Stoch Environ Res Risk Assess (2012) 26:1151–1162
DOI 10.1007/s00477-012-0563-3

ORIGINAL PAPER

Evaluation of probabilistic flow predictions in sewer systems using grey box models and a skill score criterion

Fannar Örn Thordarson · Anders Breinholt · Jan Kloppenborg Møller · Peter Steen Mikkelsen · Morten Grum · Henrik Madsen

Published online: 29 January 2012

Abstract In this paper we show how the grey box methodology can be applied to find models that can describe the flow prediction uncertainty in a sewer system where rain data are used as input, and flow measurements are used for calibration and updating model states. Grey box models are composed of a drift term and a diffusion term, respectively accounting for the deterministic and stochastic part of the models. Furthermore, a distinction is made between the process noise and the observation noise. We compare five different model candidates' predictive performances that solely differ with respect to the diffusion term description up to a 4 h prediction horizon by adopting the prediction performance measures; reliability, sharpness and skill score to pinpoint the preferred model. The prediction performance of a model is reliable if the observed coverage of the prediction intervals corresponds to the nominal coverage of the prediction intervals, i.e. the bias between these coverages should ideally be zero. The sharpness is a measure of the distance between the lower and upper prediction limits, and skill score criterion makes it possible to pinpoint the preferred model by taking into account both reliability and sharpness. In this paper, we illustrate the power of the introduced grey box methodology and the probabilistic performance measures in an urban drainage context.

F. Ö. Thordarson (✉) · J. K. Møller · H. Madsen
DTU Informatics, Bldg. 305, 2800 Kgs. Lyngby, Denmark
e-mail: ft@imm.dtu.dk

A. Breinholt · P. S. Mikkelsen
DTU Environment, Bldg. 113, 2800 Kgs. Lyngby, Denmark

M. Grum
Krüger, Veolia Water Solutions and Technologies, Gladsaxevej
363, 2860 Søborg, Denmark

Keywords Grey box modelling · Interval prediction · Reliability · Sharpness · Skill score · Urban drainage

1 Introduction

Sewer flow predictions can, in combination with Model Predictive Control (MPC), be used to minimise damages in a broad sense, e.g. to reduce combined sewer overflows to prevent sludge escaping from wastewater treatment plants and to avoid flooding of vulnerable urban areas. To the authors knowledge, most, if not all, the suggested MPC solutions that have been proposed in the literature to date are based on deterministic models, (see e.g. Ocampo-Martinez and Puig 2010; Puig et al. 2009; Giraldo et al. 2010), even though it is commonly accepted that large uncertainties are present in simulation and prediction with urban drainage models due to unreliable level or flow meters (Bertrand-Krajewski et al. 2003), non-representative rainfall inputs (Pedersen et al. 2010; Vaes et al. 2005; Willems 2001) and/or unreliable rain gauge measurements (Barbera et al. 2002; Molini et al. 2005; Shedekar et al. 2009).

For urban drainage systems, we are still awaiting this shift of paradigm from deterministic to stochastic models in predictive control. This can most likely be attributed to inadequate measurement collection, both with respect to rainfall monitoring/forecasting and in-sewer flow or level metering. However, as the number of measurement devices increase and these devices become more accurate, the potential for building suitable stochastic models also improves. A necessary first step is to derive stochastic models that can describe the predictive uncertainty sufficiently well for a certain prediction horizon of interest. Another important step is to set up a prediction performance evaluation method to be able to compare the predictive performance of different model candidates. In this paper we intend to take these necessary first steps by considering a case catchment area from where both rainfall and flow meter measurements are available for stochastic model building and prediction evaluation of sewer flows.

We apply the grey box methodology as introduced by Kristensen et al. (2004a). The grey box approach is based on a state space model where the dynamics are described using Stochastic Differential Equations (SDEs), which contain a drift term and a diffusion term. The grey box methodology has been successfully applied in numerous fields for stochastic model building, including e.g. pharmacology (Tornøe et al. 2006), chemical engineering (Kristensen et al. 2004a; Kristensen et al. 2004b), district heating (Nielsen and Madsen 2006), hydrology (Jonsdottir et al. 2001, 2006) and ecology (Møller et al. 2011). We give particular attention to the diffusion term by considering various diffusion term descriptions. Several tools have been developed to validate and compare models, especially for point forecasts that exclusively rely on the single value prediction. In contrast, little attention has been given to interval predictions, which play a crucial role in stochastic control design. We propose here to use a skill scoring criterion for interval prediction evaluation, and show how this can be applied to find the preferred model among the candidate models for a specific prediction horizon. The skill scoring criterion has previously been applied for prediction evaluation purposes in wind power generation (see Pinson et al. 2007; Møller et al. 2008).

In Sect. 2, we outline the stochastic grey box methodology. Section 3 includes a description of the interval prediction generation and how the prediction performance can be evaluated on the basis of the reliability, the sharpness and the skill score criterion. Section 4 illustrates the applicability of the grey box methodology and the use of the prediction performance criteria as important tools for model selection. Finally, in Sect. 5 we conclude on our findings.

2 The stochastic grey box model

2.1 Model structure

The model used in this study is a grey box model, or a continuous-discrete time stochastic state space model, represented by

$$d\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \quad (1)$$

$$\boldsymbol{Y}_k = \boldsymbol{g}(\boldsymbol{X}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k, \quad (2)$$

where the first equation is called the system equation, composed of a set of SDEs in continuous time. The states are partially observed in discrete time through the observation Eq. 2. The time is $t \in \mathbb{R}_0$ and t_k (for $k = 1, \ldots, K$) are the discretely observed sampling instants for the K available measurements. The states in the system equation $\boldsymbol{X}_t \in \mathbb{R}^n$ describe the system dynamics in continuous time, whereas $\boldsymbol{X}_k \in \mathbb{R}^n$ in the observation equation is the observed states in the discrete time as specified by the observations. The input variables are represented by the vector $\boldsymbol{u}_t \in \mathbb{R}^m$ and the vector of the measured output variables $\boldsymbol{Y}_k \in \mathbb{R}^l$. The vector $\boldsymbol{\theta} \in \mathbb{R}^p$ includes the unknown parameters that characterise the model, and the functions $\boldsymbol{f}(\cdot) \in \mathbb{R}^n, \boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{g}(\cdot) \in \mathbb{R}^l$ form the structural behaviour of the model. The measurement error $\boldsymbol{e}_k$ is assumed to be a l-dimensional white noise process with $\boldsymbol{e}_k \sim N(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta}))$, where $\boldsymbol{V}$ is the covariance of the measurement error, and $\boldsymbol{\omega}_t$ is a n-dimensional standard Wiener process. The first term in the system equation is the drift term, representing the dynamic structure of the system that is formulated by ordinary differential equations. The second term is the diffusion term which corresponds to the process noise related to the particular state variable in the state-space formulation.

Discrepancies between output from deterministic models and measurements are often referred to as measurement errors, even though the consecutive residuals are clearly auto-correlated. In reality, these auto-correlated discrepancies can however be explained by both non-representative and/or faulty inputs as well as model structural deficiencies. Consequently, a distinction between measurement noise and noise related to inputs and model deficiencies is required. The stochastic grey box model provides such a distinction by separating the process noise from the output measurement noise, where the process noise as described by the diffusion term is related to the state variables and accounts for noise that is not related to the output measurements.

