are then used in two separate Monte-Carlo simulations for the SM.

The two sampling methods are expected to provide an outlook on different types of uncertainties in the system. Sampling method 1 represents the full uncertainty of the SM parameters and parameter constellation and is thus the most flexible in this regard. However, moving downstream errors due to parameters might be cancelled through the system. Hence, applying this method would potentially underestimate the variation observed if using actual measured data. Sampling method 2 on the other hand represents the largest possible difference in the discharge output at downstream compartments and thus represents the other extreme. By applying both sampling methods we expect to cover the range of possible outcomes arising if we had had actual measurement data available.

2.3.4 Defining likelihood measure

A uniform likelihood function is applied as seen in Equation (2). This implies that the likelihood or degree of membership is equal for all models which are within the accepted range.

$$L(\theta | \psi_i^*) = \begin{cases} 1 & \text{if } a \leq \psi_i^* - \psi_i(\theta) \leq b \\ 0 & \text{otherwise} \end{cases}$$

(2)
where \( L(\theta | \psi^*_i) \) is the likelihood, \( \theta \) are the model variables shown in Table 1, \( \psi^*_i \) is the probability of the observed value, \( \psi_i(\theta) \) is the probability of the simulated value and \( a \) and \( b \) are lower and upper limits of acceptability.

### 2.4 Performance measures

We use the following measures to quantify the performance of the SMs: Nash-Sutcliffe Efficiency (NSE), Percent Error in Peak (PEP), Peak Difference (PDIFF) and Critical Success Index (CSI) (Bennett et al., 2013) (3-6). From this we further develop Peak Time Difference (PTDIFF) and Percent Error in Accumulated Discharge (PEAD) (7-8):

\[
\text{NSE} = 1 - \frac{\sum_{t=1}^{N}(x_{DPM,t} - x_{SM,t})^2}{\sum_{t=1}^{N}(x_{DPM,t} - \bar{x}_{DPM})^2} 
\]

\[
\text{PEP} = 100 \cdot \frac{\max(x_{DPM,i}) - \max(x_{SM,i})}{\max(x_{DPM,i})} 
\]

\[
\text{PDIFF} = \max(x_{DPM,i}) - \max(x_{SM,i}) 
\]

\[
\text{CSI} = \frac{\text{hits}}{\text{hits} + \text{missed} + \text{false positives}} 
\]

\[
\text{PTDIFF} = t_{\text{peak},DPM,i} - t_{\text{peak},SM,i} 
\]

\[
\text{PEAD} = 100 \cdot \frac{\sum_{t=1}^{N} (x_{DPM,t}) - \sum_{t=1}^{N} (x_{SM,t})}{\sum_{t=1}^{N} (x_{DPM,t})} 
\]

where \( \bar{x}_{DPM} = \frac{1}{N} \sum_{t=1}^{N} x_{DPM,t} \), \( N \) is the length of the records, \( x_{DPM,t} \) is the output from the HiFi DPM at time \( t \), \( x_{SM,t} \) is the output from the SM at time \( t \), \( i \) is the peak number.
3 Case study

3.1 Elster Creek Catchment

The Elster Creek Catchment is located in Melbourne, Australia, and covers an area of approximately 45 km². The area mainly consists of residential areas and has been subject to many large floods in recent years (Olesen et al., 2017). A 1D drainage network DPM of the catchment area presented in (Davidsen et al., 2017) is shown in Figure 4(a). The drainage network is a separate stormwater collection system and there are therefore no CSOs in the catchment. The model consists of 10011 nodes, 10415 links and 12113 sub-catchments, there are six weirs and no basins, pumps or control in the model. The weirs are internal overflows over which the water flows when the system is under high hydraulic loading and thus hydraulically speaking works just as a CSO. The network primarily consists of circular pipes. A large rectangular open channel connects the mid-eastern part of the catchment to the main outlet to the sea. The model contains three outlets to the sea. The main outlet is located in north-west and the two additional outlets are located just north and south of the main outlet. The southern outlet is connected to the open channel by a closed rectangular channel and the northern outlet is connected to the circular pipe network by a closed rectangular channel. The given DPM consists of 44718 calculation points, i.e. states of $q$ and $h$. 
Figure 4: (a) Map of the Elster Creek catchment in Australia with pipes from the MIKE URBAN model presented in (Davidsen et al., 2017) and division into 13 compartments (b) further subdivision of compartment SM1 to 17 subcompartments

The catchment is divided into 13 compartments as shown in Figure 4(a). When dividing the compartments it is ensured that all flow-limiting or altering structures (throttle pipes, weirs, pumps etc.) are placed on compartment boundaries. This is why compartment SM7 is very small; it has weirs both up- and downstream. The connections SM6-SM7 and SM7-SM8 are weir connections with the same hydraulic behaviour as CSO weirs in combined sewer systems. The computational grid has now been reduced from 44718 states to 13. Compartment SM1 represents 1009 nodes and 983 links and 4361 states in the DPM. If SM1 is further subdivided into 17 compartments as shown in Figure 4(b) additional 16 states are used. Hence the number of states in the SMs is close to 3 orders of magnitude lower than in the HiFi models and the computational scheme is reduced from iterative solution of partial differential equations to simple lookup in tables and interpolation.
3.2 Rain data

3.2.1 Training data

Steady state data

For identifying the volume-discharge parameter pairs for the steady state model a staircase rain as shown in Figure 3(a) is applied with intensities ranges from 0.5 to 10 µm/s, corresponding to 1.8 to 36 mm/hour. This range is defined for the catchment to cover system states from nearly empty to exceeding capacity i.e. surcharging. For this catchment each step is set to 4 hours for achieving steady state conditions.

