**Research Article** **Environmetrics**

Received: 30 March 2011, Revised: 13 August 2011, Accepted: 30 August 2011, Published online in Wiley Online Library: 17 October 2011

(wileyonlinelibrary.com) DOI: 10.1002/env.1135

# Grey-box modelling of flow in sewer systems with state-dependent diffusion

**Anders Breinholt[a*], Fannar Örn Thordarson[b], Jan Kloppenborg Møller[b], Morten Grum[c], Peter Steen Mikkelsen[a] and Henrik Madsen[b]**

**Generating flow forecasts with uncertainty limits from rain gauge inputs in sewer systems require simple models with identifiable parameters that can adequately describe the stochastic phenomena of the system. In this paper, a simple grey-box model is proposed that is attractive for both forecasting and control purposes. The grey-box model is based on stochastic differential equations and a key feature is the separation of the total noise into process and measurement noise. The grey-box approach is properly introduced and hypothesis regarding the noise terms are formulated. Three different hypotheses for the diffusion term are investigated and compared: one that assumes additive diffusion; one that assumes state proportional diffusion; and one that assumes state exponentiated diffusion. To implement the state dependent diffusion terms Itô's formula and the Lamperti transform are applied. It is shown that an additive diffusion noise term description leads to a violation of the physical constraints of the system, whereas a state dependent diffusion noise avoids this problem and should be favoured. It is also shown that a logarithmic transformation of the flow measurements secures positive lower flow prediction limits, because the observation noise is proportionally scaled with the modelled output. Finally it is concluded that a state proportional diffusion term best and adequately describes the one-step flow prediction uncertainty, and a proper description of the system noise is important for ascertaining the physical parameters in question.**


**Keywords:** stochastic differential equations; Lamperti transform; parameter estimation; rainfall–runoff; urban drainage

## 1. INTRODUCTION

The increasing challenges in the urban drainage sector caused by climate change, stricter environmental regulations and growing urbanisation, have triggered a need for online models to be used for warning and control purposes (see, for example, (Krämer *et al.*, 2007; Ocampo-Martinez and Puig, 2009; Puig *et al.*, 2009; Giraldo *et al.*, 2010)). However, the inherent uncertainties associated with the model predictions are rarely accounted for, although there seems to be a consensus from several sources regarding uncertainty in modelling, prediction and simulation with urban drainage models (Lei and Schilling, 1996; Willems and Berlamont, 2002; Kuczera *et al.*, 2006; Kleidorfer *et al.*, 2009; Freni and Mannina, 2010; Deletic *et al.*, 2011). Uncertainty is recognised in input data, in the choice of model structure, parameters and measurements for calibration.

In urban rainfall–runoff modelling, input uncertainties refer to the inadequate measurements of the rain input, which is a consequence of spatio-temporal variation of the rainfall events (Willems, 2001; Vaes *et al.*, 2005; Pedersen *et al.*, 2010), as well as errors and biases because of mechanical limitations of the rain gauges (Barbera *et al.*, 2002; Molini *et al.*, 2005; Shedekar *et al.*, 2009). Rainfall is commonly monitored with the nearest available tipping bucket rain gauges (Willems, 2001; Vaes *et al.*, 2005; Pedersen *et al.*, 2010) and as yet only rarely with radars.

Model structure and parameter uncertainty essentially refer to the model design and the parameter estimation method, see the discussion in Harremoës and Madsen (1999). Design and performance analysis is typically based on distributed commercial deterministic models such as MOUSE (Mike Urban)[†] (DHI, Hørsholm, Denmark), SWMM[‡] (US EPA, NW, Washington, DC, USA) and INFOWORKS[§] (Innovyze, Broomfield, CO, USA). Such models are often termed white-box models, because the considered system is formulated using only the available physical knowledge, and any stochasticity in relation to time and space is disregarded. In contrast to the white-box models, the black-box models are built solely on the consideration of the available data to derive a relation between observed input and output. This implies that

* *Correspondence to: Anders Breinholt, Department of Environmental Engineering, Bldg. 113 DTU, DK-2800 Kgs. Lyngby, Denmark. E-mail: anbre@env.dtu.dk*

a *Department of Environmental Engineering, Bldg. 113 DTU, DK-2800 Kgs. Lyngby, Denmark*

b *Department of Informatics and Mathematical Modelling, Bldg. 305 DTU, DK-2800 Kgs. Lyngby, Denmark*

c *Krüger, Veolia Water Solutions and Technologies, Gladsaxevej 363, DK-2860 Søborg, Denmark*

[†] www.dhigroup.com
[‡] http://www.epa.gov/nrmrl/wswrd/wq/models/swmm/
[§] www.innovyze.com

 

physical knowledge about the system is ignored, and both the model structure and the parameterisation are derived and validated by statistical methods, giving the possibility for developing rigorous stochastic dynamical models that can then provide methods for assessing the prediction uncertainty of the model. Black-box models usually provide sufficient short-term predictions when compared with the response time of the system; the system changes are slow, the input errors are significant, but the output errors are small (Gelfan *et al.*, 1999). There are several examples of black-box models that have been used to predict flows in sewers (see e.g. Tan *et al.* (1991), Carstensen *et al.* (1998), El-Din and Schmith (2002), Jonsdottir *et al.* (2007)).

Model-based optimal control of sewer systems presents a case where neither the white-box approach nor the black-box approach is ideal. On the one hand, a white-box model is needed, which is sufficiently accurate to be used for several time steps prediction over wide ranges of state space. On the other hand, black-box models provide access to well-developed tools for structural uncertainty identification. The corresponding model development procedure is guaranteed to converge if certain conditions of identifiability of parameters and persistency of excitation of inputs are fulfilled (Kristensen *et al.*, 2004a). In this paper, we use stochastic state space models, also termed grey-box models, which consist of a set of stochastic differential equations (SDEs), describing the dynamics of the system in continuous time and a set of discrete time measurement equations. This methodology provides a way of combining the advantages of black and white box models by allowing prior physical knowledge to be incorporated into the model structure, and subsequently apply statistical methods for parameter estimation and model validation. This typically yields models with both fewer and physically meaningful parameters. As opposed to white-box models, parameter estimation in grey-box models tends to give more consistent results and less bias because random effects, attributable to process and measurement noise, are no longer absorbed into the parameter estimates, but specifically accounted for by the diffusion and measurement noise terms (Kristensen *et al.*, 2004b). Furthermore, simultaneous estimation of the parameters of these terms provides an estimate of the uncertainty of the model, upon which further model development can be based.

In the present paper a formulation and an estimation of a simple continuous-discrete time stochastic flow model for a sewer system are proposed, which explicitly describe how the measurement and model errors enter into the model. Over the past decades, the proposed grey-box methodology has been applied in diverse disciplines, for example, pharmacology (Tornøe *et al.*, 2004), chemical engineering (Kristensen *et al.*, 2004b; 2004a), district heating (Nielsen and Madsen, 2006), hydrology (Jonsdottir *et al.*, 2001; Jonsdottir *et al.*, 2006), for modelling oxygen concentration in streams (Jacobsen and Madsen, 1996), and within urban drainage systems to model pollutant mass to wastewater treatment plant (Bechmann *et al.*, 1999; Bechmann *et al.*, 2000), flow prediction (Carstensen *et al.*, 1998) and estimation of copper loads in stormwater systems (Lindblom *et al.*, 2007). Generally, the focus of previous studies has been on the physically based part of the SDE model, the so-called drift term. However, in this article, the main focus is on developing the stochastic part of the SDE, the so-called diffusion term because this part of the SDE is significant for a proper uncertainty description of the flow predictions in an urban drainage system.

Following this introduction, the grey-box methodology and important transformations of model states and observations are outlined in Section 2. Section 3 then presents a case study of an urban drainage system with flow measurements affected by both diurnal wastewater variation and rainfall runoff and infiltration inflow. Included here is a description of the catchment area, the data and three model proposals that differ with respect to the diffusion term formulation alone. In Section 4, it is investigated which of the three models best describes the flow predictions, and it is checked if that model can be statistically validated. Finally, conclusions are drawn in Section 5.

## 2. GREY-BOX MODELLING

To ease the introduction of the grey-box methodology, we will begin by presenting the conceptual sewer flow model that later on will be confronted with data from a real catchment area. A conceptual representation of the model is depicted in Figure 1 and a nomenclature of the model is found in Table 1.

### 2.1. State-space formulation of the conceptual sewer flow model

The commercial, physically distributed urban drainage models MOUSE (Mike Urban), SWMM and INFOWORKS all build on partial differential equations (PDEs) for pipe flow calculation. However, when calculating the flow at a specific point in the sewer system, PDEs can often be simplified by substitution with a set of ordinary differential equations (ODEs), and related to the discrete time observations, using a state-space formulation. It is well known that the rainfall–runoff relationship can be modelled with linear reservoirs in series, (Jacobsen *et al.*, 1997; Mannina *et al.*, 2006; Willems, 2010). Hence, the proposed lumped conceptual model for the sewer runoff system displayed in Figure 1 consists of linear reservoirs that are based on ODEs. The first reservoir ($S_1$) represents the first state variable in the model, receiving runoff from the contributing area $A$ caused by the rainfall registered at the two rain gauges $P_{316}$ and $P_{321}$. The weighting parameter $\alpha$ is defined to account for the fraction of the measured flow that can be attributed to rain gauge $P_{316}$. Furthermore, we assume that the measured flow from the contributing area $A$ is fully described by the two rain gauges, implying that the contribution from rain gauge $P_{321}$ is equal to $1-\alpha$.

