Stoch Environ Res Risk Assess (2014) 28:505–516
DOI 10.1007/s00477-013-0768-0

ORIGINAL PAPER

Stochastic rainfall-runoff forecasting: parameter estimation, multi-step prediction, and evaluation of overflow risk

Roland Löwe · Peter Steen Mikkelsen · Henrik Madsen

Published online: 27 July 2013

Abstract Probabilistic runoff forecasts generated by stochastic greybox models can be notably useful for the improvement of the decision-making process in real-time control setups for urban drainage systems because the prediction risk relationships in these systems are often highly nonlinear. To date, research has primarily focused on one-step-ahead flow predictions for identifying, estimating, and evaluating greybox models. For control purposes, however, stochastic predictions are required for longer forecast horizons and for the prediction of runoff volumes, rather than flows. This article therefore analyzes the quality of multistep ahead forecasts of runoff volume and considers new estimation methods based on scoring rules for k-step-ahead predictions. The study shows that the score-based methods are, in principle, suitable for the estimation of model parameters and can therefore help the identification of models for cases with noisy in-sewer observations. For the prediction of the overflow risk, no improvement was demonstrated through the application of stochastic forecasts instead of point predictions, although this result is thought to be caused by the notably simplified setup used in this analysis. In conclusion, further research must focus on the development of model structures that allow the proper separation of dry and wet weather uncertainties and simulate runoff uncertainties depending on the rainfall input.

Keywords Stochastic greybox model · Skill score · Real-time control · Urban drainage · Multistep prediction · Online forecasting

1 Introduction

Real-time control (RTC) often provides a method to efficiently operate sewer systems and reduce spills of sewage into lakes, rivers, and oceans (combined sewer overflows, CSOs). This reduces the need to build storage volumes in the sewer system, which makes the method economically attractive. A multitude of control systems are in operation today. The types of setup range from rule-based strategies that are determined offline (Fuchs and Beeneken 2005; Seggelke et al. 2012), to online optimizations of storage volumes (Pabst et al. 2011) and model predictive control (MPC) (Schütze et al. 2004; Puig et al. 2009).

It is commonly expected that the combination of forecast information and global optimization as applied in MPC will yield the best control results. This is obscured by the complex side constraints that result from the operational requirements in the sewer system and by insufficient forecast quality. Recently, a new control setup was introduced in the Copenhagen area to minimize the total overflow risk from a number of storage basins in the catchment through the dynamic adjustment of the basin outflows and the pumping capacities. The decisions in this algorithm for the global control of the system are based on forecasted runoff volumes for the catchment of each basin (dynamic overflow risk assessment) (Vezzaro and Grum 2012; Grum et al. 2011).

R. Löwe (✉) · H. Madsen
Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Matematiktorvet B322, 2800 Kongens Lyngby, Denmark
e-mail: rolo@dtu.dk

P. S. Mikkelsen
Department of Environmental Engineering, Technical University of Denmark (DTU), Miljøvej B113, 2800 Kongens Lyngby, Denmark

Forecasts in such a setup need to be available at varying horizon lengths which makes models that provide multistep predictions attractive. Furthermore, the forecast uncertainties need to be considered in the decision-making process because prediction-risk relationships in urban drainage systems are typically nonlinear (Vezzaro and Grum 2012). However, no tools for the modeling of predictive uncertainties in an online setting are available. At present, the very simplified assumption that forecast uncertainties can be described by a Gamma distribution with shape parameters that depend on the predicted runoff volume is used.

Stochastic greybox models fulfill both requirements because these provide predictive uncertainties at varying horizons. For our purposes, stochastic greybox models are termed simplified models with physically interpretable parameters that provide a quantification of the model uncertainties. Several authors have demonstrated the general applicability of this class of models to urban drainage problems. Carstensen et al. (1998) applied ARMAX models to simulate the inflow to a wastewater treatment plant. Bechmann et al. (2000) simulated the first flush and later the pollutant loads (Bechmann et al. 1999) using stochastic differential equations. Breinholt et al. (2011) investigated model setups for flow predictions based on linear reservoir cascades using stochastic differential equations and took the initial steps required to quantify the predictive uncertainty. Furthermore, Thordarson et al. (2012) investigated multistep flow predictions for urban drainage systems and evaluated these using skill score criteria.