Rain event data for SM\(_{dynamic,4}\) and GLUE

37 years of rain data from a tipping bucket rain gauge is used (Madsen et al., 2017). From this series 45 events are extracted; the 15 events with the largest intensities recorded for durations of 30 and 180 minutes, respectively, and the 15 events with the largest depth. An event definition of a minimum of 10 hours of dry weather between events is used.

3.2.2 Validation data

For validation 52 year of rain data are applied (Australian Bureau of Meteorology, 2015). Similarly 3 x 15 events covering the largest depth, intensity over 30 min and 180 min are extracted. Two events were found to be outside of the calibration range of the SMs and were discarded. This could have been avoided by using higher rainfall intensities during the model calibration.

3.3 Limits of acceptability

Limits of acceptability are set up to distinguish between behavioural and non-behavioural models based on a (subjective) assessment of what is considered a well calibrated model. There are no formal guidelines on how to evaluate models in urban drainage systems. However, a book on code of practice in urban drainage modelling from (Wallingford Procedure Users Group, 1993) contains the following requirements for verification of a model with flow survey data:
i. Peak flow rate from -15% to +25%

ii. Volume of flow from -10% to +20% over the period for which the observed flows are obtained

iii. Surcharge should be from –0.1 m to +0.5 m

iv. The general shape of the two hydrographs should be similar

Moreover models should pass a routine stability test including not exceeding an overall mass balance error of 10% and flooding events that exceed 25 m³ should be verified by evidence of an actual flooding. If all of the above are valid for 2/3 of the events a model is considered validated. This evaluation scheme is, of course, intended for model-to-observations comparisons but we can still use this as an initial guide for our limits. We will include (i) with PEP for Q_{out}, (ii) is included with PEAD for Q_{out}, (iii) is not a direct output of the surrogate models and hence PEP, PTDIFF, PEAD, CSI are used instead for the Q_{spill} and (iv) is included by considering NSE. The limits will be conditioned on the DPM from the requirements of three different applications; RTC, warning and planning, see Table 1. Since the conditioning is from model to model the limits are set to be stricter than those for calibrating models to real observational data. Since the SM is only able to model volumes and discharges, limits on pollutant loads, water levels, velocities etc. that might be relevant is not included.

As previously discussed, models for RTC applications need to be able to predict the distribution of water in the network at all times with sufficient accuracy, with high emphasis on catching the dynamical behaviour of the system. Hence, limits of acceptability for this type of application are set for the discharge and volume in the compartments. This can be used as input for developing a control strategy and to monitor the system capacity during control. The spilling discharge is typically of less importance in RTC, although it is a focus of current research efforts (e.g. Lund et al., 2018; University of Sheffield, 2018), and is therefore not included in the present study. NSE is used as the performance measure for both discharge and volume while PEP is used for the volume only.
Warnings can be important for the Waste Water Treatment Plant (WWTP) for, e.g. reducing the hydraulic loading before a storm and for flooding purposes for implementing adaptive measures to reduce damage costs. For the WWTPs it is important that the model is able to predict the magnitude and timing of the peak to the treatment plant and for flood warning, likewise the magnitude of the peak discharge to the surface and timing. Thus, the peak magnitude and timing are essential and hence PEP and PTDIFF are included for both the discharge and the spilling discharge in the limits of acceptability for this application.

For planning applications the models must be able to give an accurate prediction of the accumulated discharges to the treatment plant and the surface. Thus, limits of acceptability are set for PEAD for the discharge and surcharge. Moreover, the model must be able to give an accurate statistical representation of extremes and hence CSI for spilling discharge is applied. Since the DPM only covers a separate stormwater collection system, no limits will be set on CSOs as there are none of these in the system.

Table 1: Limits of acceptability for the three applications: RTC, Warning and Planning together with the number of accepted models for each limit and application for SM1 using sampling method 1 and 2, respectively.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Limit</th>
<th>Θ</th>
<th>Sampling method 1</th>
<th>Sampling method 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{out}$</td>
<td>NSE &gt;0.8</td>
<td>500</td>
<td>437/500</td>
<td>417/500</td>
</tr>
<tr>
<td>$S$</td>
<td>NSE &gt;0.8</td>
<td>437</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PEP &lt;30%</td>
<td>489</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q_{out}$</td>
<td>PEP &lt;25%</td>
<td>500</td>
<td>496/500</td>
<td>486/500</td>
</tr>
<tr>
<td>PTDIFF</td>
<td>&lt;15min</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q_{spill}$</td>
<td>PEP &lt;50%</td>
<td>496</td>
<td>486</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PTDIFF</td>
<td>&lt;15min</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>$Q_{out}$</td>
<td>PEAD &lt;25%</td>
<td>500</td>
<td>386/500</td>
<td>343/500</td>
</tr>
<tr>
<td>$Q_{spill}$</td>
<td>PEAD &lt;25%</td>
<td>386</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSI</td>
<td>&gt;0.90</td>
<td>500</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4 Results

4.1 Computational time

The DPM network simulations was performed with dynamically adjusted timesteps ranging from 1 second to 1 minute and saving timestep of 1 min. The total number of outputs given by the DPM is nearly 55000 consisting of node water levels, weir discharges, node spilling discharges, link water levels, link discharges and link velocities at the 44718 locations in the network. The SM was computed using fixed 1 min calculation and saving timesteps. The total number of outputs given by the SM is 44 consisting of volumes, discharges and spilling discharges at the 13 locations in the network.