2.2 Parameter estimation and state transformation

For parameter estimation the Maximum Likelihood (ML) method is used, and the Kalman Filter techniques are applied to evaluate the likelihood function (Jazwinski 2007). For the grey box model in Eqs. 1 and 2, the unknown model parameters are obtained by maximising a likelihood function that is a product of the one-step conditional densities (Madsen 2008). Hence, the estimated parameters for an adequate model correspond to a fit where the distribution for the residual series for the one-step ahead prediction error is assumed to be serial independent and Gaussian. However, utilising such a model for predictions covering more than one-step ahead usually results in a residual series that is correlated, and when dealing with increasing prediction horizon, the predictive distribution for the output may divert from the assumed normality.

To estimate the unknown parameters of the model, the software CTSM[1] (Kristensen and Madsen 2003) is used. The software is well suited for estimation of linear and many nonlinear systems. In CTSM, the ordinary Kalman filter gives the exact solution for the state estimation for linear systems, whereas the extended Kalman filter provides an approximation for the states for nonlinear systems.

Parameter and state estimation is not possible with CTSM if state dependency is included in the diffusion term, as this requires higher order filtering techniques to solve the estimation than are available in the extended Kalman filter techniques implemented in the software (Vestergaard 1998). However, efficient and numerically stable estimates can be obtained by considering a transformation of the states. In particular, the transformation is well-suited for a SDE when the diffusion term is only dependent on the corresponding state variable. With such a univariate diffusion, it is always possible to transform the state description to obtain a state independent diffusion term (Baadsgaard et al. 1997).

The transformation of the ith state variable $X_{i,t}$ to $Z_{i,t}$, for $i = 1, \ldots, n$, is referred to as the Lamperti transform (Iacus 2008) and, subsequently, a corresponding SDE for the transformed variable $Z_{i,t}$, is obtained by Itô's formula (Øksendal 2003). The diffusion in the transformed SDE is state independent and the transformed grey box model is rewritten

$$d\boldsymbol{Z}_t = \tilde{\boldsymbol{f}}(\boldsymbol{Z}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \tilde{\boldsymbol{\sigma}}(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \tag{3}$$

$$\boldsymbol{Y}_k = \tilde{\boldsymbol{g}}(\boldsymbol{Z}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k, \tag{4}$$

where the functions $\boldsymbol{f}(\cdot), \boldsymbol{\sigma}(\cdot)$ and $\boldsymbol{g}(\cdot)$ in Eqs. 1 and 2 have been reformulated, respectively to $\tilde{\boldsymbol{f}}(\cdot), \tilde{\boldsymbol{\sigma}}(\cdot)$ and $\tilde{\boldsymbol{g}}(\cdot)$ in relation to the transformation of the state space. The parameters $\boldsymbol{\theta}$ and the input-output relations are, however, not affected by the transformation.

In this study, it is furthermore anticipated that flow measurement errors increase proportionally with flow magnitude and thus a log-transformation of the observations are needed to secure a Gaussian measurement noise term. This observation transformation results in an observation equation that has an additive noise term (Limpert et al. 2001).

3 Prediction, uncertainty and evaluation

3.1 Uncertainty of h-step ahead prediction

The objective with the proposed grey box model is to predict the sewer flow at time $k + h$, which is denoted as Y_{k+h}. In parallel, we have $\hat{Y}_{k+h|k}$ as the prediction of the flow at time $k + h$, given the available information at time k where h indicates the number of time steps for the prediction. By using the ML method, we find that the optimal prediction is equal to the conditional mean for the model structure (see Madsen 2008). Hence, the prediction is obtained by

$$\hat{\boldsymbol{Y}}_{k+h|k} = E\left[\boldsymbol{Y}_{k+h} | \boldsymbol{\Upsilon}_k, \boldsymbol{u}_{k+h}\right] \tag{5}$$

$$\hat{\boldsymbol{Y}}_{k+h|k} = \tilde{\boldsymbol{g}}(\hat{\boldsymbol{Z}}_{k+h|k}, \boldsymbol{u}_{k+h}, t_{k+h}, \boldsymbol{\theta}), \tag{6}$$

meaning that for a given sequence of precipitation input up to time $k + h$ and observed flow up to time $k, \boldsymbol{\Upsilon}_k = [\boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_0]^\top$, the state prediction at time $k + h$ can be estimated and consequently supply the observation equation with a suitable description for the prediction. The challenge in predicting the future flow in the system is then not directly related to predictions based on the observation equation, but rather on predicting the state variables in the system equation. The state prediction can be accomplished by considering the conditional expectation of the future state:

$$\hat{\boldsymbol{Z}}_{k+h|k} = E\left[\boldsymbol{Z}_{k+h} | \hat{\boldsymbol{\Upsilon}}_k, \boldsymbol{u}_{k+h}\right], \tag{7}$$

i.e. the conditional mean of $\boldsymbol{Z}_{k+h}$ given all measurements up to time k (Madsen 2008).

In the following study, the grey box model in Eqs. 1 and 2 is used to describe the model structure, whereas the transformed model is used for parameter estimation and model prediction in Eqs. 3 and 4. As mentioned in Sect. 2, the Gaussian assumption for the model output is only valid for one-step ahead predictions. Thus for $h \geq 1$, a numerical approach is considered, i.e. an Euler scheme for the SDEs in the system Eq. 3 is applied to predict the sewer runoff (Kloeden and Platen 1999). Thus, a sufficient probability distribution for the h-step ahead prediction is obtained by generating a number of simulations from each time step, and from this empirical distributions can be derived for the prediction intervals.

3.2 Prediction intervals

The ideal coverage of the prediction interval is defined as the nominal coverage $1 - \beta, \beta \in [0, 1]$. The upper and lower limits of the interval prediction are obtained from quantile forecasts, which are easy to obtain with a large

[1] Continuous-Time Stochastic Modelling - http://www.imm.dtu.dk/ctsm

number of simulations provided for the same prediction horizon, resulting in a reasonable empirical probability distribution for the sewer flow. If $F_{k+h|k}$ is the cumulative distribution function of the random variable $\hat{Y}_{k+h|k}$ and $\tau \in [0, 1]$ is the proportion of the relative quantile, the τ-quantile forecast for the $k + h$ prediction is obtained by

$$q^{(\tau)}_{k+h|k} = F^{-1}_{k+h|k}(\tau). \tag{8}$$

If $l = \beta/2$ and $u = 1 - \beta/2$ are defined as the lower and upper quantiles for the prediction interval at level $1 - \beta$, respectively, the prediction interval for the lead time $k + h$, issued at time k, can be described as

$$\hat{I}^{(\beta)}_{k+h|k} = \left[\hat{q}^{(l)}_{k+h|k}, \hat{q}^{(u)}_{k+h|k}\right] \tag{9}$$

where $\hat{q}^{(l)}_{k+h|k}$ and $\hat{q}^{(u)}_{k+h|k}$ are, respectively, the lower and upper prediction limits at levels $\beta/2$ and $1 - \beta/2$ (Pinson et al. 2007; Møller et al. 2008).