Previous works on stochastic forecasting of runoff in urban drainage systems have focused on the prediction of flows for one or several prediction horizons. However, the decision-making process in RTC is typically based on the predicted runoff volume. The quality of the probabilistic multistep volume predictions obtained from the stochastic greybox models has not yet been evaluated. Furthermore, it is unclear whether the currently used parameter estimation technique, which is based on the maximization of the likelihood for one-step ahead predictions, also yields a good model for multistep-ahead forecasts.

Therefore, following the scheme shown in Fig. 1, the stochastic multistep predictions of the runoff volume are generated using greybox models. New estimation approaches for stochastic greybox models that focus on multistep predictions were suggested, and the forecasts from the resulting models were compared.

A simplified assessment of the ability of the models to predict the overflow risk was subsequently performed to evaluate the possible effects of the different forecasts on RTC.

2 Methods

2.1 Data and catchments

We consider two catchments in the Copenhagen area. The Ballerup catchment has a total area of approximately 1,300 ha and is mainly laid out as a separate system, although it does have a small combined section. The runoff in this area is also strongly influenced by rainfall-dependent infiltration.

The Damhusåen catchment is located close to the Ballerup catchment but drains to a different treatment plant. We consider the northern part of this catchment, which has a total area of approximately 3,000 ha. The catchment is laid out as a combined sewer system, and a multitude of CSOs are located in the area. Flow measurements are available for both catchments at a 5-min resolution.

Numerous online rain gauge measurements are available from the Danish wastewater committee's (SVK) network in the considered catchments (Jørgensen et al. 1998). The gauges marked in Fig. 2 were used as the input for the runoff forecasting models for the two different catchments. These are the same gauges used in previous studies on the Ballerup catchment (Breinholt et al. 2011; Thordarson et al. 2012) and for radar rainfall calibration and RTC in the Copenhagen area (Grum et al. 2011). These gauge measurements are also available with a temporal resolution of 5 min.

We have selected a 3-month measurement period from 25 June 2010 to 29 September 2010 for this study. The period contains several summer storms that can be considered relevant for control applications in urban drainage systems. A modeling time step of 10 min was adopted and corresponds to the temporal resolution used in previous studies (Löwe et al. 2012a, b). The flow and rain gauge data were averaged to match this time step.

2.2 Stochastic greybox models for runoff prediction

We predicted the runoff at the catchment outlets using stochastic greybox models, which are briefly described in this section. The physical part of the models is based on lumped reservoir approaches that transform the rainfall input into the flow output. The principal model setup is described by Breinholt et al. (2011). In this work, we applied a simple two-reservoir cascade to both catchments. In a state space formulation, we used two coupled Itô stochastic differential equations, which together form the following so-called system or state equations