A simulation of one event with three peaks and a duration of 31 hours took 3.3 hours to compute for the DPM for the entire catchment. This was reduced to 8.5 milliseconds with the SM, which is a speedup of $1.4 \cdot 10^6$. To estimate the influence of the amount of output data saved from the DPM the same simulation was performed without saving any output data. This reduced the computational time for the DPM with almost an hour and thereby reducing the SM speed-up to $9.9 \cdot 10^5$. Both models were run on the same PC using a single core only.

4.2 Examination of Surrogate Model Structure

The five different surrogate model setups (presented in Section 2.1.2) are here tested on compartment SM1 (see Figure 4). The spilling discharges from the 17 small new compartments in SM$_{\text{steady},4,\text{div}}$ are aggregated into a single time series to allow for a direct comparison with the other surrogate model setups.

Figure 5 shows the resulting volume, discharge and surcharge, respectively for the SMs and DPM for a selected demonstration event (one of the validation events). SM$_{\text{steady},20}$ is not shown for visual reasons as it cannot be distinguished visually from the results from SM$_{\text{steady},4}$. The top figure shows that all SMs except SM$_{\text{steady},4+\text{UH}}$ are emulating the DPM well in terms of volume in the compartment. The SM$_{\text{steady},4+\text{UH}}$ is higher in volume as water is retained in the compartment due to the unit hydrograph. Of all the surrogate models
SM\textsubscript{steady,4,div} seems to have the best fit. For the discharge, the pink band shows a ±23% margin of the DPM results corresponding to the overall pipe discharge uncertainty found by Hansen and Liu (2004) by using Monte-Carlo simulations and a 84% confidence level. It is seen that all the SMs are within this uncertainty band and again SM\textsubscript{steady,4,div} has the best fit. The surcharge is more difficult to model as it is a more dynamic and local process. All SMs are able to catch the overall behaviour, but only SM\textsubscript{steady,4,div} is able to capture the peak with high accuracy.

Figure 5: Resulting volume, discharge and surcharge over time for a selected demonstration event for the five different SM structures and DPM for compartment SM1.
To further demonstrate the differences between the SMs, resulting volume-discharge curves for the SMs are plotted for the same demonstration event in Figure 6. Here the dynamic behaviour of the DPM is clear.

Figure 6(a) shows the difference between SM\text{steady,4} and SM\text{dynamic,4} for Q_{\text{out}}. The SMs are very similar, so for this event it appears that one can use both steady state and event based training data. Figure 6(b) shows the DPM and SM\text{steady,4+UH}. Here it is seen that the SM\text{steady,4+UH} volume-discharge curve is ambiguous, but the fit to the DPM curve is not good and results in larger volumes as also noted before. Figure 6(c) shows the volume-discharge curve for SM\text{steady,4,div} when all the 17 smaller compartments are combined into one graph. The behaviour of DPM is much better captured when increasing the compartment resolution as also seen in Figure 5. Figure 6(d) shows the volume-surcharge behaviour for all four models. SM\text{steady,4} and SM\text{dynamic,4} are both on the same line. SM\text{steady,4+UH} has a much larger volume for the surcharging. SM\text{steady,4,div} is capturing the DPM well.

Figure 6: Volume-discharge curves for the four different surrogate model structures for the same demonstration event as in Figure 5 for compartment SM1. (a-c) discharge (d) surcharge.

Table 2 shows the performance measures presented in Section 2.3 for the five different SMs as the mean (and median) of all validation events. The column to the left shows the error statistics for the demonstration event shown in Figure 5 and 6 for comparison. Most error measures are similar to those of the single demonstration event, which indicates that this event is representative of the most of the features for all events. All SMs have a high NSE of 0.90 or above for volume and discharge except for
SM\textsubscript{steady,4+UH} with a NSE of 0.86 for volume. For the aggregated surcharge only SM\textsubscript{steady,4,div} performs well with a NSE of 0.84. Both mean and median are computed for PEP. The overall performance of volume and discharge is again better than for surcharging and again SM\textsubscript{steady,4,div} performs the best. It is noticed that all SMs underestimates volume (except SM\textsubscript{steady,4+UH}) and discharge peaks. PDIFF shows that the absolute difference in peaks are <0.4 m$^3$/s for $Q\textsubscript{out}$. For $Q\textsubscript{spill}$ mean PDIFF up to -10 m$^3$/s are noticed, but the median is typically much lower indicating some significant outliers. Only SM\textsubscript{steady,4,div} performs very well here with a mean PDIFF of 1.1 m$^3$/s. The mean peak time difference for $Q\textsubscript{out}$ ranges from 11 to -0.21 min with SM\textsubscript{steady,4,div} and SM\textsubscript{steady,4+UH} with the smallest differences and for $Q\textsubscript{spill}$ from -11 down to -6.7 min for SM\textsubscript{steady,4,div}. CSI is 1 for all SMs meaning that the SMs has captured all the surcharge peaks of DPM and have no false positives. PEAD shows that all models except SM\textsubscript{steady,4,div} underestimate the accumulated surcharge while SM\textsubscript{steady,4,div} overestimates it.