The second reservoir ($S_2$), and correspondingly the second state variable in the two-state model, receives outflow from the first reservoir and diverts it to the flow gauge downstream from the catchment. The purpose of the reservoirs in the model is to simulate the time delay from a rainfall event, which is being registered at the rain gauges until a corresponding runoff is observed at the location of the flow metre. The time delay is due to both overland runoff time, transportation in the sewer, and in case of heavy rain, also internal storage of water in detention basins.

The wastewater flow $D$ is periodic with a diurnal cycle, that is, in dry weather conditions, the observed flow variation is described by the diurnal variation in the wastewater production. The following harmonic function is used:

$$D_k = \sum_{i=1}^{2} \left( s_i \sin \frac{i 2\pi k}{L} + c_i \cos \frac{i 2\pi k}{L} \right) \tag{1}$$

  

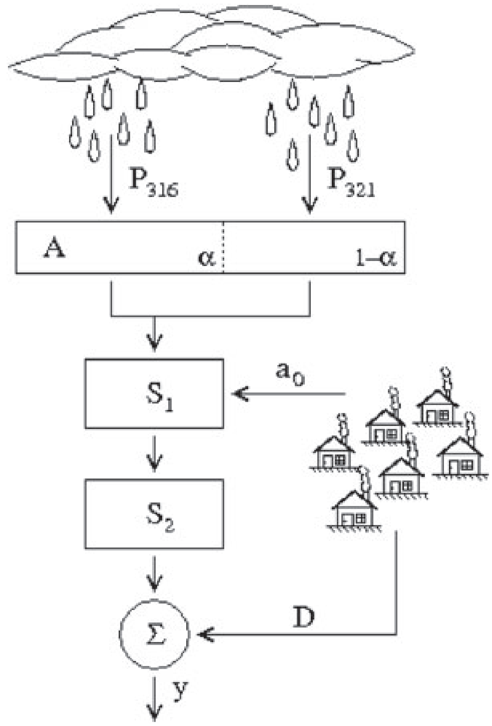


**Figure 1.** The conceptual model; a system of two linear reservoirs.

where $L$ is the period of 24 h and the parameters $s_1$, $c_1$, $s_2$ and $c_2$ are non-physical parameters to be estimated.

To fully describe the wastewater flow, a constant term for the average dry weather flow $a_0$ must be added to Equation (1). However, it was decided to attach $a_0$ to the first state $S_1$ to secure the physical interpretation of the system, that is, water is always passing through the system, securing that the reservoirs do not dry out.

By considering the conceptual model displayed in Figure 1, it follows that a state-space formulation of the model can be described as

$$\mathrm{d}\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} = \begin{bmatrix} \alpha A P_{316,t} + (1-\alpha) A P_{321,t} + a_0 - \frac{2}{K} S_{1,t} \\ \frac{2}{K} S_{1,t} - \frac{2}{K} S_{2,t} \end{bmatrix} \mathrm{d}t \tag{2}$$

and the observation equation can be formulated as

$$Y_k = \left( \frac{2}{K} S_{2,k} + D_k \right) + \varepsilon_k \tag{3}$$

The term $K$ in the system Equation (2) represents the mean retention time of the system, that is, the average time between a rainfall event being registered and the corresponding flow rise being measured by the flow gauge. Diverting the flow through two reservoirs indicates that two retention time coefficients could be used; accordingly, one for the flow from $S_1$ to $S_2$ and a second one for the flow from $S_2$ to the flow measurement station. However, we assume that the two retention times are identical and multiplying with the number of reservoirs in the series, the mean retention time for the flow through the whole sewer system is obtained. It is noted that the second state $S_2$ appears in the observation equation whereas the first state $S_1$ is unobserved, that is, a hidden state. It is furthermore seen that the error between observed and predicted flow is described by the output error term $\varepsilon_k$ that is assumed to be a white noise process with $\varepsilon_k \in N(0, S)$, where $N(0, S)$ is a normal distribution with zero mean and variance $S$.

### 2.2. Grey-box representation of the conceptual model

The model formulation as described by the Equations (2) and (3) does not distinguish observation error from input and model structural error. In the grey-box methodology, this distinction is made by introducing a *diffusion term* also referred to as a process noise term that specifically accounts for model structural deficiencies and input errors in a lumped way. In Equation (4) shown hereafter, a constant diffusion term has been introduced.

$$\mathrm{d}\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha A P_{316,t} + (1-\alpha) A P_{321,t} + a_0 - \frac{2}{K} S_{1,t} \\ \frac{2}{K} S_{1,t} - \frac{2}{K} S_{2,t} \end{bmatrix}}_{\text{drift term}} \mathrm{d}t + \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}}_{\text{diffusion term}} \mathrm{d}\boldsymbol{\omega}_t \tag{4}$$



  

**Table 1.** Nomenclature of the conceptual flow model

| Symbol | Description | Unit |
|---|---|---|
| Inputs | | |
| $P_{316,t}$ | Rain gauge input | m/h |
| $P_{321,t}$ | Rain gauge input | m/h |
| Rainfall-runoff model parameters | | |
| $A$ | Impermeable fast runoff area | $10^4$ $m^2$ |
| $K$ | Retention time, fast runoff | h |
| $\alpha$ | Rain gauge weighting coefficient | — |
| Wastewater flow model parameters | | |
| $a_0$ | Average waste water flow | $m^3/h$ |
| $s_1$,$s_2$ | Sine constants | — |
| $c_1$,$c_2$ | Cosine constants | — |
| Model states | | |
| $S_{1,t}$ | State of first linear reservoir | $m^3$ |
| $S_{2,t}$ | State of second linear reservoir | $m^3$ |
| Process noise | | |
| $\sigma_1$ | Standard deviance, state 1 | $m^3$ |
| $\sigma_2$ | Standard deviance, state 2 | $m^3$ |
| Model output | | |
| $Y_k$ | Observed flow at time step $k$ | $m^3/h$ |
| Observations | | |
| $\mathcal{Y}_N$ | $N$ number of flow observations | $m^3/h$ |
| Observation noise | | |
| $\epsilon_k$ | $N(0, S)$ | $m^3/h$ |
| Time | | |
| $k$ | Time step counter | — |
| $t$ | Continuous time | h |

and the observation equation then changes to

$$Y_k = \left(\frac{2}{K} S_{2,k} + D_k\right) + e_k \tag{5}$$

The diffusion term adds two standard deviations ($\sigma_1$ and $\sigma_2$) that account for prediction uncertainty on $S_1$ and $S_2$. $\boldsymbol{\omega}_t$ is, in this case, a two-dimensional standard Wiener process, that is, $\mathrm{d}\boldsymbol{\omega}_t \sim \sqrt{\mathrm{d}t}\,N(0, 1)$, where $N(0, 1)$ is a normal distribution with zero mean and unit variance. The deterministic part of the state equations are referred to as the *drift term*. In Equation (4), the input uncertainty is primarily related to $\sigma_1$ because the rain input enters this first reservoir, whereas the model structural uncertainty will appear in both $\sigma_1$ and $\sigma_2$. The only change in the observation equation (Equation 5) is that $\varepsilon_k$ is substituted with $e_k$ because now the total output error ($\varepsilon_k$ in Equation (3)) has been divided into a process noise represented by $\sigma_1$ and $\sigma_2$ and an *observation noise* term ($e_k$).

  

In the grey-box terminology, it is also possible to let the uncertainty on the state predictions depend on the current state level, the inputs or some parameters instead of using a constant diffusion term. In the case of urban drainage systems, it seems reasonable to expect that the uncertainty on the state prediction must somehow be related to the rain input. We will return to this in Section 3.2 and now introduce the grey-box methodology in its general form:

$$dX_t = \underbrace{f(X_t, u_t, t, \theta)}_{\text{drift term}}\,\mathrm{d}t + \underbrace{\sigma(X_t, u_t, t, \theta)}_{\text{diffusion term}}\,\mathrm{d}\omega_t \tag{6}$$

$$Y_k = h(X_k, u_k, t_k, \theta) + e_k \tag{7}$$

where Equation (6) is the *system equation*, describing the dynamic time evolution ($t \in \mathbb{R}_O$) of the physical state of the system in continuous time, and Equation (7) is again the *observation equation* that relates the modelled output to the observations $Y_k \in \mathbb{R}^l$ at discrete sampling instants $t_k$ ($k = 1, \dots, N$) for $N$ number of measurements. Note that in the system equation, $f(\cdot) \in \mathbb{R}^n$ represents the drift term and $\sigma(\cdot) \in \mathbb{R}^{n \times n}$ the diffusion term. Here, $\omega_t$ is a $n$-dimensional standard Wiener process. In the system equation, $X_t \in \mathbb{R}^n$ represents the state variables of the model, the input variables are $u_t \in \mathbb{R}^m$ and the parameters are $\theta \in \mathbb{R}^p$. As seen, the diffusion term $\sigma(\cdot)$ can be a function of the states, the inputs, the time or some parameter. In the observation equation, the observation error term $e_k$ is assumed to be a one-dimensional white noise process with $e_k \in N(\mathbf{0}, S(u_k, t_k, \theta))$. It is seen that the observation noise can be a function of the inputs, the time and parameters.