3.3 Reliability

For the prediction interval to be of practical usage for decision makers it is a primary requirement for the interval to be reliable, indicating that the upper and lower limits have to correspond to the nominal coverage rate of $1 - \beta$.

To obtain an evaluation of the reliability of the interval we define a counter that rewards prediction intervals that are able to capture the observations. For a given prediction interval, as represented in Eq. 9, and corresponding measured flow in the system Y_{k+h}, the binary indicator variable $n^{(\beta)}_{k,h}$ is obtained by

$$n^{(\beta)}_{k,h} = \begin{cases} 1, & \text{if } Y_{k+h} \in \hat{I}^{(\beta)}_{k+h|k} \quad \text{for } k \le K - h \\ 0, & \text{otherwise} \end{cases}, \tag{10}$$

corresponding to hits and misses of the h-step prediction interval. The mean of the binary series then corresponds to the actual proportion of hits in the estimation period, i.e. for prediction horizon h the proportion of hits for a flow series of length K, is given by

$$\bar{n}^{(\beta)}_h = \mathrm{E}\left[n^{(\beta)}_{k,h}\right] = \frac{1}{K-h}\sum_{k=1}^{K-h} n^{(\beta)}_{k,h}. \tag{11}$$

The discrepancy between the nominal coverage and the observed proportion of hits is measured by the bias

$$b^{(\beta)}_h = 1 - \beta - \bar{n}^{(\beta)}_h, \tag{12}$$

where a perfect fit is defined as $b^{(\beta)}_h = 0$, i.e. that the empirical coverage is equal to the nominal coverage, $\bar{n}^{(\beta)}_h = 1 - \beta$, and a perfect reliability is obtained. However, when the empirical coverage is larger than the nominal, i.e. $\bar{n}^{(\beta)}_h > 1 - \beta$, we talk about an overestimation in the coverage. This means that, since the empirical coverage is subtracted from the nominal coverage, we obtain $b^{(\beta)}_h < 0$ when the predictions overestimate the coverage. When the opposite is the case, this is referred to as underestimation, i.e. $b^{(\beta)}_h > 0$.

3.4 Sharpness

Sharpness is an accuracy measure of the prediction interval where smaller values indicate that the model is better suited to generate predictions (Gneiting et al. 2007). The size of the interval prediction, issued at time k for lead time $k + h$ is measured as the difference between the corresponding upper and lower quantile forecast, and averaging over the whole time series, defines the average sharpness. For the horizon h and coverage $1 - \beta$, the sharpness is calculated by

$$\bar{\delta}^{(\beta)}_h = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{q}^{(u)}_{k+h|k} - \hat{q}^{(l)}_{k+h|k}\right) \tag{13}$$

and by calculating $\bar{\delta}^{(\beta)}_h$ at relevant coverages, a δ-diagram can be viewed to summarise the evaluation of the sharpness. When comparing interval predictions generated from different models, the one with the smallest distance between upper and lower bound is the sharpest.

3.5 Interval score criterion and resolution

The skill score combines the performance measures discussed above in a single numerical value, which enables us to compare the predictive performance of different models directly. The skill score for interval predictions is outlined in detail by Gneiting and Raftery (2007), where the score of the individual prediction interval is also referred to as an interval score. The skill score Sc for the interval prediction, at time instant k, is calculated as

$$\begin{aligned} Sc^{(\beta)}_{I,k,h} = {} & (\hat{q}^{(u)}_{k+h|k} - \hat{q}^{(l)}_{k+h|k}) \\ & + \frac{2}{\beta}(\hat{q}^{(l)}_{k+h|k} - Y_{k+h})\mathbf{1}\{Y_{k+h} < \hat{q}^{(l)}_{k+h|k}\} \\ & + \frac{2}{\beta}(Y_{k+h} - \hat{q}^{(u)}_{k+h|k})\mathbf{1}\{Y_{k+h} > \hat{q}^{(u)}_{k+h|k}\}, \end{aligned} \tag{14}$$

where the indicator $\mathbf{1}\{\cdot\}$ is equal to one if the inequality within the brackets is fulfilled, but zero otherwise. As the objective is to evaluate the predictive performance of each model by a single number, an extension is required to account for the whole considered period. Hence, we average the scores for all time instants where observations are available, and thus the score becomes independent of the length of the time series. The average interval score criterion for h-step prediction is written

$$\begin{aligned}\overline{Sc}^{(\beta)}_{I,h} &= \frac{1}{K}\sum_{k=1}^{K} Sc^{(\beta)}_{I,k,h} = \bar{\delta}^{(\beta)}_h \\ &+ \frac{2}{\beta(K-h)}\sum_{k=1}^{K-h}\Big[(\hat{q}^{(l)}_{k+h|k} - Y_{k+h})\mathbf{1}\{Y_{k+h} < \hat{q}^{(l)}_{k+h|k}\} \\ &+ (Y_{k+h} - \hat{q}^{(u)}_{k+h|k})\mathbf{1}\{Y_{k+h} > \hat{q}^{(u)}_{k+h|k}\}\Big]. \qquad (15)\end{aligned}$$

It follows from Eq. 15 that for any observation that falls outside the predefined prediction interval, the skill score is increased by the distance between the interval and the observation at each considered quantile. Hence, the skill score gives a positive penalisation, which indicates that an increase in the score criterion will result in a reduced fit of the prediction interval. Therefore, we select the prediction interval with the lowest skill score.

The indication of the individual observation in relation to the prediction interval can be merged into an indicator, corresponding to the reliability indicator in Eq. 10. Thus, the interval score in Eq. 15 can be written as an indirect function of the prediction interval in Eq. 9 by including the reliability indicator from Eq. 10, i.e.

$$\begin{aligned}\overline{Sc}^{(\beta)}_{I,h} &= \bar{\delta}^{(\beta)}_h + \frac{2}{\beta(K-h)}\sum_{k=1}^{K-h}\left(1 - n^{(\beta)}_{k,h}\right) \\ &\times \left(\min\left|Y_{k+h} - [\hat{q}^{(l)}_{k+h|k}, \hat{q}^{(u)}_{k+h|k}]\right|\right), \qquad (16)\end{aligned}$$

where the second term under the summation accounts for the minimum distance between the observed value and the prediction interval, which is always either the lower or the upper limit of the interval.

The score is still a function of the prediction horizon h. This indicates that there are just as many $\overline{Sc}^{(\beta)}_{I,h}$ as there are h's. To evaluate the performance independently of h, we simply average over all horizons, obtaining the interval score criterion $\overline{\overline{Sc}}^{(\beta)}_{I}$.

We talk about resolution when conditioning the predictive distributions on some particular property. For urban drainage systems, it is expected that the skill score (or the sharpness and reliability) depends on the weather, i.e. the predictive performance is assumed to be different in periods of dry weather than in periods of wet weather.

4 Application results

In the previous sections, the model framework and tools for assessing the uncertainty and the performance of the model have been described. In the following we introduce the catchment area and the data, the applied grey box models, and finally present and discuss our results.