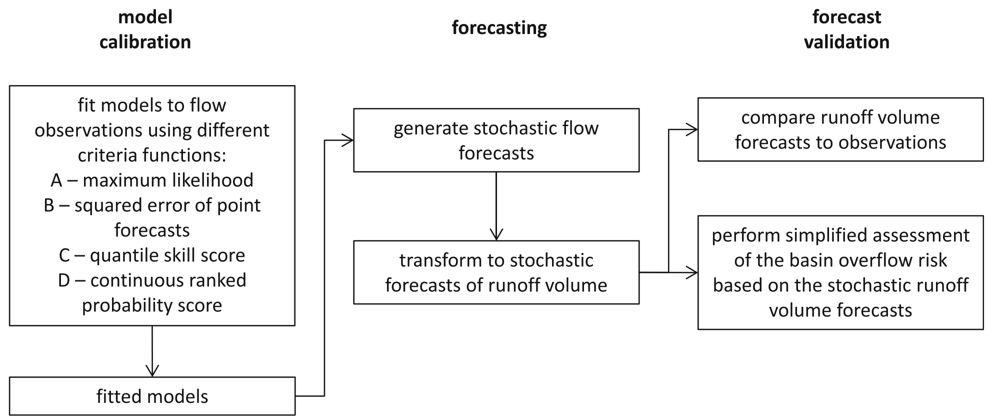

Fig. 1 Flow scheme for comparing model estimation approaches and evaluation of multistep forecasts of runoff volume

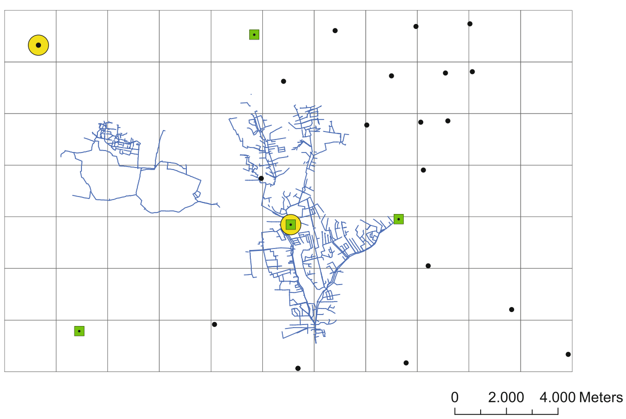

Fig. 2 Ballerup (*left*) and Damhusåen (*right*) catchments with online rain-gauge measurements in the area (*black dots*) and the gauges used as the input data for the Ballerup (*circle*) and the Damhusåen (*rectangle*) catchments

$$d\begin{bmatrix} S_{1,t} \\ S_{2,t} \end{bmatrix} = \underbrace{\begin{bmatrix} A \cdot P_t + a_0 - \frac{1}{K} S_{1,t} \\ \frac{1}{K} S_{1,t} - \frac{1}{K} S_{2,t} \end{bmatrix} dt}_{\textit{Drift term}} + \underbrace{\begin{bmatrix} \sigma_1 S_{1,t}^{\gamma_1} & 0 \\ 0 & \sigma_2 S_{2,t}^{\gamma_2} \end{bmatrix} d\omega_t}_{\textit{Diffusion term}} \tag{1}$$

and the observation equation

$$Y_k = \log(Q_k) = \log\left(\frac{1}{K} S_{2,k} + D_k\right) + e_k \tag{2}$$

where

$$D_k = \sum_i^2 \left(s_i \sin \frac{i2\pi k}{24h} + c_i \cos \frac{i2\pi k}{24h} \right) \tag{3}$$

S_1 and S_2 correspond to the states of the system, i.e. virtual storage fillings, A is the sealed area in the catchment, a_0 refers to the mean dry weather flow at the catchment outlet, and K corresponds to the travel time constant. The rainfall input P_t was determined as the mean area rainfall by averaging the rain-gauge measurements considered for every catchment. In the diffusion term, the variance of the state values was scaled depending on the state value itself because the model predictions are more uncertain in wet weather. The scaling was exponential to avoid extreme increases in the variance in situations with high runoff.

In the observation equation, Q_k corresponds to the flow observation at time step k in discrete time, and D_k describes the variation of the dry weather flows as a harmonic function with parameters s_1, s_2, c_1, and c_2. A log transform was used to avoid negative flow predictions. Please refer to Breinholt et al. (2011) for a detailed derivation and description of the model structure.

The open source software CTSM (Kristensen and Madsen 2003) was used for the parameter estimation and the forecast generation. State-dependent diffusion terms, such as those in Eq. (1), cannot be modeled in this setup (Vestergaard 1998). Therefore, a Lamperti transform was applied to the system equations (1), as described by Breinholt et al. (2011).