### 2.3. Parameter estimation

Given the model structure in Equations (6) and (7), the unknown parameters can be determined by finding the parameters that maximise the likelihood function for a given sequence of measurements (Kristensen *et al.*, 2004b).

For time series models, the likelihood function is based on the product of conditional densities, (Madsen, 2008). To express the likelihood as product of conditional densities, the rule $P(A \cap B) = P(A|B)P(B)$ is applied, and with a sequence of measurements, denoted as $\mathcal{Y}_N = [Y_N, \dots, Y_0]$, the likelihood function is the joint probability density:

$$L(\theta; \mathcal{Y}_N) = P(\mathcal{Y}_N|\theta) = \left(\prod_{k=1}^{N} P(Y_k|\mathcal{Y}_{k-1}, \theta)\right) P(Y_0|\theta) \tag{8}$$

which is seen by repeated use of $P(A \cap B) = P(A|B)P(B)$. From (8), it is recognised that the likelihood function consists of a product of one-step-ahead conditional densities. The likelihood function can only be evaluated if the initial probability density $P(Y_0|\theta)$ is known, and all subsequent conditional probability densities can then be assessed by successively solving Kolmogorov's forward equation and applying Bayes' rule, (Jazwinski, 2007). In practice, however, this approach is not computationally feasible and an alternative approach is required. Because the system equations in Equation (6) are driven by a Wiener process, which has Gaussian increments; it seems reasonable to assume that the conditional densities can be approximated by Gaussian densities. For linear systems, the conditional probabilities in the likelihood function in Equation (8) are Gaussian, but for nonlinear systems, this remains an approximation.

The Gaussian density is completely characterised by its mean and covariance of the one step prediction, which are denoted by $\hat{Y}_{k|k-1} = E\{Y_k|\mathcal{Y}_{k-1}, \theta\}$ and $R_{k|k-1} = V\{Y_k|\mathcal{Y}_{k-1}, \theta\}$, respectively, and by introducing an expression for the innovation formula, $\epsilon_k = Y_k - \hat{Y}_{k|k-1}$, the likelihood function can be rewritten as (Madsen, 2008)

$$L(\theta; \mathcal{Y}_N) = \left(\prod_{k=1}^{N} \frac{\exp\left(-\frac{1}{2}\epsilon_k^\top R_{k|k-1}^{-1}\epsilon_k\right)}{\sqrt{\det(R_{k|k-1})}\left(\sqrt{2\pi}\right)^l}\right) P(Y_0|\theta)$$

where the conditional mean and covariance are calculated using a Kalman Filter (KF) for linear models or an Extended Kalman Filter (EKF) for nonlinear models. Finally, the parameter estimates can be obtained by conditioning on the initial values and solving the optimisation problem

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \{\log(L(\theta; \mathcal{Y}_N|Y_0))\} \tag{9}$$

In general, it is not possible to optimise the likelihood function analytically, and numerical methods must be applied, (Kristensen and Madsen, 2003).

The maximum likelihood method also provides an assessment of the uncertainty for the parameter estimates in Equation (9) because the maximum likelihood estimation is asymptotically and normally distributed with mean $\theta$ and covariance matrix

$$\hat{\Sigma}_\theta = H^{-1}$$

The matrix $H$ is the Fisher information matrix and is given by

$$h_{ij} = -E\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(L(\theta|\mathcal{Y}_{k-1}))\right\} \qquad i, j = 1, \dots, p \tag{10}$$

  

where in practice, an approximation for $\boldsymbol{H}$ is obtained by the observed Hessian $h_{ij}$ evaluated for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Due to the asymptotic Gaussianity of the estimator in Equation (9), a $t$-test can be performed to ascertain if the estimated parameters are statistically significant.

When estimating the unknown parameters of the model from a set of data, the continuous discrete time formulation enables the model to function flexibly with possibilities for varying sample times and missing observations in the data series.

### 2.4. Transforming the state

To solve the estimation problem, the open source software CTSM[¶] (Kristensen and Madsen, 2003) is used. Most physical systems have natural constraints in the model structure, for example, the mass balance in the system cannot be neglected, or states need to be positively defined. The restrictions related to positively defined states can partly be dealt with by state dependent diffusion terms in the SDEs. However, this requires a higher order KF, which has not been implemented in CTSM because it was shown to become numerically unstable (Vestergaard, 1998). Hence, it is not directly possible to estimate parameters in models with state dependent diffusion terms. To obtain efficiency and numerical stability in the estimation, a transformation of the SDEs is required to generate a new set of equations, where the diffusion term can be independent of the state variable, (Baadsgaard *et al.*, 1997).

The procedure of transforming a general SDE into a form with state independent diffusion term is frequently referred to as the Lamperti transform, (Iacus, 2008). Existence is only guaranteed for one-dimensional diffusion processes, whereas for multi-dimensional diffusion processes, existence depends on the structure of the diffusion term (Luschgy and Pagés, 2006). The one-dimensional diffusion is the simplest case of a state-dependent diffusion term in SDEs, and only the univariate transformation is considered here. For the multivariate transform, we refer to (Møller and Madsen, 2010).

For any given $t$, assume that the drift term $f_i(\cdot) = f_i(\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{\theta})$, and the diffusion term $\sigma_{ii}(\cdot) = \sigma_{ii}(X_i, \boldsymbol{u}, \boldsymbol{\theta})$, $\sigma_{ij} = 0$ for $i \neq j$, then, the SDE for the transformed state $Z_i = \phi(X_i) = \phi$ is obtained by Itô's formula (Øksendal, 2003):

$$\mathrm{d}Z_i = \left( \frac{\partial \phi}{\partial t} + f_i(\cdot) \frac{\partial \phi}{\partial X_i} + \frac{\sigma_{ii}^2(\cdot)}{2} \frac{\partial^2 \phi}{\partial X_i^2} \right) \mathrm{d}t + \sigma_{ii}(\cdot) \frac{\partial \phi}{\partial X_i} \mathrm{d}\omega_i \tag{11}$$

where $\phi$ is a twice continuously differentiable function for $(t, X_i) \in (\mathbb{R}_+, \mathbb{R})$. Focusing on the diffusion term in the transformed SDE in Equation (11) shows that the state dependence can be removed from the equation by solving

$$\frac{1}{\sigma_{ii}(\cdot)} = \frac{\partial \phi}{\partial X_i}$$

and the Lamperti transform for the $i$th state becomes

$$Z_i = \phi(t, X_i) = \int \mathrm{d}\phi(t, \xi) \bigg|_{\xi = X_i} = \int \frac{\partial \phi}{\partial \xi} \mathrm{d}\xi \bigg|_{\xi = X_i} = \int \frac{\mathrm{d}\xi}{\sigma_{ii}(\xi, \boldsymbol{u}_t, t, \boldsymbol{\theta})} \bigg|_{\xi = X_i} \tag{12}$$

The Lamperti transform in Equation (12) provides a system equation with a state independent diffusion term, but the parameters are the same as in the original SDE, and the model is still describing the same input–output relationship. Thus, considering a transformation for all system equations in a model, the transformed grey-box model is written as follows:

$$\mathrm{d}\boldsymbol{Z}_t = \tilde{\boldsymbol{f}}(\boldsymbol{Z}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})\mathrm{d}t + \tilde{\boldsymbol{\sigma}}(\boldsymbol{u}_t, t, \boldsymbol{\theta})\mathrm{d}\boldsymbol{\omega} \tag{13}$$

$$\boldsymbol{Y}_k = \tilde{\boldsymbol{h}}(\boldsymbol{Z}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{14}$$

where $\boldsymbol{Z}$ is a vector including the transformed states and the function $\tilde{\boldsymbol{f}}$ is a description for the drift terms of the transformed state space model. $\tilde{\boldsymbol{h}}$ represents the new observation equation, but now, as a function of the transformed states, and $\tilde{\boldsymbol{\sigma}}$ is a state independent diffusion term.

### 2.5. Example of the Lamperti transform

In what follows, the properties of the Lamperti transform will be exemplified and later applied in a case study. The notation for the SDE is simplified by omitting input dependence for the diffusion, because focus is on state dependence. Hence, the $i$th SDE of the system equation in Equation (6) is written as

$$\mathrm{d}X_i = f_i(\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{\theta})\mathrm{d}t + \sigma_{ii}(X_i, \boldsymbol{\theta})\mathrm{d}\omega \tag{15}$$

The drift term is assumed to be linear. The function $f_i$ can then be separated into two terms, one describing the linear relation to the state $(a_i)$ and a second term $(b_i)$ counting for the relation to any other variable influencing the state $X_i$, that is, the input variables $\boldsymbol{u}$ and the remaining states $\boldsymbol{X}^*$, where $\boldsymbol{X}^* = \boldsymbol{X} \backslash X_i$. Using Equation (15), the $i$th SDE becomes

$$\mathrm{d}X_i = \left(b_i(\boldsymbol{X}^*, \boldsymbol{u}, \boldsymbol{\theta}) + a_i(\boldsymbol{\theta})X_i\right) \mathrm{d}t + \sigma_{ii}(X_i, \boldsymbol{\theta})\mathrm{d}\omega \tag{16}$$

The focus is now on the diffusion term $\sigma_{ii}$, whereas the drift term is considered as displayed in Equation (16). Only the system equation is considered because the observation equation remains unchanged.