4.1 Description of the case study

The considered catchment area, which receives both wastewater and rainfall-runoff, is located in the Municipality of Ballerup west of Copenhagen in Denmark; see Fig. 1. It is connected to the second largest wastewater treatment plant in Denmark, located in Avedøre. Flow was measured downstream from the catchment area with a semi-mobile ultrasonic Doppler type flow meter. The flow meter was placed in an interceptor pipe with a dimension of

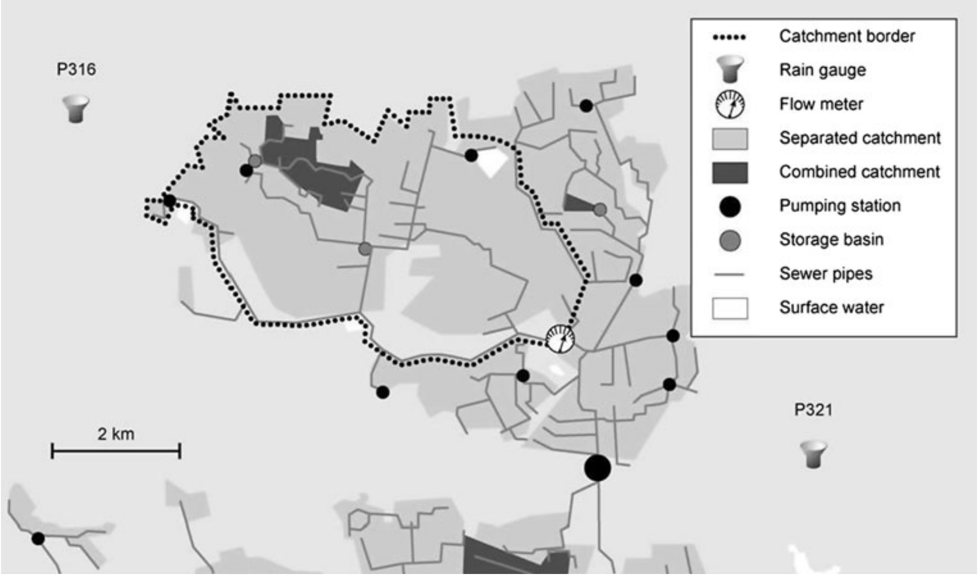

Fig. 1 The Ballerup catchment area

1.4 m. The flow meter logs every 5 min, but in this study a temporal resolution of 15 min was considered and, thus, only every third available measurement is used.

Precipitation was measured using two tipping bucket gauges with a volumetric resolution of 0.2 mm. The rain gauges are located just outside the considered catchment area, approximately 12 km apart from each other (Fig. 1). Data of flows and rain for almost 3-month period were used in the case study, i.e. from April 1 2007 to June 21 2007. The considered grey box models were estimated for all 3 months. However for prediction uncertainty assessment only data from May and June were utilised as very few rain events were logged by the rain gauges in April, and because the rain periods are the most important, it was decided to leave out this month. When generating predictions with the models we used the measured precipitation up to 4 h ahead of current time assuming a perfect rain forecast was available. This assumption is obviously unrealistic but serves an illustrative purpose here by showing how the skill score terminology can be applied to select the preferred model.

4.2 The stochastic model

The model should be kept simple and identifiable from data to facilitate the parameter estimation. In hydrology it is well known that the rainfall-runoff relationship can often be modelled with a series of linear reservoirs (e.g. Jacobsen et al. 1997; Mannina et al. 2006; Willems, 2010). A model with just two reservoirs is considered here, where the volume in each reservoir corresponds to a state variable in the grey box model. There is also a contribution of wastewater from the connected households to the sewer flow that needs to be accounted for. The model is written as

$$d\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} = \begin{bmatrix} \alpha A P_{1,t} + (1-\alpha) A P_{2,t} + a_0 - \frac{2}{K} S_{1,t} \\ \frac{2}{K} S_{1,t} - \frac{2}{K} S_{2,t} \end{bmatrix} dt + \begin{bmatrix} \sigma_1 S_{1,t}^{\gamma_1} & 0 \\ 0 & \sigma_2 S_{2,t}^{\gamma_2} \end{bmatrix} d\omega_t, \quad (17)$$

$$\log(Y_k) = \log\left(\frac{2}{K} S_{2,k} + D_k\right) + e_k, \quad (18)$$

where D_k is the wastewater flow variation formulated as a periodic function with diurnal cycles of length L, i.e.

$$D_k = \sum_{i=1}^{2} \left(s_i \sin \frac{i2\pi k}{L} + c_i \cos \frac{i2\pi k}{L} \right) \quad (19)$$

and s_1, s_2, c_1 and c_2 are parameters. The first reservoir $S_{1,t}$ receives runoff from the contributing area A at time t, caused by the rainfall registered at the two rain gauges $P_{1,t}$ and $P_{2,t}$. A weighting parameter α is defined to account for the fraction of the measured runoff that can be attributed to rain gauge $P_{1,t}$, whereas the remaining $1 - \alpha$ is attributed to $P_{2,t}$ assuming that the rainfall input area A is fully described by the two rain gauges. The second reservoir, $S_{2,t}$, receives outflow from the first reservoir and diverts it to the flow gauge downstream from the catchment.

To fully account for the wastewater flow in the grey box model, a constant term for the average dry-weather flow a_0 is included. The constant enters the first state to secure the physical interpretation of the system, i.e. water is always passing through the system, also in dry weather, which means that the reservoirs always contain water. From a modelling point of view this is important because the state variance from the diffusion term in Eq. 17—if large enough—could lead to predicted states that are negative, which is physically impossible. This risk of receiving negative states is especially high if an additive diffusion term is used and therefore we focus on state dependent diffusion terms only; see Breinholt et al (2011) for more details. When rainwater enters the system, the volume of water in the reservoir increases and the diffusion term is scaled accordingly (see Eq. 17), which means that the state prediction uncertainty rises.

The observation Eq. 18 depends on the second state variable only, since the output from the second reservoir corresponds to the flow measured downstream from the catchment area. The observation equation is log-transformed to account for proportional observation variance as mentioned in Sect. 2. In the following we will investigate various state dependencies through the γ parameter in each state dependent diffusion in the system Eq. 17. Different γ parameters will produce different prediction intervals and, subsequently, different skill scores. This is useful for model prediction comparison.

The diffusion parameters γ_1 and γ_2 are restricted to $\gamma_i \in [0.5, 1]$, for $i = 1{,}2$ in the system equation. The reasons are that for $\gamma_i \leq 0.5$ there is a positive probability of reaching zero and the risk of obtaining a non-stationary diffusion process is increased, whilst for $\gamma_i > 1$ the system existence and uniqueness is not guaranteed because the behaviour of the solution might explode in finite time (Iacus 2008).

Five models are proposed with different combinations of the diffusion parameters γ_1 and γ_2. These are (0.5, 0.5), (1, 0.5), (0.5, 1), (0.75, 0.75) and (1,1). The minimum γ parameter is actually slightly higher than 0.5 (i.e. 0.5001) in order to fulfill the parameter restriction, but for practical reasons is rounded to 0.5 in the text below. It is not possible to estimate the γ parameters with CTSM because each combination of γ parameters has its own restricted $Z_{i,t}$ domain. To distinguish between the models, they have been designated "M1", "M2", etc., as in the first line in Table 1; the corresponding sets of γ parameters are indicated in the next two rows (highlighted in bold).