In conclusion, all the SM structures are able to sufficiently accurate describe the behaviour of the DPM. It does not seem to make a big difference if $L$ is set to 4 or 20 volume-discharge parameter pairs, or whether steady state or dynamical training data are used. When trying to describe the ambiguousness of the dynamic volume-discharge curves shown in Figure 6, subdividing the compartments further proved to give the largest improvement. However, this increases the size and complexity of the SMs, and since the other SMs yield reasonably accurate results we proceed with the simplest of these models, SM\textsubscript{steady,4}. 

447
Table 2: NSE, mean and median PEP, PDIFF, PTDIFF, CSI and PEAD for the volume over time, $S$, the outgoing discharge over time, $Q_{out}$, the spilling discharge over time, $Q_{spill}$ for compartment SM1 for the demonstration event from Figure 5 and 6 and the total 3x15 validation events. Numbers in bold and underline have the best score and numbers in bold italic the second best score.

<table>
<thead>
<tr>
<th>Event</th>
<th>SM\text{steady,4}</th>
<th>SM\text{steady,4}</th>
<th>SM\text{steady,20}</th>
<th>SM\text{steady,4,div}</th>
<th>SM\text{steady,4+UH}</th>
<th>SM\text{dynamic,4}</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSE [-]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S$</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.86</td>
<td>0.97</td>
</tr>
<tr>
<td>$Q_{out}$</td>
<td>0.99</td>
<td>0.95</td>
<td>0.95</td>
<td>0.99</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>$Q_{spill}$</td>
<td>0.55</td>
<td>0.33</td>
<td>0.35</td>
<td>0.84</td>
<td>0.33</td>
<td>0.32</td>
</tr>
<tr>
<td>PEP [%]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S$</td>
<td>-4.2 / -4.2</td>
<td>4.1 / 1.6</td>
<td>5.0 / 4.5</td>
<td>0.66 / 0.68</td>
<td>-16 / -19</td>
<td>10 / 9.8</td>
</tr>
<tr>
<td>mean/median $Q_{out}$</td>
<td>-2.5 / -2.5</td>
<td>7.6 / 4.3</td>
<td>5.6 / 2.4</td>
<td>5.8 / 4.3</td>
<td>7.4 / 4.3</td>
<td>7.3 / 6.4</td>
</tr>
<tr>
<td>$Q_{spill}$</td>
<td>44 / 44</td>
<td>-2.1 / 24</td>
<td>-0.2 / 27</td>
<td>-1.6 / -2.0</td>
<td>-2.1 / 23</td>
<td>-5.0 / 17</td>
</tr>
<tr>
<td>PDIFF [m$^3$/s]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean/median $Q_{out}$</td>
<td>-0.37 / -0.37</td>
<td>0.23 / 0.27</td>
<td>0.16 / 0.20</td>
<td>0.31 / 0.37</td>
<td>0.22 / 0.27</td>
<td>0.36 / 0.30</td>
</tr>
<tr>
<td>$Q_{spill}$</td>
<td>6.0 / 6.0</td>
<td>-9.8 / 2.4</td>
<td>-9.3 / 2.7</td>
<td>1.1 / -0.50</td>
<td>-9.8 / 2.4</td>
<td>-10 / 1.7</td>
</tr>
<tr>
<td>PTDIFF [min]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean/median $Q_{out}$</td>
<td>18 / 18</td>
<td>10 / 13</td>
<td>11 / 13</td>
<td>-0.21 / 2.0</td>
<td>-0.28 / 2.0</td>
<td>8.8 / 13</td>
</tr>
<tr>
<td>$Q_{spill}$</td>
<td>8.0 / 8.0</td>
<td>-8.6 / -6.0</td>
<td>-8.3 / -6.0</td>
<td>-6.7 / -2.0</td>
<td>8.6 / -6.0</td>
<td>-11 / -6.0</td>
</tr>
<tr>
<td>CSI [-]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q_{spill}$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>PEAD [%]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Q_{out}$</td>
<td>-0.39</td>
<td>-0.37</td>
<td>-0.57</td>
<td>1.9</td>
<td>-0.42</td>
<td>-0.14</td>
</tr>
<tr>
<td>$Q_{spill}$</td>
<td>9.5</td>
<td>9.5</td>
<td>11</td>
<td>-9.7</td>
<td>9.5</td>
<td>7.4</td>
</tr>
</tbody>
</table>

4.3 Spatial performance

The performance for all 13 compartments is illustrated in Figure 7. This shows (a) NSE for volume and discharge and (b) NSE for spilling discharge. The arrows on Figure 7(a) are scaled according to the magnitude of the mean discharge and their colour show the SM performance for the discharges. Figure 7 (c-d) shows similar results for PEP.
Figure 7: NSE (fig. a-b) and PEP (fig. c-d) for all Elster Creek compartments for SMsteady.4. Figure a and c show values for the volume (colour within compartment) and discharge between compartments (colours in arrows). Figure b and d shows values for the discharge to the surface from each compartment.