[¶]Continuous-Time Stochastic Modelling - www.imm.dtu.dt/ctsm

  

Example: $\sigma_{ii}(\cdot) = \sigma_i X_i^{\gamma_i}$

One of the simplest diffusion formulations in SDEs is to assume linear dependence between the state and corresponding noise, but linearity is not always a satisfactory state dependence. Therefore, the diffusion is a function of the state to the power of $\gamma_i$, where, for now, $\gamma_i$ is arbitrary. The system equation then becomes

$$dX_i = \left(b_i(\boldsymbol{X}^*, \boldsymbol{u}, \boldsymbol{\theta}) + a_i(\boldsymbol{\theta})X_i\right)dt + \sigma_i X_i^{\gamma_i} d\omega \tag{17}$$

where $\sigma_i$ is a constant term. According to the Lamperti transform in Equation (12), the function $\sigma_{ii}(\cdot) = \sigma_i X_i^{\gamma_i}$ should be considered to obtain the transformed state $Z_i$, but because $\sigma_i$ is a constant and not influencing the result of the integration, it can be neglected in the transformation. Consequently, $\sigma_i$ remains in the system equation in Equation (17) and only the part of the diffusion term with state $X_i$ involved is reflected in the state transformation.

Using Equation (12), the Lamperti transform for the SDE in Equation (17) is then

$$Z_i = \phi(t, x_i) = \int \frac{d\xi}{\xi_i^{\gamma_i}}\bigg|_{\xi = X_i} = \frac{X_i^{1-\gamma_i}}{1-\gamma_i} \Leftrightarrow X_i = ((1-\gamma_i)Z_i)^{\frac{1}{1-\gamma_i}} \tag{18}$$

To obtain the SDE of the transformed state, Itô's formula is applied, as described in Equation (11), but here, it utilises both the first and second derivatives of the transformed state $Z_i$ with respect to the original state $X_i$, as well as the first time derivative of the transformed state. For the transformation in Equation (18), the derivatives in Equation (11) become

$$\frac{\partial \phi}{\partial X_i} = \phi_x = \frac{1}{X_i^{\gamma_i}} \qquad \frac{\partial^2 \phi}{\partial X_i^2} = \phi_{xx} = -\frac{\gamma_i}{X_i^{\gamma_i - 1}} \qquad \frac{\partial \phi}{\partial t} = \phi_t = 0$$

and Itô's formula then gives the transformed state:

$$\begin{aligned} dZ_i &= \left(\phi_t + \phi_x f_i + \frac{1}{2}\phi_{xx}\sigma_i^2\right)dt + \phi_x \sigma_i d\omega \\ &= \left(0 + \frac{b_i(\cdot) + a_i(\cdot)X_i}{X_i^{\gamma_i}} + \frac{1}{2}\left(-\frac{\gamma_i}{X_i^{\gamma_i+1}}\right)\sigma_i^2 X_i^{2\gamma_i}\right)dt + \frac{\sigma_i X_i^{\gamma_i}}{X_i^{\gamma_i}} d\omega \\ &= \left(\frac{b_i(\cdot)}{X_i^{\gamma_i}} + a_i(\cdot)X_i^{1-\gamma_i}) - \frac{1}{2}\gamma_i \sigma_i^2 X_i^{\gamma_i - 1}\right)dt + \sigma_i d\omega \end{aligned} \tag{19}$$

Substitute the state transformation in Equation (18) into the transformed SDE in Equation (19), and obtain,

$$\begin{aligned} dZ_i &= \left(\frac{b_i(\cdot)}{((1-\gamma_i)Z_i)^{\frac{\gamma_i}{1-\gamma_i}}} + a_i(\cdot)\,((1-\gamma_i)Z_i)^{\frac{1-\gamma_i}{1-\gamma_i}} - \frac{1}{2}\gamma_i\sigma_i^2\,((1-\gamma_i)Z_i)^{\frac{\gamma_i - 1}{1-\gamma_i}}\right)dt + \sigma_i d\omega \\ &= \left(b_i(\cdot)\,((1-\gamma_i)Z_i)^{-\frac{\gamma_i}{1-\gamma_i}} + a_i(\cdot)(1-\gamma_i)Z_i - \frac{1}{2}\frac{\gamma_i}{1-\gamma_i}\sigma_i^2 z^{-1}\right)dt + \sigma_i d\omega \\ &= \tilde{f}_i(\boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{\theta})dt + \sigma_i d\omega \end{aligned} \tag{20}$$

corresponding to the $i$th state in the transformed system equation in Equation (13).

By setting $\gamma_i$ equal to 1, a linear state dependence in $X_i$ can be obtained by applying Equation (12),

$$Z_i = \log(X_i) \Leftrightarrow X_i = e^{Z_i} \tag{21}$$

The Lamperti transform for an SDE with a diffusion term that is linearly dependent on the state is the logarithmic transform (or log-transform) because the integral in the Lamperti transform results in a logarithmic relation between the original state and the transformed one. To find the SDE of the transformed state, Equation (11) is again applied to obtain the following:

$$\begin{aligned} dZ_i &= \left(b_i(\cdot)e^{-Z_i} + a_i(\cdot) - \frac{1}{2}\sigma_i^2\right)dt + \sigma_i d\omega \\ &= \tilde{f}_i(\boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{\theta})dt + \sigma_i d\omega \end{aligned} \tag{22}$$

Notice that the diffusion parameters $\sigma_i$ and $\gamma_i$ in Equations (20) and (22), as well as the model parameters in $b_i(\cdot)$ and $a_i(\cdot)$ are unaffected by the transformation. Hence, the estimated parameters in the transformed model can be directly implemented into the original model.

To estimate the $\gamma_i$ parameters, a restriction is required to obtain proper prediction intervals for coverage of the variation in the observations. With $\gamma_i = 0.5$, the state has a positive probability of reaching zero, if the input parameters are small compared with the diffusion parameters, (Iacus, 2008; Møller and Madsen, 2010), and the EKF is not suited for such distributions, whereas for $\gamma_i > 1$, existence and uniqueness of the system are not guaranteed (Øksendal, 2003). Thus, the $\gamma_i$ parameters need to take values between 0.5 and 1 during estimation.

### 2.6. Transforming the observations

The implicit assumption of using a constant observation noise term is that the observation noise is independent of states. However, for many physical systems, this is unrealistic and a noise term that increases proportionally with the output is more appropriate, that is,

$$\boldsymbol{Y}_k = \boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta})\boldsymbol{\epsilon}_k$$

where $\epsilon$ is log-normally distributed, and the observation functions $\boldsymbol{h}$ are the same as shown in Equation (7). Consequently, the observation noise is scaled with the size of the measured model output. This is beneficial because studies of flow metre uncertainty have shown that measurement uncertainty increases proportionally with the flow magnitude (Bertrand-Krajewski *et al.*, 2003).

One of the benefits of expressing the observation equation with an additive Gaussian noise, as in Equation (7), is that the assumption of Gaussianity for the residuals enables the use of the EKF and statistical tests to verify the proposed model, (more regarding model validation in the following section). CTSM utilises these tests, and in the implementation, only additive noise terms in the observation equations are allowed. Thus, to separate the noise term from the model, where the noise is multiplicative and log-normally distributed, a logarithmic transform of the measurements is required:

$$\begin{aligned}\log(\boldsymbol{Y}_k) &= \log\left(\boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta})\boldsymbol{\epsilon}_k\right) \\ &= \log\left(\boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta})\right) + \log\left(\boldsymbol{\epsilon}_k\right) \\ &= \log\left(\boldsymbol{h}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta})\right) + \boldsymbol{e}_k\end{aligned} \tag{23}$$

The log-transformed observations can then be applied in CTSM.

### 2.7. Model validation

One of the main aspects of the grey-box modelling framework is its predictive ability, which implies that the output errors are examined for any systematic pattern for further extension of the model. Several statistical tools are utilised for the validation procedure, in which all have their own properties for identifying the lack of fit in the model. The statistical tools used in the paper are all well described in Madsen (2008).

The model residuals are useful for the validation. The general assumption for the residuals for an adequate model is that they are white noise. Plotting the sample autocorrelation function (ACF), and the sample partial autocorrelation function (PACF) for the residuals will show if the residuals eventually are autocorrelated. In the frequency domain, the cumulative periodogram is useful for detecting the deviation from the white noise assumption for the residuals. With the cumulative periodogram, any hidden periodicities, including seasonality, in the residuals can be detected. For more details on the cumulative periodogram, see Madsen (2008) and Priestley (1981).

## 3. CASE STUDY AND MODEL PROPOSALS

The grey-box methodology is applied to find a satisfactory flow model for a sewer system. As already seen in Section 2.2, the proposed model has a rather limited physical structure, and therefore, the advantages of adequately formulating the diffusion term of the SDEs to cope with model deficiencies and input uncertainties will be emphasised.