The multistep flow predictions were generated using the extended Kalman filter with updating. This setup provides a log-transformed flow prediction $\hat{Y}_{i+k|i}$ with variance $R_{i+k|i}$ that is assumed to be normally distributed.

2.3 Parameter estimation for stochastic runoff prediction models

The purpose of the runoff prediction models considered is to describe the expected runoff volume over a horizon of variable extent, which is defined by the control scheme. When estimating the model parameters from historical data, we need to design the objective function such that the resulting model is actually optimal for the generation of predictions for different horizons. Below, we introduce possible objective functions.

All of the suggested objective functions focus on flow values rather than runoff or even overflow volumes because the conversion from stochastic flow to runoff predictions is computationally demanding. In addition, the models should be estimated to correctly describe the physical behavior of the system and thus reduce the risk of overfitting (Weijs et al. 2010). The physical system behavior is captured when focusing on flow values during the parameter estimation, whereas focusing on overflow volumes would likely introduce a partial loss of the information provided by the measurements.

Parameter estimation was performed automatically in all cases using a genetic algorithm based on the concepts described by Whitley (1994), Spall (2003) and Hallam (2010).

2.3.1 Maximum likelihood estimation (model A)

The most common approach for the estimation of parameters in stochastic greybox models is to maximize the likelihood function for a given series of measurements (Kristensen et al. 2004; Breinholt et al. 2011). The computation of the likelihood function is based on the computation of the one-step prediction errors or innovations under the assumption that the one-step-ahead conditional densities are Gaussian:

$$\varepsilon_i = Y_i - \hat{Y}_{i|i-1} \tag{4}$$

This approach may be difficult in the context of the estimation of models for multi-step predictions in the urban runoff setting. The parameters found may not to be optimal for multi-step predictions because these are based on the one-step prediction errors. Furthermore, there is a clear risk of overfitting if the physical part of the model fails to completely capture the system behavior. The one-step predictions are strongly influenced by the updating of the states in the extended Kalman filter, and we may identify parameters that are optimal for this updating but do not actually match the physical system behavior, which would result in bad forecasts and simulations.

One may argue that, in these situations, the modeler should attempt to improve the physical part of the model and make it more suitable to the actual behavior of the catchment. However, in practical applications, we will often face the situation that a simple model will be sufficient for the forecasting purpose; moreover, the tailoring of a model to each new catchment may be too time-consuming. This also indicates a need for more robust estimation methods that focus on the forecasting purpose.

2.3.2 Minimizing the error of the predicted runoff volumes (model B)

The fitting of forecast models in hydrology is typically performed by minimizing the forecast error variance (see, e.g., Nash and Sutcliffe 1970). We suggest an objective function based on the sum of the squared errors between the predicted and the observed runoff volumes over the prediction horizon:

$$S(\theta) = \sum_{i=1}^{N}\left(\sum_{j=1}^{k} Q_{i+j} - \sum_{j=1}^{k}\hat{Q}_{i+j|i}(\theta)\right)^2 \Delta t \tag{5}$$

At every time step i of length Δt, a k-step ahead flow prediction $\hat{Q}$ is generated. The flow values are integrated to a runoff volume over the prediction horizon and compared to the observations Q. The minimization of the sum of these volume differences for all N time steps gives an objective function for the estimation of the model parameter set θ.

This objective function optimizes the model to give a good point forecast of the expected runoff volumes over the maximum prediction horizon of k steps (e.g. $k = 10$ steps). Implicitly, we assume that we also obtain good predictions for shorter horizons.

2.3.3 Estimation based on the interval score (model C)

Minimizing the squared error of the predicted runoff volumes tunes the forecast models to give good point predictions of the runoff volume for multistep prediction horizons. The quality of the forecast uncertainties is not evaluated in this criterion. However, the forecast objective in the described setup is to obtain a probabilistic description of the predicted runoffs. The predictive distribution should be as narrow (sharp) as possible and at the same time match the observations (be calibrated or reliable).