Figure 7 shows that the SM performs well with regard to volume and discharge. The scaling of the arrows underlines that the SM performs the best for the major (and most important) discharges through the system. Furthermore, discharge performances for pure weir connections (SM6-SM7, SM7-SM8) are very
good indicating that the SM can easily be applied for modelling CSOs. In Figure 7 it is also clear that the SM has most difficulties in emulating the spilling discharge. The highest PEP values for the spilling discharge are found in compartments SM2, SM4, SM6, SM8 and SM11. There are several reasons for this behaviour. First, the topography of the area will have an impact as the area close to the shoreline is less steep. These compartments will thus be more prone to effects such as backwatering. Moreover, these shoreline compartments function both as storage and transportation, while upstream compartments with higher gradients only serve as transportation. Secondly, the compartments holding the highest PEP values are also some of the compartments with the smallest surcharge amounts. Table 3 shows the size of the accumulated spilling discharge for all compartments relative to the largest compartment, SM3. Compartments SM2, SM4, SM6, SM7, SM8 and SM11 are together with SM13 the compartments with the smallest accumulated spilling discharges. Thus the relative error will be larger for these compartments. In Table 3 the Peak Difference is also shown. These results underline that the PEP values by themselves do not necessarily give a good indication of the SM performance as the compartments with the highest PEP values do not have the highest PDIFF values. The CSI for the spilling discharge is also provided in Table 3. CSI values indicate that the SM is able to capture most peaks in the system. For compartment 7 the SM have six false positives and SM8, SM9 and SM13 all have one miss.

Table 3: Relative size of the spilling discharge in relation to SM3 for all compartments and their PDIFF and CSI values.

<table>
<thead>
<tr>
<th>Compartment</th>
<th>SM1</th>
<th>SM2</th>
<th>SM3</th>
<th>SM4</th>
<th>SM5</th>
<th>SM6</th>
<th>SM7</th>
<th>SM8</th>
<th>SM9</th>
<th>SM10</th>
<th>SM11</th>
<th>SM12</th>
<th>SM13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative size Acc Q_{spill} [%]</td>
<td>32</td>
<td>16</td>
<td>100</td>
<td>8.3</td>
<td>34</td>
<td>12</td>
<td>0.07</td>
<td>5.1</td>
<td>41</td>
<td>38</td>
<td>5.2</td>
<td>24</td>
<td>4.0</td>
</tr>
<tr>
<td>PDIFF [m^3/s] mean</td>
<td>-9.8</td>
<td>3.8</td>
<td>-0.37</td>
<td>2.8</td>
<td>-2.1</td>
<td>5.9</td>
<td>-5.9</td>
<td>1.8</td>
<td>3.3</td>
<td>-3.2</td>
<td>2.1</td>
<td>3.1</td>
<td>0.88</td>
</tr>
<tr>
<td>CSI [-]</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.82</td>
<td>0.97</td>
<td>0.97</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.97</td>
</tr>
</tbody>
</table>
4.4 GLUE surrogate model evaluation

4.4.1 Defining prior distributions of surrogate model parameters

Figure 8 shows the resulting dynamic volume-discharge curves for compartments SM1, SM2, SM6 and SM8 for all validation events. Each 2 min data point is coloured corresponding to the density of the volume-discharge points. The mean and standard deviation of the discharge at varying volumes are also displayed. We have chosen to apply $L=4$ volume-discharge parameter pairs as this number was found to be sufficient for describing the shape of the volume-discharge curve.

A clear pattern is seen for all four compartments: The standard deviation increases with the volume up to a certain point and then it decreases again. The small standard deviations for the larger volumes are due to capacity limitations in the drainage system. At one point the maximum discharge capacity is reached and this limits the variation of the discharge. Another clear pattern, which is observed for all four compartments, is the skewed distribution around the mean.

The steady state points previously used for identifying the volume-discharge parameter pairs in the SM are also shown. In all four compartments the steady state points are close to the and all within the standard deviations. Vertical lines at the fixed volume parameters are shown and the discharge distributions will be created at these cross-sections. The further downstream, the larger the discharge is through the compartment and these compartments therefore have a smaller coefficient of variation. This observation is in agreement with the findings in Section 4.2 when studying the behaviour for a given physical catchment, i.e. that the dynamics are simulated more accurately when applying several SMs in series.
Figure 8: Volume-discharge curves generated with rain presented in Section 3.3.1 for four compartments: (a) SM1, (b) SM2, (c) SM6 and (d) SM8. The volume-discharge points from the DPM are marked with a colour corresponding to the density of points and connected with lines. The steady state points are given as well as the mean and standard deviation. The fixed volume parameters used are shown with vertical lines.

Histograms used for creating prior distributions for each of the four discharge parameters at the fixed volume parameters are shown in Figure 9 for the four compartments for SM1, SM2, SM6 and SM8. Again it is noticed that the sample size is much larger for the lower range and that some of the discharge parameters overlap with their min and max values.
Figure 9: Histograms for creating prior distributions for the discharge parameters for each of the four compartments: (a) SM1, (b) SM2, (c) SM6 and (d) SM8

4.4.2 Performance of surrogate model realizations

The simulation results for SM1 from 500 SM realizations created using sampling method 1 are shown in Figure 10. The prediction bands are wider for the volume than the discharge. Some of the SM realizations take much longer to empty the compartments which results in extended tails for the volume plot. The DPM is within the SM band most of the time. However for the discharge and the spilling discharge the SMs appear to be too soon and too late, respectively.
Similar resulting model simulations for the SM realizations for SM1 with sampling method 2 can be seen in the supplementary material in Figure S.1. The main difference between the results are that when using sampling method 2 the realizations are not intersecting due to the sorting of the parameters. This results in
slightly larger bands and less flexibility, indicating that the choice of sampling method does not have a
major role to play within a simple upstream compartment.