### 3.1. Catchment, drainage system and data

Figure 2 gives an overview of the study catchment, which is situated in the north-western part of greater Copenhagen in Ballerup Municipality. The total area is $1320 \cdot 10^4$ m$^2$. Most of the catchment area (93%) utilises a separate system with two parallel pipes for wastewater and stormwater, whereas the remaining 7% is served by a combined sewer system in which both wastewater and stormwater flow through the same pipe. A significant amount of infiltration inflow into the sewer network is taking place, probably because of worn-out pipes and faulty connections. A flow metre has been installed downstream of the catchment area to attempt to ascertain the extent of this leakage. The flow metre is a semi-mobile ultrasonic Doppler type. It is placed in an interception pipe with a diameter of 1.4 m. The flow metre logs every 5 min.

There are around 50,000 inhabitants living inside the catchment area, which is one of several sub-catchment areas that diverts water to the second largest wastewater treatment plant (WWTP) in Denmark, called Avedøre WWTP. There are a couple of small pumping stations and one storage basin inside the catchment area, with an approximate capacity of 4000 m$^3$. The two closest rain gauges from the national Danish tipping bucket network, (Jørgensen *et al.*, 1998), indicated $P316$ and $P321$ in Figure 2, have a 0.2 mm resolution and are located outside the studied catchment area, approximately 12 km apart.

A nearly 3 month period, (1 April–21 June, 2007) is used for estimation. The measured precipitation varies considerably from one rain gauge to the other, and spatio-temporal rainfall variation is clearly identified. This is illustrated in Figure 3 that shows the accumulated precipitation measured at each rain gauge, $(P_{316})$ and $(P_{321})$ plotted on a log scale. If a given rainfall registration at the two gauges is separated by more than 1 h, they are considered to be separated events. Note how this distinction results in some rainfall events being registered at only one of the rain gauges.



  


**Figure 2.** The Ballerup catchment area.

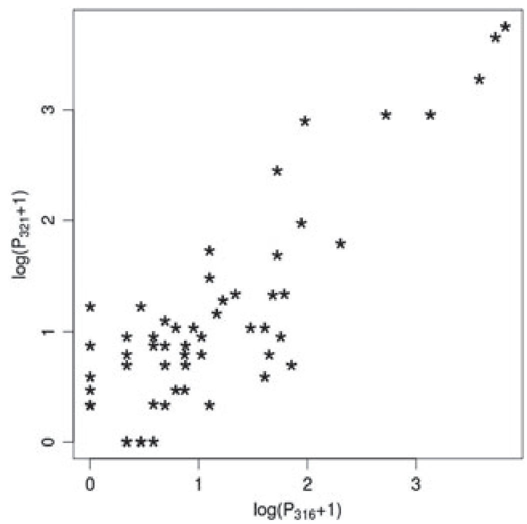


**Figure 3.** Correlation between the two rain gauges. The measured precipitation varies considerably from one rain gauge to the other.

### 3.2. Diffusion term proposals

Comparing the drift term of the SDE in Equation (16) with the drift term of the system equation in Equation (4), it is seen that the flow model can be rewritten

$$\mathrm{d}\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} = \begin{bmatrix} b_1(\boldsymbol{u}_t, \boldsymbol{\theta}) + a_1(\boldsymbol{\theta}) S_{1,t} \\ b_2(S_{1,t}, \boldsymbol{\theta}) + a_2(\boldsymbol{\theta}) S_{2,t} \end{bmatrix} \mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\omega}_t \tag{24}$$

where

$$a_i(\boldsymbol{\theta}) = a_i(K) = -\frac{2}{K} \qquad \text{for } i = 1, 2$$

$$b_1(\boldsymbol{u}_t, \boldsymbol{\theta}) = b_1(P_{316,t}, P_{321,t}, A, \alpha, a_0) = \alpha A P_{316,t} + (1-\alpha) A P_{321,t} + a_0$$

$$b_2(S_{1,t}, \boldsymbol{\theta}) = b_2(S_{1,t}, K) = \frac{2}{K} S_{1,t}$$

The observation equation remains the same for all model proposals and we refer to the grey-box model represented by Equations (4) and (5), where in the following, only the diffusion matrix $\boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})$ is modified to obtain an improved description of the flow uncertainty. The models are estimated on a 15 min time resolution.

  

Model 1 The first model proposal is a model where the diffusion term is considered constant corresponding to the model presented in Section 2.2. Model 1 is then represented by the diffusion matrix

$$\boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \tag{25}$$

and the diffusion parameters $(\sigma_1, \sigma_2)$ are estimated as described in Section 2.3. Because the diffusion in the model is state independent, no transformation of the states is required to estimate the model parameters.

Model 2 The drift term of the model is driven by transient rain events, implying that most of the time, the flow in the sewer system consists of wastewater flow only. In that case, the variance of the diffusion term is expected to be rather small, but when a rain event occurs, the variance is expected to increase significantly owing to the uncertainty in the actual rain input to the system. It is furthermore anticipated that the uncertainty increases with the magnitude of the rainfall, (both duration and magnitude), which is captured by state-dependent diffusion.

Introducing a state-dependent diffusion term has the desired implication that the diffusion is scaled with the state magnitude. This makes physical sense because the diffusion terms (especially the first one), primarily represent the uncertainty in the rain input, and therefore should not contribute any uncertainty to the output (the flow), in dry weather periods. Another implication is that the risk of receiving negative state values is avoided as discussed in Section 2.5. Model 2 is represented by the state proportional diffusion matrix

$$\boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{bmatrix} \sigma_1 S_{1,t} & 0 \\ 0 & \sigma_2 S_{2,t} \end{bmatrix} \tag{26}$$

With the addition of state dependence, it is expected that the diffusion parameters will be reduced because the state variation is adjusted with the state magnitude. The states in model 2 need to be transformed to avoid numerical instability and to be able to implement the model in CSTM. The transformed states in model 2 are identical to Equation (22) with $a_1$, $a_2$, $b_1$ and $b_2$ as defined in Equation (24).

Model 3 Because of the risk that the uncertainty intervals might become too large, it was decided to investigate a reduced state dependence and introduce a $\gamma_i$ parameter. More specifically, model 3 is expressed with the diffusion matrix

$$\boldsymbol{\sigma}(\boldsymbol{S}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{bmatrix} \sigma_1 S_{1,t}^{\gamma_1} & 0 \\ 0 & \sigma_2 S_{2,t}^{\gamma_2} \end{bmatrix} \tag{27}$$

Here, the $i$th diffusion term is assumed to be dependent on the $i$th state to the power of $\gamma_i$. The Lamperti transform is also required for model 3 because the diffusion is state dependent. The transformation is identical to Equation (20) with $a_1$, $a_2$, $b_1$ and $b_2$ as defined in Equation (24).

# 4. RESULTS

## 4.1. Searching for optimal $\gamma_i$ parameters in model 3

Because of instability-related problems with estimating the $\gamma_i$ parameters in model 3, an iterative approach had to be adopted to pinpoint the optimal $\gamma_i$ parameters. Repeatedly, the $\gamma_i$ parameters were adjusted and the corresponding log-likelihood value calculated in search of the maximum log-likelihood area. Figure 4 displays the resulting surface for the profile log-likelihood, varying with the two diffusion parameters $\gamma_1$ and $\gamma_2$.

Figure 4 shows that an increase in $\gamma_2$ causes a linear increase in the log-likelihood, implying that optimal diffusion parameter $\gamma_2$ is 1. A similar linear correspondence appears between the values of $\gamma_1$ and the log-likelihood, but for higher values of the parameter, the contour lines even out, meaning that a rather minor increase in the log-likelihood is obtained for further increases in $\gamma_1$. It should be recalled that $\gamma_1$ is important for controlling the variance of the modelled flow during rain because most of the uncertainty is expected to origin from an insufficient rain input. However, the argument for introducing the $\gamma_i$ parameters in the first place was to downsize the uncertainty boundaries, which might be important when a prediction horizon of more than one step is needed. Therefore, to test the influence on the uncertainty bounds (in this paper, only on the one step prediction) $(\gamma_1, \gamma_2) = (0.6, 0.95)$ was selected for further analysis with model 3.

## 4.2. Estimation results

Table 2 displays the mean and standard deviation of the estimated parameters. Considering the runoff parameters of the drift term ($A$, $K$ and $\alpha$), it is noticed that the model parameters differ considerably, particularly between models 1–3, even though the models differ solely with respect to the diffusion term. The drift term of the model remains the same in all three models, but the estimated drift term parameters still differ. This shows the importance of selecting a proper description of the diffusion term. The size of the contributing catchment area $A$ is estimated in the range of 35–51 $\cdot 10^4$ m$^2$, the retention time $K$ in the range of 3–5.3 h and the rain gauge weighting parameter $\alpha$ range between 0.3–0.4, that is to say, rain gauge $P_{321}$ represents most of the runoff. This is a little surprising because $P_{316}$ is located much closer to the largest paved area of the catchment (cf. Figure 2). Considering the estimated wastewater parameters ($a_0$,$s_1$, $s_2$,$c_1$,$c_2$), it is noticeable that all models returned similar values for $a_0$ and $c_2$, whereas the rest of the parameters differed.