To account for the quality of the probabilistic predictions, we can modify the criterion developed in Sect. 2.3.2. Assuming normality, we compute a $(1-\beta)\cdot 100\,\% = 95\,\%$ prediction interval for forecast horizon j for the log-transformed flow values as

$$\begin{aligned} \widehat{uY}_{i+j|i} &= \hat{Y}_{i+j|i} + 1.96\cdot\sigma_{\hat{Y}_{i+j|i}} \\ \widehat{lY}_{i+j|i} &= \hat{Y}_{i+j|i} - 1.96\cdot\sigma_{\hat{Y}_{i+j|i}} \end{aligned} \tag{6}$$

where $\sigma_{\hat{Y}_{i+j|i}}$ is the standard deviation of the j-step predictions.

The quality of this prediction interval can be evaluated using a number of methods, e.g. the interval score described by Gneiting and Raftery (2007), which was applied to stochastic flow forecasts in urban drainage systems by Thordarson et al. (2012). The score for the j-step prediction generated at time step i is

$$\begin{aligned} SC_{i,j}^{\beta} = {} & \widehat{uY}_{i+j|i} - \widehat{lY}_{i+j|i} + \frac{2}{\beta}(\widehat{lY}_{i+j|i} - Y_{i+j})\mathbb{1}\{Y_{i+j} < \widehat{lY}_{i+j|i}\} \\ & + \frac{2}{\beta}(Y_{i+j} - \widehat{uY}_{i+j|i})\mathbb{1}\{Y_{i+j} > \widehat{uY}_{i+j|i}\} \end{aligned} \tag{7}$$

In Eq. (8), a reasonable scoring rule based on Eq. (7) is suggested and accounts for several forecast horizons. More weight is placed on the flow forecasts for shorter horizons, which have a stronger influence on forecasts of runoff volume because the latter are generated as an integral over the flow forecasts for different horizons.

$$SC_i^{\beta} = \frac{1}{\sum_{j=1}^{k}(k-j+1)} \left(\sum_{j=1}^{k}(k-j+1) \cdot SC_{i,j} \right) \tag{8}$$

By averaging over all N time steps, we obtain the objective function for parameter estimation in model C.

$$S(\theta) = \frac{1}{N}\sum_{i=1}^{N} SC_i^{\beta} \tag{9}$$

2.3.4 Estimation based on continuous ranked probability score (model D)

The interval score criterion described above was previously applied to flow forecasts in urban drainage systems, but focuses on a 95 % prediction interval, i.e., only the tails of the predictive distribution. This may lead to a dislocation of the center of the predicted flow distribution. The continuous ranked probability score (CRPS) is a measure of the fit of the overall distribution; therefore, we introduce this score here as the last objective function for parameter estimation in the stochastic runoff forecasting models. Gneiting et al. (2005) suggested the use of the CRPS in the fitting of post-processing models for ensemble predictions and consider it robust toward extreme events and outliers. A discussion of the score can be found in the manuscript published by Gneiting and Raftery (2007).

We obtained the CRPS for the j-step flow prediction generated at time step i as

$$CRPS_{i,j} = \int_{-\infty}^{\infty} \left(F(s) - \mathbb{1}\{s > Y_{i+j}\}\right)^2 ds \tag{10}$$

where F is the cumulative distribution function for the (assumed normally distributed) log-transformed flow prediction $\hat{Y}_{i+j|i}$, and Y_{i+j} is the corresponding $(i+j)$th value in the time series of the observations. $\mathbb{1}$ denotes the Heaviside function and takes the value 0 when $s < Y_{i+j}$ and 1 otherwise. There exists a closed-form solution for Eq. (10) if the predicted value is normally distributed. However, we do not expect to be able to always rely on this assumption in practical situations and therefore chose to evaluate the integral numerically.