Additional simulation results for the SM realizations for discharge and spilling discharges for compartments
SM1, SM2, SM4 and SM8 can be seen in the supplementary material in Figures S.2-S.3 and S.4-S.5 for
sampling method 1 and 2, respectively. As we move downstream in the system, the difference between the
two sampling methods becomes larger in two ways: Sampling method 2 gets a wider band of models
downstream for discharge compared to sampling method 1 as expected due to the discharge parameter
correlation. However, sampling method 1 results in a larger band of models for the spilling discharge
downstream than sampling method 2. This might seem surprising as the motivation behind using sampling
method 2 was to create the largest model span. The narrow bands for the spilling discharge is because the
largest spilling discharge for a compartment is obtained when the water is held back in the given
compartment, i.e. low discharge parameters, while discharge(s) to the given compartment from upstream
compartments are high, i.e. high discharge parameters. Such a constellation of parameters between
compartments is not possible for sampling method 2 since all parameters within and between the
compartments are correlated. Therefore we actually get the largest band of models for the spilling
discharge when allowing parameters to vary freely as in sampling method 1 for the downstream
compartments.

To further investigate the influence of the discharge parameters, the four different discharge parameters
are plotted with their resulting error in dotty plots. First the discharge parameters from sampling method 1
are plotted with the limits of acceptability from Table 1 for compartment SM1 in Figure 11.

For RTC the limits of acceptability for NSE for both volume and discharge and PEP for volume appear to
accept most of the SM realizations with NSE for the volume being the strictest requirement. For warning
limits all models, except for very few models in PEP for spilling discharge, appear to be accepted. For
planning limits the PEAD for the spilling discharge is the strictest requirement.
Figure 11: NSE, PEP, PTDIFF, PEAD and CSI values as functions of density of discharge parameters 1, 2, 3, 4, respectively, for sampling method 1 with limits of acceptability for the three applications from Table 1. The top 3x4 figures show NSE for the volume, discharge and spilling discharge. The next 3x4 figures similarly show the PEP values. The next 2x4 figures show the PTDIFF values for the discharge and spilling discharge.
The next 2x4 figures show the PEAD values similarly. The final 1x4 figures show the CSI values for spilling discharge. It should be noted that the y-axis for the same error measures might vary for visual ability.

For the entire discharge parameter spectrum there appears to be an upper limit for the performance for NSE for $Q_{out}$ and $Q_{spil}$ and PTDIFF for $Q_{out}$ preventing a perfect error score. All other measures show close to perfect fits somewhere in their range of parameters. This is a result of the model structure and not the parameter ranges. As also seen in Figure 10 none of the SM realizations excel in capturing the behaviour of DPM results for spilling discharge regardless of the discharge parameters. Thus to improve these, the model structure would have to be improved as described in Section 4.2.

Taking a closer look on each of the four discharge parameters, for discharge parameter 1 the NSE plots show an optimum of ~2-3 m$^3$/s for volume and discharge, while for spilling discharge the opposite pattern is seen. For PEP for discharge parameter 1 a “V” shape for the volume is seen peaking at ~3 m$^3$/s, the opposite shape is seen for the discharge while the spilling discharge appears more random. For PTDIFF the lower parameter ranges result in the smallest errors while it for PEAD appears to be the upper range. For discharge parameter 2 and 3 a trend towards the higher values is obtained in NSE, PEP and PEAD. Discharge parameter 4 does not show an optimal range of values as low errors are present in the entire parameter spectrum. This is also the discharge parameter with the smallest range.

For the discharge parameters from sampling method 2 similar dotty plots are shown in Figure 12. Since the sampling is done systematically, the four plots for the four discharge parameters will be the same when regarding the percentiles. Instead we show dotty plots for four compartments: SM1, SM2, SM6 and SM8. Here the same pattern for SM1 is observed as in Figure 12. However, as a result of the sampling method, the SM realizations are now much more bound. This is reflected in the graph as one model realization with a specific discharge parameter cannot have different error measures for that same discharge parameter, and therefore the patterns become much clearer.
It is seen that especially the PEP limit for $Q_{spill}$ for the warning applications reject all models in SM6. As discussed earlier, this error measure might be too strict on these compartments, due to their low surface discharge amounts. No model realizations in the entire parameter range can meet this criterion. It might be that these compartments are highly governed by the larger upstream compartment. Compartment SM8, however, appears to be more sensitive to the discharge parameters, but this compartment also fails for the PTDIFF criterion for warning with SM2 and SM6.

Results for SM$_{steady,4}$ are also shown to evaluate the volume-discharge parameter pairs found using the steady state method. It is seen that the SM$_{steady,4}$ is mostly in the good range of the plot indicating that this method for finding volume-discharge parameter pairs is decent.
Figure 12: NSE, PEP, PTDIFF, PEAD and CSI values as functions of density of percentile discharge parameters for compartments SM1, SM2, SM6, SM8 respectively, for sampling method 2 with limits of acceptability for the three applications from Table 1. The top 3x4 figures show NSE for the volume, discharge and spilling discharge. The next 3x4 figures similarly show the PEP values. The next 2x4 figures show the PTDIFF values.
for the discharge and spilling discharge. The next 2x4 figures show the PEAD values similarly. The final 1x4 figures show the CSI values for spilling discharge. It should be noted that the y-axis for the same error measures might vary for visual ability.