Turning to the diffusion parameters, it is seen that for all three models, $\sigma_1$ are larger than $\sigma_2$, which is reasonable because the input uncertainty primarily can be assigned to $\sigma_1$. However, the model structure limitations can probably be equally attributed to both states, and thus,

  

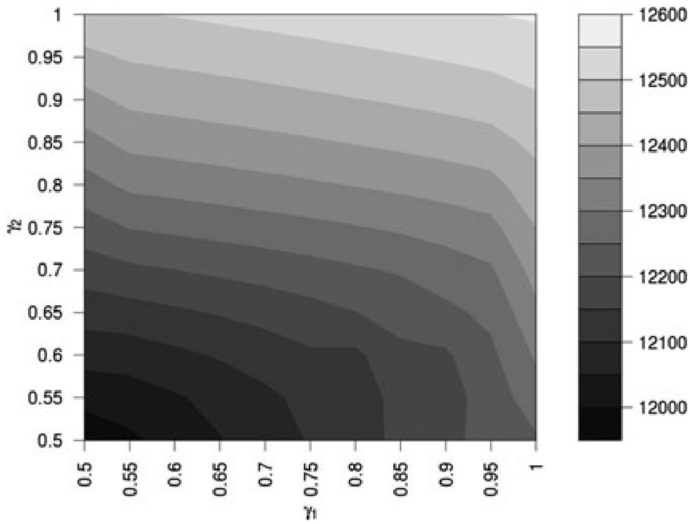

**Figure 4.** Contour plot of the log-likelihood as a function of the two diffusion parameters $\gamma_1$ and $\gamma_2$.

**Table 2.** Estimation results

| Parameter | Model 1 | | Model 2 | | Model 3 | |
|---|---|---|---|---|---|---|
| | $\hat{\theta}_{M1}$ | sd($\hat{\theta}_{M1}$) | $\hat{\theta}_{M2}$ | sd($\hat{\theta}_{M2}$) | $\hat{\theta}_{M3}$ | sd($\hat{\theta}_{M3}$) |
| $s_1$ | −46.641 | 5.288 | −65.645 | 2.876 | −63.580 | 2.824 |
| $c_1$ | −96.282 | 5.089 | −51.814 | 3.564 | −56.725 | 2.386 |
| $s_2$ | −48.185 | 3.528 | −35.459 | 1.882 | −39.047 | 1.544 |
| $c_2$ | 17.934 | 3.726 | 17.576 | 1.926 | 18.039 | 1.829 |
| $\log(A)$ | 3.567 | 0.035 | 3.934 | 0.060 | 3.856 | 0.064 |
| $\alpha$ | 0.398 | 0.056 | 0.305 | 0.081 | 0.269 | 0.034 |
| $a_0$ | 314.290 | 4.172 | 317.330 | 5.002 | 308.890 | 4.154 |
| $K$ | 2.999 | 0.068 | 5.286 | 0.220 | 5.261 | 0.202 |
| $\log(\sigma_1)$ | 5.240 | 0.031 | −1.414 | 0.052 | 1.107 | 0.050 |
| $\log(\sigma_2)$ | 3.053 | 0.072 | −2.444 | 0.011 | −2.082 | 0.010 |
| $\log(S)$ | −7.519 | 0.047 | −19.020 | 11.559 | −19.070 | 8.845 |

a significant $\sigma_2$ is found in all three models. The estimated diffusion parameters of the three models cannot be directly compared because in model 1, the diffusion parameters are constants, whereas for model 2 the diffusion parameters are scaled with the states and, for model 3, the state to the power of $\gamma_i$. This explains why a decrease of their values are realised with increasing state dependence. The variance of the observation noise $S$ is significant for model 1 and insignificant for models 2 and 3. This indicates that the state dependent models cannot separate uncertainty that originate from input and model structural errors from uncertainty that origins from flow measurement errors.

### 4.3. Model comparison and validation

Table 3 shows that for the one-step-ahead prediction, model 2 gives the best fit and uncertainty description according to the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). This means that the state proportional scaling of the diffusion parameters is the preferred diffusion term, although the scaling of the prediction bounds might become a problem if several prediction steps are needed. This is however not investigated in this paper, but will be examined in a future study.

Model validation is only considered for the best model (model 2). A structural behaviour in the residuals would suggest that more physics is needed in the drift term. Figure 5 displays the results of the residual analysis. From the standardised residual plot of model 2 shown in Figure 5(a), it seems that the Gaussian assumption is satisfied, because the residuals are randomly distributed around zero. Even though few data points appear to depart from the assumption, they are not considered to violate the Gaussianity.

Inspecting the autocorrelation functions in Figure 5(b) and (d), a minor significance for lags 2 and 3 is visible, but considered small enough to be neglected. However, it is also noticed from the ACF and PACF plots that there is a periodicity in the residual series; note the peaks around lag 96 and 672 corresponding to 1 day and 1 week, respectively. These values are also very small and thus ignored here, although they



  

**Table 3.** Model Comparison

| | $\log(L)$ | DF | AIC | BIC |
|---|---|---|---|---|
| Model 1 | 11379.81 | 13 | −22733.62 | −22643.12 |
| Model 2 | 12555.67 | 13 | −25085.34 | −24994.84 |
| Model 3 | 12461.81 | 15 | −24893.62 | −24789.19 |

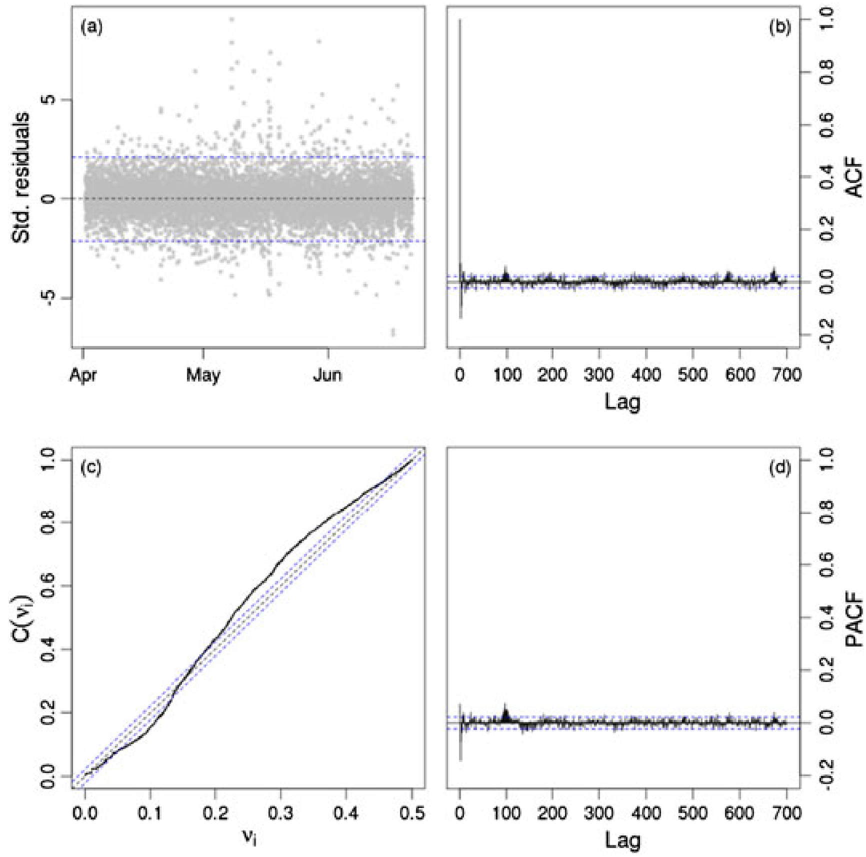


**Figure 5.** Model validation. (a) standardised residual plot; (b) autocorrelation function (ACF) (c) cumulative periodogram; and (d) partial autocorrelation function (PACF).

point to a need for further model development of the dry weather flow parameterisation. In the adopted modelling approach, no distinction between weekends (holidays) and working days or between consecutive working days was tested, although the wastewater diurnal pattern changes accordingly. Thus, the periodicity would be a good starting point to improve the dry weather part of the model, but this is beyond the scope of this paper.

The cumulative periodogram for the residuals is shown in Figure 5(c). For the residuals to be considered white, the black solid line should be close to the dashed diagonal line and within the two off-diagonal dashed lines, which correspond to 95% confidence limits for the assumed Gaussianity. In the plot, a minor periodicity is detected on each side of the straight line, but these effects are rather limited and can be ignored.

To sum up, the minor deviation for the residuals from the Gaussian assumption for the residuals does not give solid basis for model rejection, and model 2 can be considered sufficiently accurate for assessing the one step prediction uncertainties.