Table 1 The results from the parameter estimation, for various values of (γ_1, γ_2), for all five models

θ	Unit	M1	M2	M3	M4	M5
γ_1	–	**0.500**	**1.000**	**0.500**	**0.750**	**1.000**
γ_2	–	**0.500**	**0.500**	**1.000**	**0.750**	**1.000**
s_1	–	−59.355	−65.313	−63.303	−63.909	−65.545
		(3.927)	(3.861)	(2.764)	(3.310)	(2.709)
s_2	–	−41.363	−34.090	−39.143	−37.341	−34.904
		(2.537)	(3.049)	(1.989)	(2.377)	(2.133)
c_1	–	−61.618	−49.407	−56.898	−51.593	−50.884
		(4.321)	(8.038)	(3.169)	(4.062)	(3.397)
c_2	–	17.437	17.120	18.407	17.133	17.785
		(2.537)	(2.927)	(1.913)	(2.220)	(1.889)
a_0	m^3/h	313.310	345.510	307.000	314.390	319.080
		(4.321)	(1.217)	(4.524)	(5.263)	(5.686)
α	–	0.359	0.374	0.288	0.334	0.335
		(0.068)	(0.080)	(0.070)	(0.059)	(0.067)
A	ha	42.406	39.694	49.591	46.479	51.413
		(1.059)	(1.221)	(1.062)	(1.080)	(1.104)
K	h	4.253	4.104	5.237	4.763	5.221
		(0.148)	(0.472)	(0.200)	(0.201)	(0.274)
σ_1	–	6.510	0.373	5.866	1.313	0.254
		(1.042)	(1.078)	(1.051)	(1.048)	(1.050)
σ_2	–	2.186	1.817	0.087	0.449	0.085
		(1.027)	(1.079)	(1.010)	(1.016)	(1.011)

Standard deviance is indicated in brackets

4.3 Estimation results

The parameter estimation is shown in Table 1. It is seen that the choice of diffusion term description affects all the parameters to some extent. However, the dry weather parameters s_1, s_2, c_1, c_2 and a_0 are not noticeably influenced, even though a_0 is slightly higher in M2 than it is in the other models. Considering the wet weather parameters A, K and α, it is seen that A and K are positively correlated with γ_2 while α is estimated to have more or less the same value. The largest area is estimated with M5. Regarding the estimates for the diffusion parameters σ_1 and σ_2, a higher expected parameter value follows a lower state dependency.

4.4 Overall reliability assessment

The average reliability bias is studied in Fig. 2, both as a function of the nominal coverage (Fig. 2a), and as a function of the prediction horizon of up to 4 h ahead (Fig. 2b). In Fig. 2a the reliability bias is calculated as the average for all the considered prediction steps, whereas in Fig. 2b, the reliability bias is calculated as an average of all the nominal coverages. No definite deviation is observed between the models, neither at the chosen prediction steps, nor at different nominal coverages. At coverage up to 80–90%, Fig. 2a shows that all five models slightly overestimate the nominal coverage, whereas for higher nominal coverage the bias is underestimated. Furthermore, the models approach the nominal coverage at around 85–90%.

Regarding the reliability bias for the individual models, Fig. 2a reveals that M1 deviates the most from the ideal as it exhibits the largest positive bias at intermediate coverage rates, and the most negative bias at higher nominal coverage rates. M2 is the most reliable model on average, the average bias from ideal reliability is −0.01 for all coverage rates up to 95% coverage. This indicates that ($\gamma_1 = 1$, $\gamma_2 = 0.5$) provides the best reliability across all the considered horizons.

Turning to the average reliability bias as a function of the prediction horizon, Fig. 2b shows that all five models produce almost the same reliability structure; i.e. for shorter horizons the reliability bias of the model predictions is overestimated, whereas for horizons longer than 1.5 h reliability is increasingly underestimated. Thus, the almost identical shift from overestimation to underestimation implies that all the models are reliable at 1.5 h lead time, but it is recalled that this is an average for all nominal coverages and, thus, it can vary for each nominal coverage. In contrast to what was concluded from Fig. 2a, the most

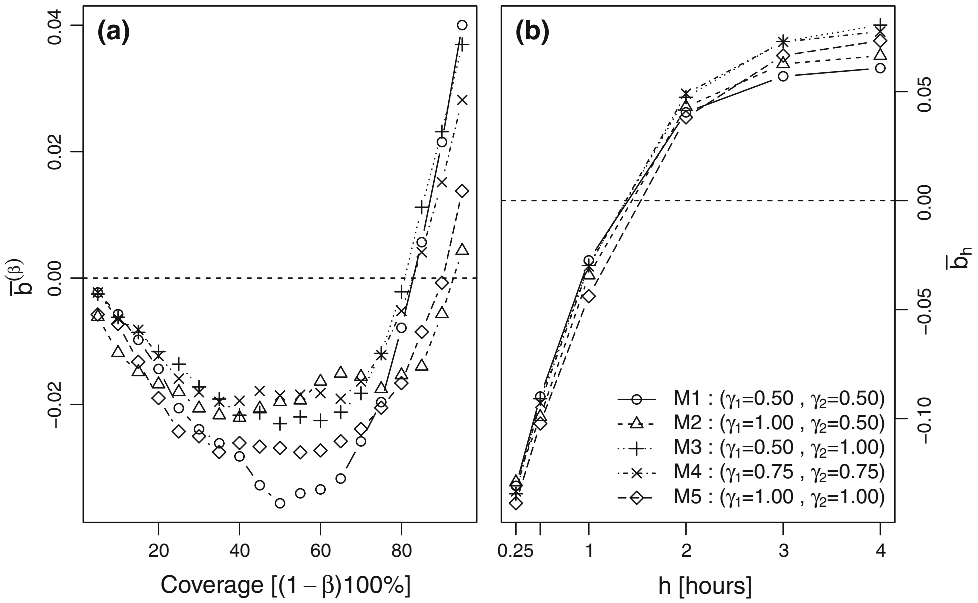

Fig. 2 Reliability bias for all five models of interest: **a** averaged over the entire prediction horizon, plotted as a function of the nominal coverage rate, **b** averaged over the coverage rates for each prediction step considered in the study. Coverage rates calculated for the nominal coverage rates: {5%, 10%...95%}

reliable model in Fig. 2b is M1. However, differences in reliability bias between the models are very small, suggesting that the minor discrepancies for the longer horizons are unimportant.

Here, a single nominal coverage is chosen for further investigation. From the reliability assessment above, it was detected that, on average, the 85–90% coverages are reliable. Therefore, the 90% coverage is selected for further investigation, which is also a typical value for interval prediction within hydrology.