As in Eq. (8), we performed a weighting of the CRPS values derived for different forecast horizons to obtain an average value for every time step. Ultimately, we averaged the values for all of the considered time steps as in Eq. (9) to obtain the value of the objective function. The optimal parameter set is found by minimizing this value.

2.4 Generating stochastic forecasts of runoff volumes

The applied greybox models provide flow forecasts for horizons 1 up to k. To derive probabilistic forecasts of the runoff volume, we used a multivariate sampling approach. The correlations between the flow forecasts for different horizons are derived from past forecast errors. The following steps were taken.

- Generate a 10-step forecast at time step i from the greybox models. We obtained a vector of (assumed normal) log-transformed flow predictions $\hat{Y}_i$ containing the forecast values for horizons 1 through 10. The corresponding observations are denoted Y_i.
- Find the error covariance contribution from this time step (Madsen 2008):

$$V_i = (Y_i - \hat{Y}_i) \cdot (Y_i - \hat{Y}_i)^T \tag{11}$$

- Estimate the overall error covariance structure for time step i using exponential smoothing. This allows for time variation of the considered correlations. We applied $\lambda = 0.99$.

$$\Sigma_i = \lambda \cdot \Sigma_{i-1} + (1-\lambda) \cdot V_i \tag{12}$$

- Scale Σ_i to the predictive variances provided by the greybox model. We obtained a covariance structure with variances according to those predicted by the model and correlation values estimated from the forecast errors.
- Create 100,000 multivariate flow samples from the $N(\hat{Y}_i, \Sigma_{S,i})$ distribution [using the R-package MASS (Venables and Ripley 2002)], each of which represents a possible flow scenario for horizons 1 through 10. Integrate each sample into runoff volumes and empirically derive the distribution of the runoff volumes.

2.5 Forecast evaluation

A set of measures was applied to evaluate the quality of the prediction intervals generated by the stochastic greybox models. These are described by Thordarson et al. (2012) and Jin et al. (2010). All of the measures were applied not to flow predictions as in Thordarson et al. (2012) but to runoff volume predictions for different forecast horizons.

- Reliability

$$REL = \frac{1}{N}\sum_{i=1}^{N} n_i^{\beta} \tag{13}$$

where N is the number of observations, β is the significance level, and n^{β} is an indicator variable with value 1 if an observation is not contained in the $(1-\beta)\%$ prediction interval and 0 otherwise. The measure corresponds to the percentage of observations not contained in the $(1-\beta)\%$ prediction interval. A reliability bias can be defined as

$$RB = \beta - REL \tag{14}$$

and becomes negative if the prediction bands fail to include more than $\beta\%$ of the observations (it is otherwise positive). Ideally, the reliability bias should be 0. The bias is bounded depending on the significance level.

– Average relative interval length

$$ARIL = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{U}_{i+k|i} - \hat{L}_{i+k|i}}{V_{i+k|i}} \tag{15}$$

This refers to the average width of a volume prediction interval with lower bound $\hat{L}_{i+k|i}$ and upper bound $\hat{U}_{i+k|i}$ generated for a forecast horizon of k time steps relative to the observed value $V_{i+k|i}$. We consider 95 % prediction intervals.

– *CRPS* (10)

In general, a good stochastic forecast will be calibrated, i.e., generate reliabilities close to the significance level of the required prediction interval. Given a calibrated model, the spread of the prediction bounds should of course be as narrow as possible, which is indicated by low *ARIL* values. As an overall criterion, we aim to obtain the minimal *CRPS* for the forecasts of runoff volume.

2.6 Evaluating the overflow risk for different forecast types

To evaluate the effect of the considered forecasting models on the RTC, a simplified assessment of the model ability to correctly predict the overflow cost according to Eq. (16) was used. We assumed a basin at the outlet of both catchments studied. The basin outlet capacity is fixed. The outlet capacity and volume were both chosen somewhat arbitrarily but such that a reasonable amount of overflow is obtained in the summer period considered. The selected values are shown in Fig. 3.