### 4.4. State and flow uncertainty in dry and wet weather periods

In Figure 6, a comparison of the 95% one step ahead prediction interval of the states is shown for a large rain event, (left column), and a dry weather period, (right column). Notice the scale difference of the vertical axis. For model 1, the prediction interval of the states remains constant in dry and wet weather and at one point encloses negative state volumes in dry weather. This shows why a state dependent diffusion term is needed. Furthermore, it is clearly seen that the prediction interval is wider for $S_1$ than $S_2$, which is related to the uncertain rain input that primarily influences $S_1$. The prediction interval of models 2 and 3 reveals that the lower boundary stays positive in dry weather and that



  

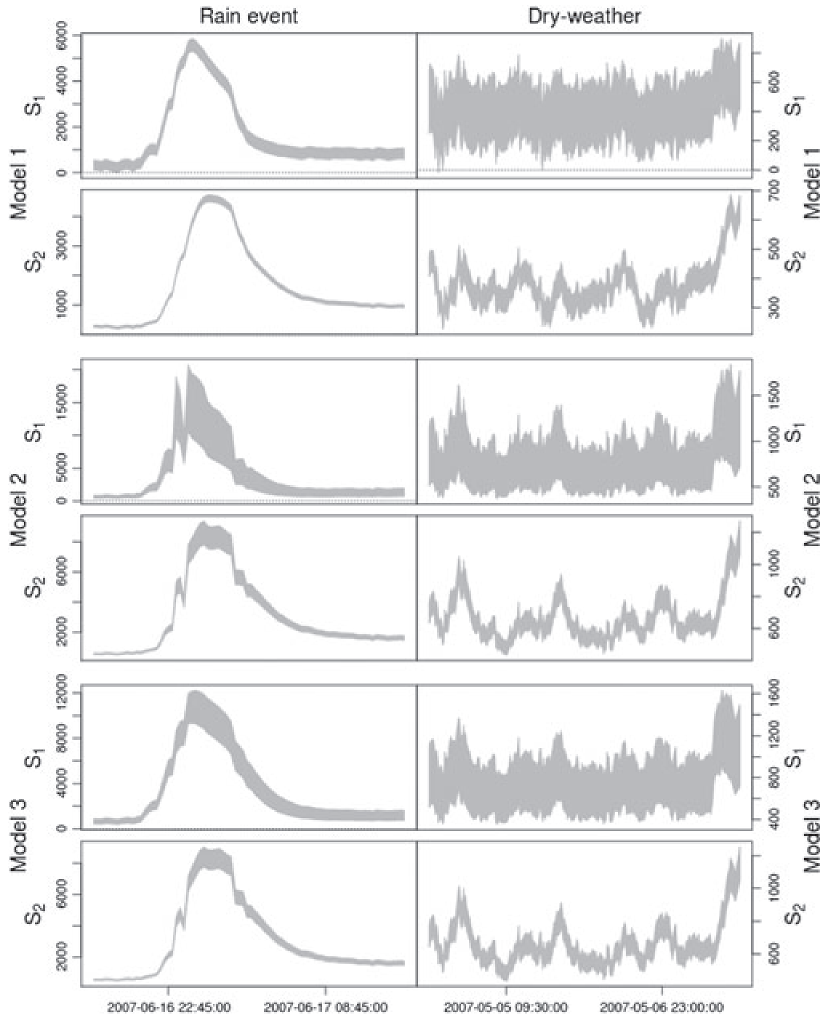


**Figure 6.** Ninety-five per cent state prediction intervals for all three considered models. State predictions during a rain event is displayed in the left column and in dry-weather in the right column.

the uncertainty increases considerably with the state magnitude; but as expected, less in model 3 than model 2. Generally, much more water is stored in the states of model 2 and 3 than was the case with model 1. This is reasonable because the estimated catchment area is larger for models 2 and 3 than for model 1. Moreover, the estimated retention time in models 2 and 3 is also larger, that is, to obtain the same average dry weather flow, a larger amount of water is stored in both states.

In the left column of Figure 7, the observed flow rate and the corresponding one step ahead 95% prediction interval are displayed for all three models during a rain event and in the right column during a dry weather period. The prediction interval for model 1 is seen to increase with flow magnitude, which is a consequence of scaling the variance of the observation noise $S$ with the observation function $\boldsymbol{h}$. The prediction interval of model 1 is the most narrow for large flows, whereas the opposite holds in dry weather periods. Comparing model 2 with model 3, the downsizing of the prediction interval is only recognised at the flow peak during rain. However, a longer prediction horizon would probably lead to a more substantial difference.

Considering how the model assimilates the observations, it can be shown that the observation noise plays an important role. In model 1, the belief in the drift term of the model is quite good as the updating of the states in the model is not overly aggressive. The predictions are clearly not tracking the latest observation whereas, in the case of models 2 and 3, the states are updated in accordance with the latest observation because observations are taken to be almost 100% accurate. The problem with identifying the observation noise is probably related to both inadequate rain inputs, as well as periods with poor or erroneous flow metre observations.

## 5. CONCLUSIONS

This study has shown that a simple grey-box model consisting of two linear reservoirs for rainfall–runoff flow and a harmonic function for wastewater flow can be successfully applied to model the one step prediction uncertainty when an appropriate diffusion term is identified.



  

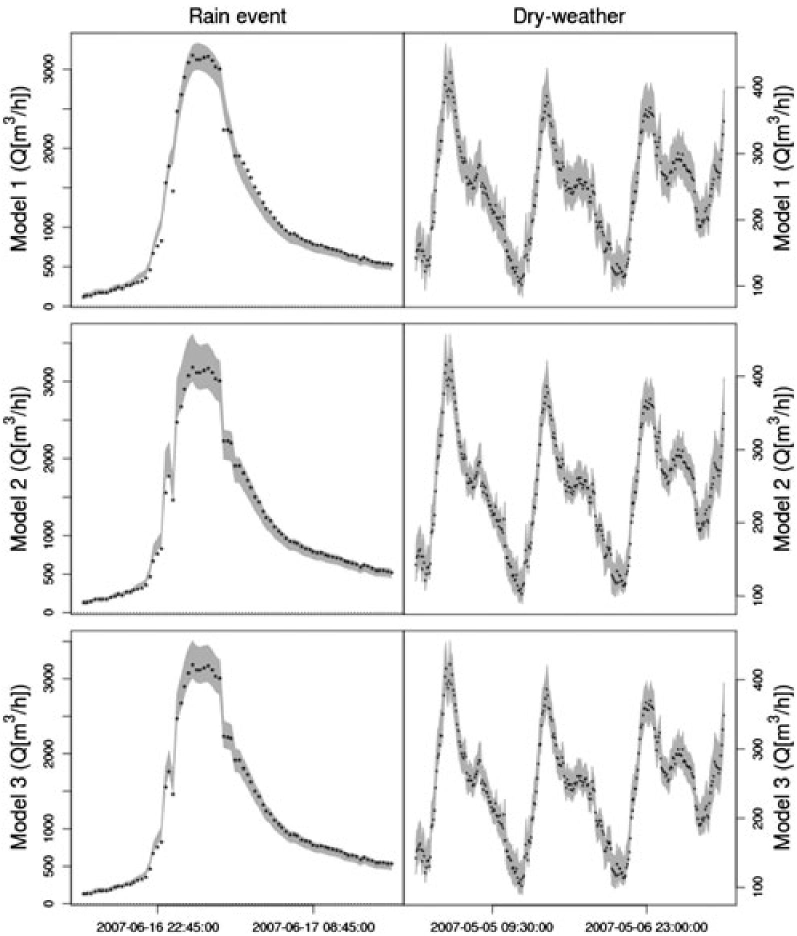


**Figure 7.** Ninety-five per cent flow prediction intervals (grey area) during wet-weather (left column) and dry-weather (right column) conditions for all three considered models. Measured values are displayed as stars.

Such a simple model is attractive for forecasting and control. Three different models were compared that differed with respect to the diffusion term formulation only: one with additive diffusion, one with state proportional diffusion and one with state exponentiated diffusion. To implement the state dependent transformations, it was necessary to apply Itô's formula and the Lamperti transformation. The state proportional diffusion was found to best and adequately describe the one step flow prediction uncertainty, whereas the additive diffusion term resulted in a violation of the physical constraints of the model states that are positively restricted. In a similar manner the risk of obtaining negative flows from an additive observation noise description was avoided by a logarithmic transformation of the observations. This ensured that the observation noise was scaled with the model output. Finally, it was found that a proper description of the diffusion term is important for estimation of the physical parameters.

## Acknowledgements

We appreciate the help and support of flow metre data from Spildevandscenter Avedöre I/S, and the help with the graphical layout by Lisbeth Brusendorf, DTU Environment. This research project was financially supported by a PhD fellowship co-funded by Krüger A/S, DTU Environment and the Ministry of Science, Technology and Innovation through the graduate school for Urban Water Technology (UWT), and by the Danish Council for Strategic Research (SWI).

## REFERENCES

Baadsgaard M, Nielsen JN, Spliid H, Madsen H, Preisel M. 1997. Estimation in stochastic differential equations with a state dependent diffusion term. *11th IFAC Symposium on System Identification* **3**: 1425–1430. Kitakyushu, Fokuoka.

Barbera PL, Lanza LG, Stagi L. 2002. Tipping bucket mechanical errors and their influence on rainfall statistics and extremes. *Water Science & Technology* **45**(2): 1–9.

Bechmann H, Nielsen MK, Madsen H, Poulsen NK. 1999. Grey-box modelling of pollutant loads from a sewer system. *Urban Water* **1**: 71–78.

  

Bechmann H, Madsen H, Poulsen N, Nielsen MK. 2000. Grey box modeling of first flush and incoming wastewater at a wastewater treatment plant. *Environmetrics* **11**: 1–12.

Bertrand-Krajewski JL, Bardin JP, Mourad M, Béranger Y. 2003. Accounting for sensor calibration, data validation, measurement and sampling uncertainties in monitoring urban drainage systems. *Water Science & Technology* **47**(2): 95–102.