4.5 Performance evaluation of the 90% prediction interval

In Fig. 3, the reliability bias of the 90% prediction interval ($\beta = 0.1$) as a function of the prediction horizon is seen. The same shift in reliability from overestimation to underestimation is observed for all models as the prediction horizon increases. The deviation from the nominal coverage is generally not that big, although M3 deviates almost 10% at the 4 h prediction step. On average, M1 is the most reliable model with mean distance from ideal reliability of 0.043. This can be hard to envisage from Fig. 3, because at larger prediction horizons, i.e. more than 1 h, M1 is clearly less reliable than M2 and M5.

In Fig. 4, the sharpness of the 90% prediction intervals is plotted for all the models as a function of the prediction horizon. As expected, all models become less sharp with increasing prediction horizon, i.e. the uncertainty of the prediction rises, but only up to 2 h. Hereafter the uncertainty levels out. When considering all prediction horizons, M2 is the least sharp model (the one with the largest uncertainty), and already at the 0.5 h prediction step it deviates considerably from the other models. Figure 4 also reveals that the models with $\gamma_1 = 0.5$ prove to be the sharpest for all prediction horizons, and M3 is visually slightly sharper than M1. Thus, M3 provides the sharpest average 90% prediction interval (187.3 m^3/h), whereas M2 provides the least sharp average prediction interval (286.3 m^3/h).

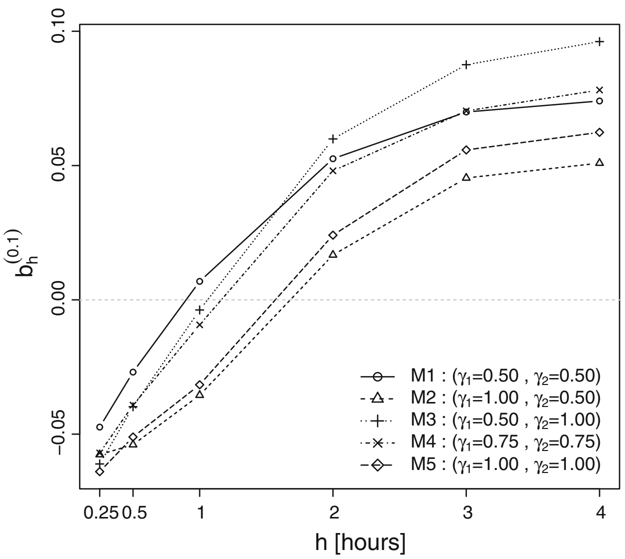

Fig. 3 Reliability bias of the 90% prediction interval, as a function of the prediction horizon

From studying the reliability and the sharpness it is not immediately clear which model should be preferred. However, this can be unravelled by calculating the skill score for each model for every prediction step and as an average for the entire prediction horizon. Table 2 shows the skill score for the generated 90% prediction intervals

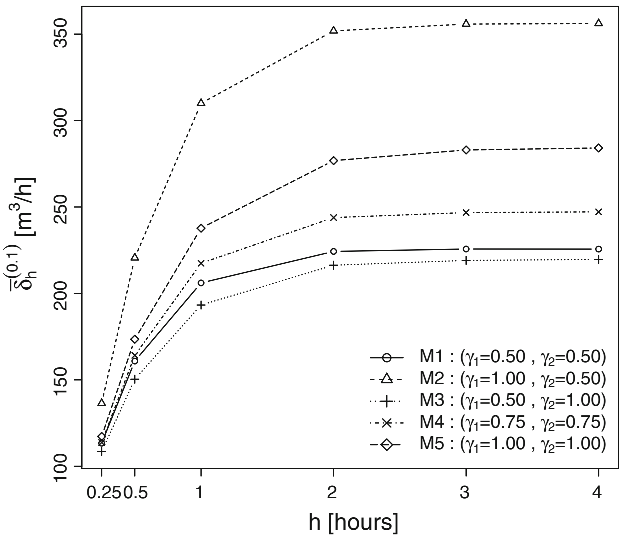

Fig. 4 Sharpness for the 90% prediction intervals, as a function of the prediction horizon for all five models

calculated for various prediction steps, and as an average for the maximum prediction horizon of 4 h. Note that all 16 prediction steps (every 15 min for 4 h) are included in the average skill score but only 6 prediction steps are presented in Table 2. M3 is seen to perform best at prediction steps 0.25 and 0.5 h (recalling that the smaller skill score is the preferred score), while M5 (the model with state proportional dependency for both states) performs best at larger prediction horizons up to 4 h. Surprisingly, the most reliable model M1 is seen to perform rather poorly compared to the other models for the prediction horizons of 1–4 h. Apparently, the sharpness for M1 is too narrow because many observations fall too far away from the lower and upper prediction bounds incurring a high penalty when calculating the skill score. When considering the average skill score for the entire prediction horizon of 4 h, it is furthermore seen that M2–M4 perform rather similarly, whereas M1 has a significantly higher score value.

4.6 Resolution analysis: conditioning on dry and wet weather periods

From a MPC point of view it is especially of interest to evaluate how well the models perform during wet weather periods. Separation of wet weather flow measurements from dry weather flow measurements using a rough flow threshold, i.e. wet weather interpreted as flows above 540 m^3/h and dry weather flows below, a conditional reliability is obtained as shown in Fig. 5. By introducing this threshold, 90% of the flow data is catagorised in the dry weather period and the remaining 10% in the wet weather period. For shorter prediction steps, the dry weather reliability (see Fig. 5a) is overestimated, whereas it is underestimated for longer prediction steps. This shift in reliability was also observed in the unconditional case seen in Fig. 3, and thus emanates from dry weather periods. In wet weather periods, the underestimated reliability increases with the length of the prediction horizon; see Fig. 5b. The only exception appears at the one-step prediction (0.25 h), where M3 and M5 both are reliable. At the 4 h prediction step the reliability bias is around 50–75%, compared with just 10% in the unconditional case. The models with $\gamma_1 = 1$ (M2, M5) are significantly less biased than the remaining models, but still underestimate the coverage by approximately 50% at the 4 h prediction step. This observed discrepancy in reliability bias between the unconditional case and the wet weather periods reveals the importance of the resolution analysis, and show that the relatively low reliability bias at the 4 h prediction horizon for the unconditional case is a result of the dry weather period, constituting 90% of the whole data set.