4.4.3 Rejection of non-behavioural models

Table 1 shows number of accepted models for compartment SM1 for each of the three applications and for each of the limits of acceptability within each application for both sampling methods. For RTC the restricting limit is NSE for volume which yields an overall acceptance of 437 and 417 models for method 1 and 2 respectively. For warning, all models are accepted for all limits except PEP for spilling discharge which gives an overall acceptance of 496 and 486 models for method 1 and 2, respectively. For planning, again all models are accept for all other limits than PEAD for spilling discharge, which yields a number of 386 and 343 accepted models for sampling method 1 and 2, respectively. The warning application has the highest number of accepted models for compartment 1 while planning has the lowest number of accepted models. Generally, fewer models are accepted with sampling method 2 than sampling method 1 for the upstream compartment.

Figure 13 shows the number of accepted models for all compartments in the catchment for the two sampling methods. The red colour indicates than none of the 500 model realizations are accepted. Since most compartments are connected to more than one compartment and the acceptance and rejection of SMs depend on the performance of discharge between the compartments, only the largest connection is considered if there are multiple connections. As expected, the downstream compartments along the shore get none to very few accepted models due to severe backwater effects in this part of the catchment. The RTC always have accepted models for all compartments. For the planning applications downstream compartments near the shore have few to none accepted models. The difference between the two sampling methods becomes more apparent as we move downstream. Sampling method 2 generally results in fewer accepted models than sampling method 1, primarily because spilling is not modelled with
sufficient accuracy. If the specific planning purpose was to dimension a downstream structure such as e.g. a treatment plant, one could choose to not include spilling in the conditioning, in which case there would be plenty of accepted models. If the low number of accepted models is not satisfactory different model formulations could be explored, including even simpler models where the compartments dominated by backwater is merged with their upstream compartments to avoid highlighting the problematic hydraulic features that may not be important in the actual use of the model.

![Figure 13: Number of accepted models for each of the three applications. Red colour indicates that none of the 500 SMs were accepted. First 1x3 figures shows the number of accepted models using sampling method 1 and final figures 1x3 shows the number of accepted models using sampling method 2](image)

### 4.4.4 Prediction uncertainty for applications

Figure 14 shows the discharge prediction uncertainty bands for the same demonstration event as shown in Figure 5 and onwards for the three applications with sampling method 1 for the four compartments SM1, SM2, SM6 and SM8. There are models accepted for all three applications in SM1, and hereafter only for planning and RTC as also seen in Figure 13. For SM1 the warning band is the widest, while for SM2, SM6
and SM8, RTC has the widest band with close to all models accepted. It is noticed that the planning application have the best fit around the DPM, but that all accepted applications bands capture the behaviour of the DPM. Furthermore it is observed that all accepted application bands are within the \( \pm 23\% \) DPM uncertainty bands visible in Figure S. In the supplementary material in Figure S.7, the same figure can be seen for sampling method 2. Only for SM1 the bands are narrower, otherwise the bands are similar or wider. Figure S.6 and S.8 in the supplementary material shows the spilling discharge for the four compartments for sampling method 1 and 2, respectively. Here it is seen that the bands from sampling method 1 has a better coverage of the DPM than the narrower bands from sampling method 2.
Figure 14: Results for the DPM (black line) and the SM with prediction uncertainty for the three applications using the limits of acceptability from Table 1 for the outgoing discharge over time, $Q_{out}$, for the four compartments SM1, SM2, SM6, SM8, respectively with sampling method 1.

The prediction bands for the three applications are similar upstream for both the three applications and the two sampling methods. Downstream, the prediction bands differ for both applications and sampling methods. The large differences between the different applications are due to the fact that some of the applications are conditioned on spilling discharge which is difficult to model in downstream backwater-prone areas. Overall sampling method 2 results in larger model bands downstream for discharge than sampling method 1. The bands for spilling discharge are in contrary to the discharge bands wider for sampling method 1 due to the correlation of discharge parameters between compartments in sampling method 2. We would have expected the largest overall model span with sampling method 2 but the correlation of discharge parameters between the compartments hindered maximum variation of the spilling discharge. Hence, in this regard sampling method 1 proved to be the best option with the largest span of models for the spilling discharge indicating that both the internal parameter constellation in a compartment and the parameter constellation between compartments need to be considered. However, the numbers of accepted models for each sampling method did not differ substantially due to the strict limit of acceptability for spilling.
5 Discussion

In recent years fit-for-purpose models in urban drainage modelling has received increased attention. Three issues are prohibiting widespread application. First, no strict formal guidelines are found on how to evaluate models and thus many different performance measures are found throughout literature varying from statistical measures as NSE, RMSE etc. to visual inspection, e.g. (Vanrolleghem et al., 2009; Wolfs et al., 2013). Secondly, the applications of the models are often not included in the choice of these error measures. Finally, no limits are set for the different error measures to differentiate between acceptable models and rejected models. All this means that it has been challenging to develop limits of acceptability for the GLUE analysis both in terms of selection of indicators, the applications and the actual limits of acceptability. These have thus only been made roughly for the three applications without further details on e.g. lead time, and input data quality and uncertainty. Such factors may have a large influence on the model criteria and these might vary greatly for different case study areas and projects. The SMs will have limitations for all three applications areas, as the model criteria do not cover all possible sub-applications. Hence, when modelling e.g. areas highly influenced by backwater, pollutant transport, velocity etc. other more suitable models should be applied. For RTC, the SMs can be applied for static control as the model is able to model flow gates, pumps etc. However, when implementing the control this might cause backwatering and affect upstream compartments. In these situations, these compartments will either have to be recalibrated for the given control strategy or to calibrate and employ several SM structures to accommodate for different control strategies. For warning, the model criterion of peak spilling discharge might be changed to a measure that is more easily interpreted like flood volume. For planning, specific criteria for CSOs will be essential for combined sewer systems.