We considered a prediction horizon of 10 time steps or 100 min. Evaluating the basin mass balance with the selected characteristics and the measured time series of catchment outflows, we determined a series of true 'predicted' overflow volumes over a 100-min horizon at every time step. Assuming a unit cost of overflow volume, this amount also corresponds to the true 'predicted' overflow cost $C_{f,t}$:

$$C_{f,t} = \int C(V_f)\cdot p(V_f)_t dV_f \tag{16}$$

where $C(V_{f,t})$ corresponds to the cost value forecasted at time step t and $p(V_f)_t$ is the forecasted probability that a runoff volume V_f occurs.

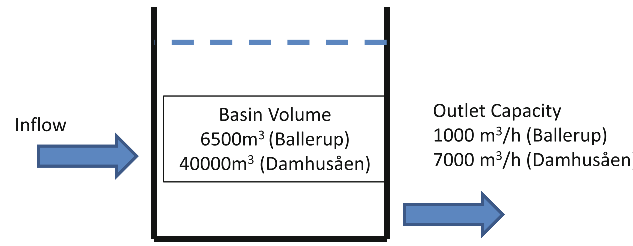

Fig. 3 Simplified model setup used for the evaluation of the predicted overflow cost for the different models and catchments

Forecasts of the runoff volumes were again derived from the probabilistic flow forecasts using the sampling approach described in Sect. 2.5. Each sample forms a time series of flow predictions for the different horizons. We can evaluate the basin mass balance for this time series and compute the predicted overflow cost for each sample. Ideally, the predicted overflow cost derived from the stochastic models will match the reference derived from the observations at every time step.

3 Results

3.1 Forecast performance of different objective functions

3.1.1 Runoff predictions for a number of rain events

Figures 4 and 5 compare the predicted runoff volumes from the different models to the observed runoff volumes in the Damhusåen catchment. We also included prediction intervals that are based on the point prediction of model A, which describe the uncertainty of the runoff forecasts by a Gamma distribution, as detailed by Vezzaro and Grum (2012).

We found that model A satisfactorily captures the characteristics of the observed runoff curve. The prediction intervals, however, appear to be rather small. Model B provides very wide prediction intervals, whereas model C gives wider prediction intervals than model A. The forecasts from model D appear similar to those from model A, although the 50 % quantile of the forecasts appears to match the observations slightly better than model A. The prediction bounds from model D are narrower than those from model A.

With the exception of model B, all of the models appear to provide better estimates of forecast uncertainty than the Gamma distribution.

The estimated model parameters, which are shown in Table 1, exhibit the following tendency: models estimated using multistep predictions produce more pronounced runoff peaks as a result of the smaller K values in the Ballerup