The conditional sharpness is shown in Fig. 6. In dry weather periods (Fig. 6a) the sharpness is very close to the unconditional sharpness, albeit slightly more sharp. In wet weather periods (Fig. 6b), the sharpness decreases considerably, i.e. the prediction intervals are approximately twice the size in dry weather periods. It is seen that the prediction uncertainty grows rapidly during the first prediction steps and then levels out at 2 h. The effect of the

Table 2 Skill score calculated from 90% prediction intervals at several prediction steps and averaged for the entire prediction horizon of 4 h

	γ_1	γ_2	Prediction horizon						
			0.25 h	0.5 h	1 h	2 h	3 h	4 h	Average
M1	0.50	0.50	166.0	292.7	491.2	680.8	724.7	732.7	514.7
M2	1.00	0.50	201.9	324.9	455.2	563.6	602.6	610.1	459.7
M3	0.50	1.00	**137.2**	**228.3**	391.2	603.8	675.8	691.7	454.7
M4	0.75	0.75	155.1	264.3	429.7	606.4	663.6	673.6	465.4
M5	1.00	1.00	150.4	247.1	**383.8**	**535.2**	**593.8**	**608.2**	**419.7**

The preferred model candidate for each prediction horizon is highlighted in bold

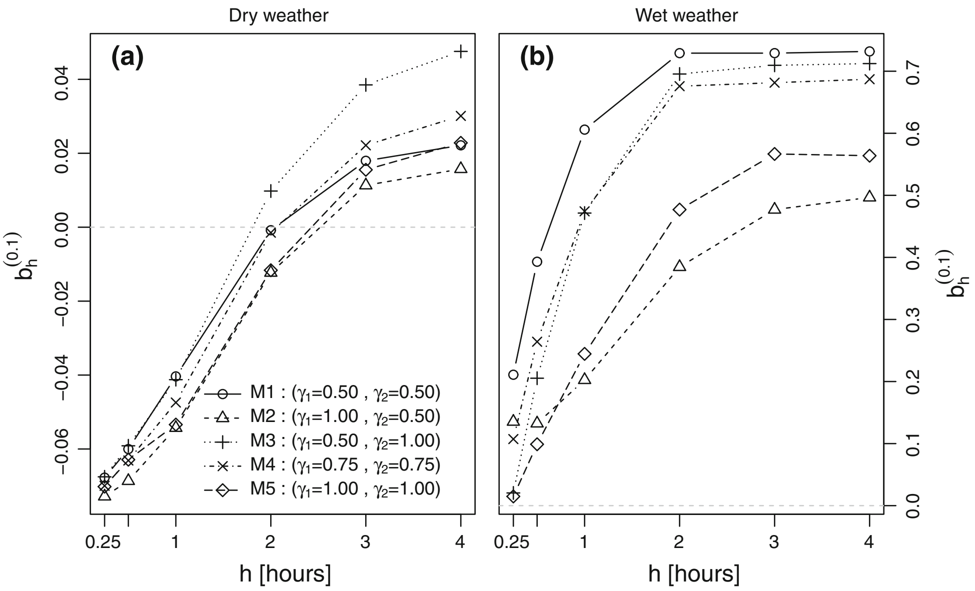

Fig. 5 Reliability of the 90% prediction intervals, as a function of the prediction horizon and conditioned on the weather: **a** for dry weather periods; **b** for wet weather periods. A flow threshold of 540 m^3/h was applied for conditioning

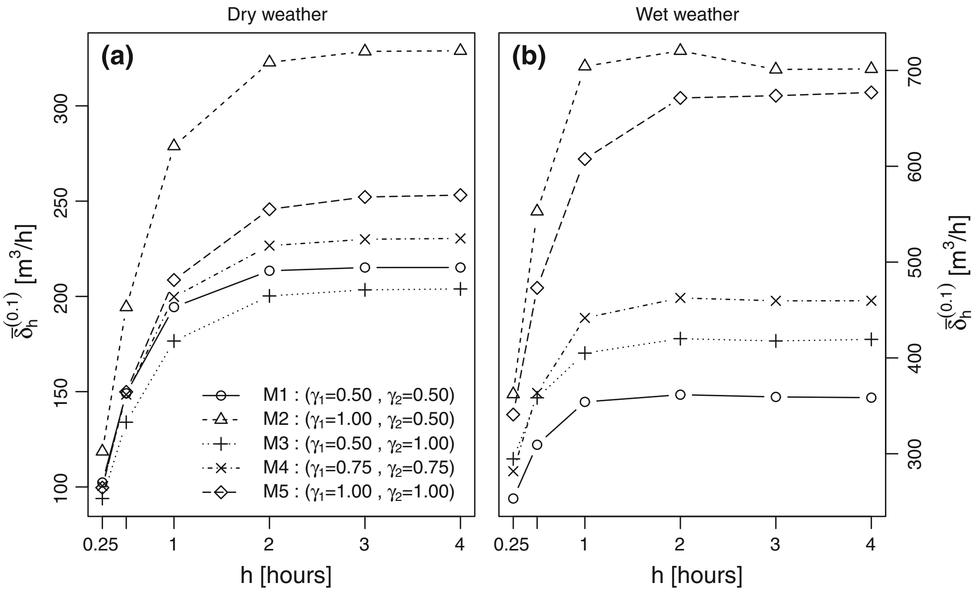

Fig. 6 Sharpness of the 90% coverage, as a function of the prediction horizon and conditioned on the flow: **a** for dry weather periods; **b** for wet weather periods. A flow threshold of 540 m^3/h was applied

diffusion term is clearly identified. The models M2 and M5 are seen to be the least sharp, but both models have state proportional diffusion in the first reservoir ($\gamma_1 = 1$). In contrast, the models M1 and M3, with $\gamma_1 = 0.5$, generate the sharpest prediction intervals.

The dry weather conditional skill score for the five model candidates is seen in Table 3. It is readily seen that M3 is the preferred model candidate both at each prediction step and as an average for the entire prediction horizon. As the reliability bias was found to be close to zero at all considered prediction steps, we conclude that M3 is very useful for making 90% prediction intervals in dry weather periods.

When conditioning on wet weather periods alone, Table 4 yields more ambiguous results. M3 is the best model for prediction steps of less then 1h, which is the same as obtained when conditioning on dry weather periods alone. However, for 1–4 h, models M2 and M5 provide better results (lower skill score). Note the large difference in average skill score between dry and wet weather periods when comparing Tables 3 and 4. The best model on average when considering all prediction horizons of interest is M5, but it should be kept in mind that the reliability bias showed that none of the models are able to generate satisfactory 90% prediction intervals, and thus cannot be fully trusted when considering prediction horizons larger than one. If focusing on the one-step ahead prediction only in wet weather periods, M3 must be the preferred model; both because it was shown to be reliable and because it has the lowest skill score.

Table 3 Skill score calculated for the 90% prediction interval conditioned on dry weather periods

	γ_1	γ_2	Prediction horizon						
			0.25 h	0.5 h	1 h	2 h	3 h	4 h	Average
M1	0.50	0.50	68.3	114.2	176.4	225.0	236.5	238.8	176.5
M2	1.00	0.50	74.6	128.5	198.9	253.3	266.8	269.2	198.6
M3	0.50	1.00	**62.2**	**101.9**	**159.5**	**214.3**	**233.2**	**237.7**	**168.1**
M4	0.75	0.75	65.7	109.6	170.8	224.4	240.9	244.4	176.0
M5	1.00	1.00	64.2	106.9	166.7	222.4	241.7	246.9	174.8

The preferred model candidate for each prediction horizon is highlighted in bold

Table 4 Skill score calculated for the 90% prediction interval conditioned on wet weather periods

	γ_1	γ_2	Prediction horizon						
			0.25 h	0.5 h	1 h	2 h	3 h	4 h	Average
M1	0.50	0.50	397.3	709.9	1243.0	1859.4	1996.8	2015.7	1370.4
M2	1.00	0.50	572.9	820.9	1044.5	**1328.0**	**1439.9**	**1456.1**	1110.4
M3	0.50	1.00	**289.1**	**489.3**	913.1	1607.8	1825.9	1874.5	1166.6
M4	0.75	0.75	372.3	631.8	1051.3	1619.2	1785.3	1815.3	1212.5
M5	1.00	1.00	365.5	583.0	**903.5**	1374.9	1547.9	1583.5	**1059.7**