In many cases an iterative approach is needed to set up a SM given constraints of acceptability and important locations in the system, i.e. the reverse work flow than we have applied in identifying a suitable model structure. A key component in the work flow should be to choose the right compartment resolution with respect to the applications and desired outcome. A simple approach will be to choose only the points
of interest as boundaries between compartments and supplement with compartments to describe key
disruptors (e.g. division of flows) and where a better description of complex flow patterns can be obtained
by using a finer compartment resolution. Currently the SM only models the aggregated surcharge from
each compartment, with no information on the spatial distribution of these flows to the surface within the
given compartment. To achieve more detailed spatial information, the compartment sizes may be reduced
or multiple connections to the surface from each compartment can be used. On the contrary, for
application where only the downstream point is considered, we have shown that the compartment
resolution can be very coarse and still provide sufficiently accurate discharges at the drainage outlet for
most applications. With the 13 compartments we have a good model for the overall state of the drainage
network of the considered domain, but as shown in compartment SM1, if this was increased to e.g. 100
compartments we would most likely have an even better model. Backwater effects represent a challenge,
though, and most likely the best way to circumvent it is to avoid a division into compartments where
backwater effects dominate the flows or include this in the SM structure. The process of setting up the
compartments and calibrating them is currently a manual process. It does require time to examine and
understand the network, deciding on compartment shapes and sizes and reporting links crossing the
compartment boundaries. Once a suitable compartment design has been obtained, the results of this paper
show that the surrogate model can be determined fairly easy and quick. We see a potential for
automatizing the workflow in the future.

We introduced two sampling methods to account for the shortcoming of using data from model-to-model
rather than model-to-observations, i.e. to account for model biases as well as model uncertainty. Although
the difference between the sampling methods were visible, especially downstream in the catchment, the
overall impact on accepted models were smaller than expected and it is doubtful that this procedure is
sufficient to account for all the variation that is present in actual measurement due to, e.g. spatial variation
of rainfall within each compartment.
Both the DPM and the SM introduce model errors. The high level of details in the DPM does not necessarily reduce the model uncertainty, but it prohibits us from exploring this uncertainty because of high computational costs. Compared to the expected uncertainty of the DPM, the uncertainty bands of the surrogate model are well within for the ranges of accepted models. From these reflections one can argue that surrogate models can be applicable for all three types of applications given that sufficient care is taken in setting up the model to the catchment in question. Hence the objectives of the modelling should always be taken into consideration to compute tailor-made fit-for-purpose modelling, including the cost of setting up the SM compared to using a HiFi model.

6 Conclusion

In this paper, lumped drainage surrogate models based on unambiguous volume-discharge curves from an original HiFi 1D network DPM are presented. The models are applied on the Elster Creek Catchment in Melbourne, Australia, with varying training data and model complexity. The key findings are:

- The proposed SMs are approximately six orders of magnitudes faster than the original DPM when reducing the number of states simulated by three orders of magnitude and the SM results are within expected uncertainty bands of results obtained by the HiFi DPM.

- The SMs are robust with respect to extraction of parameters, indicating that choice of training data seems to be of minor importance. The simplest SMs based on steady state training data perform well. We therefore recommend the usage of steady state data when training the SM.

- The results indicate that the most important factor is the compartment resolution. The selection of compartment size should be determined from the model objective as this is a trade-off between accuracy, computational time, and time to set up the SM.

- The SM structure is limited in predicting spilling in flat areas with backwater effects. In other areas the performance is sufficient. The spilling discharge is lumped for the compartments so it does not contain
any spatial information on how the flows to the surface are distributed within each compartment. For this, a finer compartment resolution or multiple surface connections can be applied.

- GLUE demonstrates that the SMs can be tailored to all major types of applications, e.g. RTC, warning and planning. Two sampling methods are applied to compensate for the model-to-model uncertainty. Upstream, different limits of acceptability result in similar prediction uncertainty bands for the three applications. For downstream applications the differences are higher, but probably none of the methods are able to cover the actual uncertainty when validating models against observations.

Despite their simplicity the presented SMs show great potential for usage in urban drainage modelling for various applications. The SMs proved to be within the overall uncertainty bands of the DPM supporting their usage in urban drainage modelling when computational demands of DPMs are prohibitive.

**Acknowledgement**

This research has been financially supported by the Australian Government through the Cooperative Research Centre for Water Sensitive Cities, Realdania through the Smart Cities Water Solutions project under the Klimaspring programme and the Innovation Fund Denmark through the Water Smart Cities project under the Grand Solutions programme (grant no 5157-00009B). Melbourne Water and City of Port Phillip kindly provided the catchment data. Furthermore, we wish to acknowledge Professor Keith Beven for providing his ideas and inputs to this work.

**References**


Australian Bureau of Meteorology, 2015. Rain data extracted from eWater MUSIC for Melbourne, Australia,
using the Pluviograph Rainfall Data Tool.


DHI (Danish Hydraulic Institute), 2014b. MOUSE Runoff Reference Manual