Carstensen J, Nielsen MK, Strandbæk H. 1998. Prediction of hydraulic load for urban storm control of a municipal WWT plant. *Water Science & Technology* **37**(12): 363–370.

Deletic A, Dotto CBS, McCarthy DT, Kleidorfer M, Freni G, Mannina G, Uhl M, Henrichs M, Fletcher TD, Rauch W, Bertrand-Krajewski JL, Tait S. 2011. Assessing uncertainties in urban drainage models. *Physics and Chemistry of the Earth Parts A/B/C*. In Press. DOI:10.1016/j.pce.2011.04.007.

El-Din AG, Schmith DW. 2002. A neural network model to predict the wastewater inflow incorporating rainfall events. *Water Research* **36**(5): 1115–1126. DOI:10.1016/S0043-1354(01)00287-1.

Freni G, Mannina G. 2010. Uncertainty in water quality modelling: the applicability of variance decomposition approach. *Journal of Hydrology* **394**(3–4): 324–333.

Gelfan A, Hajda P, Novotny V. 1999. Recursive system identification for real-time sewer flow forecasting. *Journal of Hydrologic Engineering* **4**(3): 280–287.

Giraldo JM, Leirensa S, Díaz-Granados MA, Rodríguez J. 2010. Nonlinear optimization for improving the operation of sewer systems: the Bogotá case study. International Environmental Modelling and Software Society (iEMSs). In *2010 International Congress on Environmental Modelling and Software*. Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada. Swayne DA, Yang W, Voinov AA, Rizzoli A, Filatova T (eds). http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings.

Harremoës P, Madsen H. 1999. Fiction and reality in the modelling world-balance between simplicity and complexity, calibration and identifiability, verification and falsification. *Water Science & Technology* **39**(9): 1–8.

Iacus SM. 2008. *Simulation and Inference for Stochastic Differential Equations: With R Examples*. Springer: New York.

Jacobsen JL, Madsen H. 1996. Grey box modelling of oxygen levels in a small stream. *Environmetrics* **7**: 109–121.

Jacobsen JL, Madsen H, Harremoës P. 1997. A stochastic model for two-station hydraulics exhibiting transient impact. *Water Science & Technology* **36**(5): 19–26.

Jazwinski AH. 2007. *Stochastic Processes and Filtering Theory*. Dover Publications: Mineola, New York, USA.

Jonsdottir H, Jacobsen JL, Madsen H. 2001. A grey-box model describing the hydraulics in a creek. *Environmetrics* **12**: 347–356.

Jonsdottir H, Madsen H, Palsson OP. 2006. Parameter estimation in stochastic rainfall-runoff models. *Journal of Hydrology* **326**(1–4): 379–393.

Jonsdottir H, Nielsen HA, Madsen H, Eliasson J, Palsson OP. 2007. Conditional parametric models for storm sewer runoff. *Water Resources Research* **43**: 1–9.

Jørgensen HK, Rosenørn S, Madsen H, Mikkelsen PS. 1998. Quality control of rain data used for urban runoff systems. *Water Science & Technology* **37**(11): 113–120.

Kleidorfer M, Deletic A, Fletcher TD, Rauch W. 2009. Impact of input data uncertainties on urban stormwater model parameters. *Water Science & Technology* **60**(6): 1545–1554. DOI:10.2166/wst.2009.493.

Krämer S, Fuchs L, Verworn HR. 2007. Aspects of radar rainfall forecasts and their effectiveness for real time control—the example of the sewer system of the city of vienna. *Water Practice & Technology* **2**(2). Selected papers from the Sewer Operation and Maintenance conference held in Vienna, Austria, 26 – 28 October 2006, and the 7th Urban Drainage Modelling and 4th Water Sensitive Urban Design conferences, held concurrently in Melbourne, Australia, 2 – 7 April 2006. DOI:10.2166/wpt.2007.042.

Kristensen NR, Madsen H. 2003. *Continuous Time Stochastic Modeling - CTSM 2.3 - Mathematics Guide*, Technical Unversity of Denmark. *Technical Report DTU*, Lyngby, Denmark. http://www2.imm.dtu.dk/ctsm/MathGuide.pdf

Kristensen NR, Madsen H, Jørgensen SB. 2004a. A method for systematic improvement of stochastic grey-box models. *Computers and Chemical Engineering* **28**(8): 1431–1449.

Kristensen NR, Madsen H, Jørgensen SB. 2004b. Parameter estimation in stochastic grey-box models. *Automatica* **40**: 225–237.

Kuczera G, Kavetski D, Franks S, Thyer M. 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: characterising model error using storm-dependent parameters. *Journal of Hydrology* **331**(1–2): 161–177. DOI:10.1016/j.jhydrol.2006.05.010.

Lei JH, Schilling W. 1996. Preliminary uncertainty analysis—a prerequisite for assessing the predictive uncertainty of hydrologic models. *Water Science & Technology* **33**(2): 79–90.

Lindblom E, Madsen H, Mikkelsen PS. 2007. Comparative uncertainty analysis of copper loads in stormwater systems using GLUE and grey-box modeling. *Water Science & Technology* **56**(6): 11–18.

Luschgy H, Pagés G. 2006. Functional quantization of a class of Brownian diffsuions: a constructive approach. *Stochastic Processes and their Applications* **116**: 310–336.

Madsen H. 2008. *Time Series Analysis*. Chapman & Hall/CRC: London.

Mannina G, Freni G, Viviani G, Saegrov S, Hafskjold LS. 2006. Integrated urban water modelling with uncertainty analysis. *Water Science & Technology* **54**(6–7): 379–386. DOI:10.2166/wst.2006.611.

Molini A, Lanza LG, Barbera PL. 2005. The impact of tipping-bucket rain gauge measurement errors on design rainfall for urban-scale applications. *Hydrological processes* **19**: 1073–1088. DOI:10.1002/hyp.5646.

Møller JK, Madsen H. 2010. *From state dependent diffusion to constant diffusion in stochastic differential equations by the Lamperti transform*, IMM-Technical Report-2010-16, 25. DTU Informatics, Building 321, Kgs. Lyngby, Denmark. http://www.imm.dtu.dk/English/Service/Phonebook.aspx?lg=showcommon&id=271164

Nielsen HA, Madsen H. 2006. Modelling the heat consumption in district heating systems using a grey-box approach. *Energy and Buildings* **38**(1): 63–71.

Ocampo-Martinez C, Puig V. 2009. On modelling approaches for receding-horizon control design applied to large-scale sewage systems. *Proceedings of the IEEE Conference on Decision & Control, including the Symposium on Adaptive Processes*: Shanghai, PR China; 8052–8058. December 16-18.

Øksendal B. 2003. *Stochastic differential equations - an introduction with applications*, 6th ed. Springer: Berlin.

Pedersen L, Jensen NE, Christensen LE, Madsen H. 2010. Quantification of the spatial variability of rainfall based on a dense network of rain gauges. *Atmospheric Research* **95**(4): 441–454. DOI:10.1016/j.atmosres.2009.11.007.

Priestley MB. 1981. *Spectral Analysis and Time Series*, Series of Monographs and Textbooks. Academic Press: London.

Puig V, Cembrano G, Romera J, Quevedo J, Aznar B, Ramón G, Cabot J. 2009. Predictive optimal control of sewer networks using coral tool: application to Riera Blanca catchment in Barcelona. *Water Science & Technology* **60**(4): 347–354.

Shedekar VS, King KW, Brown LC, Fausey NR, Heckel M, Harmel RD. 2009. Measurement errors in tipping bucket rain gauges under different rainfall intensities and their implication to hydrologic models. *American Society of Agricultural and Biological Engineers Annual International Meeting 2009* **10**: 6429–6438.

Tan PC, Berger CS, Dabke KP, Mein RG. 1991. Recursive identification and adaptive prediction of wastewater flows. *Automatica* **27**(5): 761–768.

Tornøe CW, Jacobsen J, Pedersen O, Hansen T, Madsen H. 2004. Grey-box modelling of pharmacokinetic/pharmacodynamic systems. *Journal of Pharmacokinetics and Pharmacodynamics* **31**(5): 401–417.

Vaes G, Willems P, Berlamont J. 2005. Areal rainfall correction coefficients for small urban catchments. *Atmospheric Research* **77**(1–4): 48–59. DOI:10.1016/j.atmosres.2004.10.015.


  

Vestergaard M. 1998. Nonlinear filtering in stochastic volatility models. *Master Thesis*. Technical University of Denmark. Department of Mathematical Modelling, Lyngby, Denmark.

Willems P. 2001. Stochastic description of the rainfall input errors in lumped hydrological models. *Stochastic Environmental Research and Risk Assessment* **15**: 132–152. DOI:10.1007/s004770000063.

Willems P. 2010. Parsimonious model for combined sewer overflow pollution. *Journal of Environmental Engineering* **136**(3): 316–325. DOI:10.1061/(ASCE)EE.1943-7870.0000151.

Willems P, Berlamont J. 2002. Probabilistic emission and immission modelling case study of the combined sewer—WWTP—receiving water system at Dessel (Belgium). *Water Science & Technology—WST* **45**(3): 117–124.

  