The preferred model candidate for each prediction horizon is highlighted in bold

The resolution study has clearly demonstrate the importance of conditioning the drainage model performance on relative weather situations; in our case rain. When considering the model performance on the whole data series, altogether it appeared as though the best model is able to provide quite reliable 90% prediction limits. However, when conditioning separately on wet weather periods it becomes clear that even the best model is unable to generate reliable prediction limits beyond 0.25 h. This can primarily be ascribed to a poor rain input that does not represent the actual rainfall on the whole catchment area. If the rain input used in the models is improved by, e.g., placing rain gauges inside the catchment area or by using rain radars a different description for the diffusion term in the model would be preferred and a larger prediction horizon would probably be shown to be reliable. With more representative rain input, it is possible to extend the diffusion term by considering both the states and the rain input in its description, which would contribute to more reliable probabilistic predictions.

5 Conclusions

This study has demonstrated how simple stochastic models suitable for making interval flow predictions in urban drainage systems can be built using the grey box methodology, and the models capabilities for providing interval predictions evaluated by the performance measures: reliability, sharpness and skill score. Reliability concerns the coverage ratio of the prediction intervals that must correspond to the nominal coverage, sharpness concerns the size of the prediction interval, and finally the skill score utilises both reliability and sharpness to evaluate the prediction performance in a single score value. This is useful for model prediction comparison. Grey box models are tailored to derive the one-step prediction interval, but can, presuming a representative rain input is given and the model describes the processes well, be used to make interval predictions several time steps into the future, given that the interval predictions are reliable.

Five different grey box models, that only differed with respect to the diffusion term description, were estimated and their probabilistic prediction performance was evaluated using data from a case catchment area. A model was found that was able to predict the 90% flow prediction interval up to 4 h ahead when all the observations were included in the study. The skill score criterion was applied to compare the prediction performance of the models and eventually to select the preferred model. However, when conditioning the model performance on wet weather periods (accounting for 10% of the whole data series), it was shown that solely the one-step prediction (15 min) was reliable. This can most likely be attributed to a poor rain input that does not represent the actual rainfall on the catchment area very well. In a control context, since wet weather periods are the most important periods, more representative rain inputs and rain forecasts are needed to derive models that can reliably describe the prediction uncertainty several time steps into the future. Nevertheless, this particular case study should not detract from the power of the proposed methodology.

Acknowledgements We appreciate the help and support of flow meter data from Spildevandcenter Avedøre I/S. The research was funded by a PhD fellowship, including DTU Informatics and DTU Environment, and by the Danish Strategic Research Council (Sustainable Energy and Environment Programme).

References

Baadsgaard M, Nielsen JN, Spliid H, Madsen H, Preisel M (1997) Estimation in stochastic differential equations with a state dependent diffusion term. SYSID’97—11th IFAC symposium of system identification, IFAC

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

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 Sci Technol 47(2):95–102

Breinholt A, Thordarson FÖ, Møller JK, Mikkelsen PS, Grum M, Madsen H (2011) Grey box modelling of flow in sewer systems with state dependent diffusion. Environmetrics 22(8):946–961

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction and estimation. J Am Stat Assoc 102(477):359–378

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. J R Stat Soc B 69(2):243–268

Giraldo JM, Leirens S, Díaz-Grenados MA, Rodríguez JP (2010) Nonlinear optimization for improving the operation of sewer systems: the Bogotá Case Study. International Environmental Modelling and Software Society (iEMSs). 2010 International Congress on Environmental Modelling and Software Modelling for Environments Sake, Fifth Biennial Meeting, Ottawa, Canada

Iacus SM (2008) Simulation and Inference for stochastic differential equations—with R examples. Springer series of Statistics

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

Jazwinski AH (2007) Stochastic processes and filtering theory. Dover Publications, Mineola, NY

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. J Hydrol 326(1–4):379–393

Kloeden P, Platen E (1999) Numerical solutions of stochastic differential equations. Springer

Kristensen NR, Madsen H (2003) Continuous time stochastic modeling—CTSM 2.3—mathematics guide. Technical Unversity of Denmark

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

Kristensen NR, Madsen H, Jørgensen SB (2004b) A method for systematic improvement of stochastic grey-box models. Comput Chem Eng 28(8):1431–1449

Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the sciences: keys and clues. BioScience 51(5):341–352

Madsen H (2008) Time series analysis. Chapman & Hall/CRC

Mannina G, Freni G, Viviani G, Saegrov S, Hafskjold L (2006) Integrated urban water modelling with uncertainty analysis. Water Sci Technol—WST 54(6–7):379–386

Møller JK, Nielsen HA, Madsen H (2008) Time-adaptive quantile regression. Comput Stat Data Anal 52:1292–1303

Møller JK, Madsen H, Carstensen J (2011) Parameter estimation in a simple stochastic differential equation for phytoplankton modelling. Ecol Model 222:1793–1799

Molini A, Lanza LG, Barbera Pl (2005) The impact of tipping-bucket raingauge measurement errors on design rainfall for urban-scale applications. Hydrol Process 19:1073–1088

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

Ocampo-Martinez C, Puig V (2010) Piece-wise linear functions-based model predictive control of large-scale sewage systems. IET Control Theory Appl 4(9):1581–1593

Øksendal B (2003) Stochastic differential equations—an introduction with applications, 6th edn. Springer

Pedersen L, Jensen NE, Christensen LE, Nielsen HA, Madsen H (2010) Quantification of the spatial variability of rainfall based on a dense network of rain gauges. Atmospheric Res 95(4):441–454

Pinson P, Nielsen HA, Møller JK, Madsen H (2007) Non-parametric probabilistic forecasts of wind power: required properties and evaluation. Wind Energy 10(6):497–516

Puig V, Cembrano G, Romera J, Quevedo J, Aznar B, Ramoń G, Cabot J (2009) Predictive optimal control of sewer networks using CORAL tool: application to Riera Blanca catchment in Barcelona. Water Sci Technol 60(4):869–878

Shedekar VS, King KW, Brown LC, Fausey NR, Heckel M, Harmel DR (2009) Measurement errors in tipping bucket rain gauges under different rainfall Intensities and their implication to hydrologic models. Conf. paper, ASABE Annual International Meeting, June 21–24, pp 1–9

Tornøe CW, Jacobsen J, Pedersen O, Hansen T, Madsen H (2006) Grey-box Modelling of pharmacokinetic/pharmacodynamic systems. J Pharmacokinet Pharmacodyn 31(5):401–417

Vaes G, Willems P, Berlamont J (2005) Areal rainfall correction coefficients for small urban catchments. Atmospheric Res 77(1–4):48–59

Vestergaard M (1998) Nonlinear filtering in stochastic volatility models. Master’s 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. Stoch Environ Res Risk Assess 15:132–152

Willems P (2010) Parsimonious model for combined sewer overflow pollution. J Environ Eng 136(3):316–325