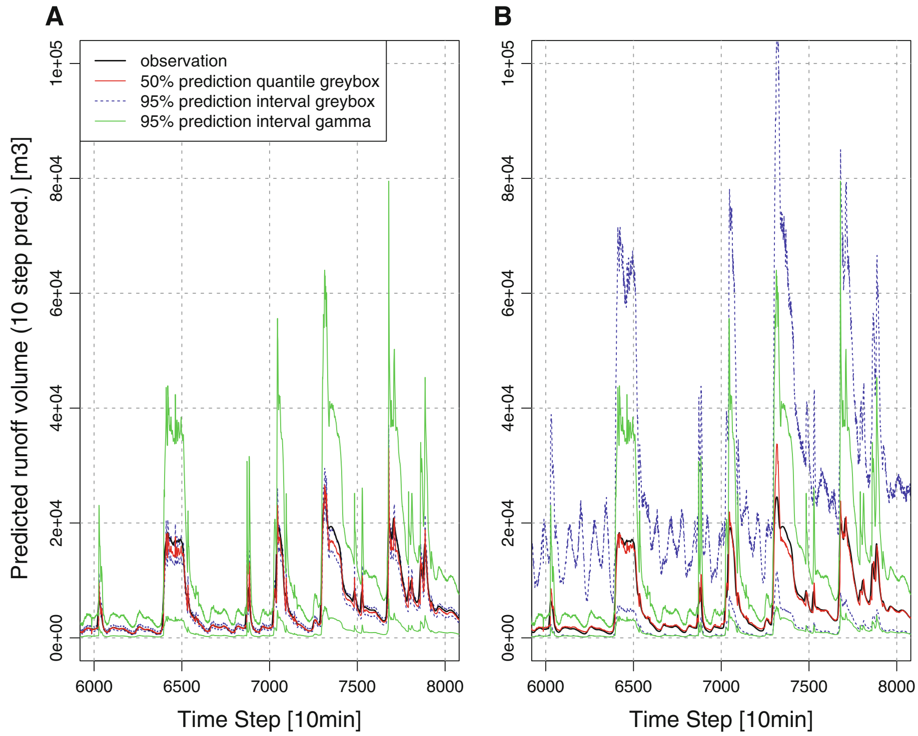

Fig. 4 95 % prediction intervals for the 10-step runoff volume forecasts from model A (*left*) and model B (*right*). The 50 % prediction quantile and the observation and the prediction intervals derived from the Gamma distribution are also shown (Damhusåen catchment)

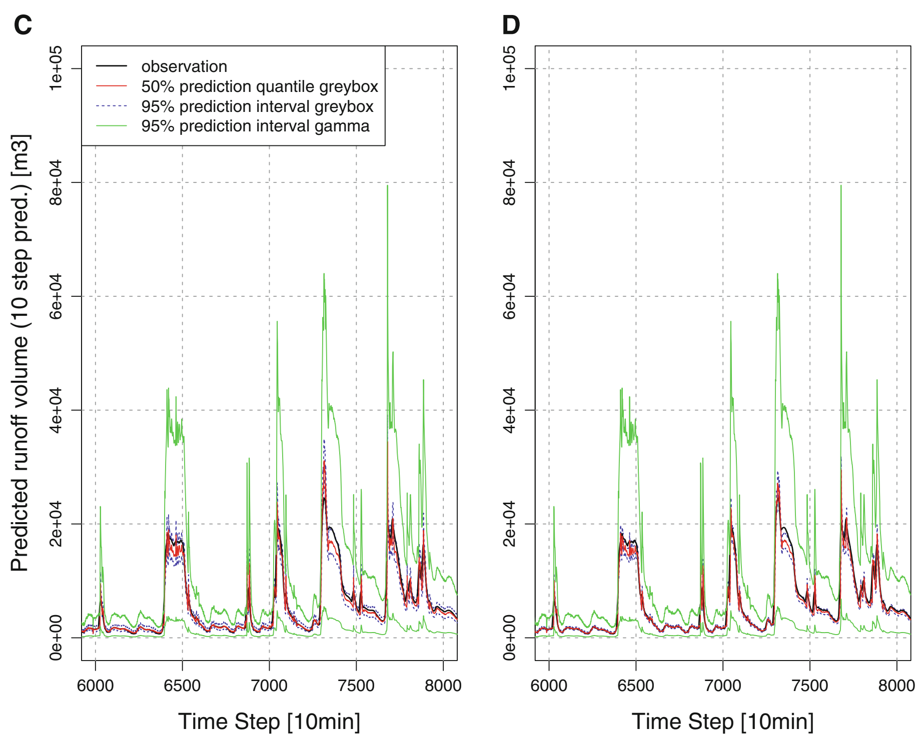

Fig. 5 95 % prediction intervals for the 10-step runoff volume forecasts from model C (*left*) and model D (*right*). The 50 % prediction quantile and the observation and the prediction intervals derived from the Gamma distribution are also shown (Damhusåen catchment)

catchment and the larger effective areas *A* in the Damhusåen catchment. Note that the models do not necessarily respect the mass balance due to the state updating. For all models, we obtained rather small observation uncertainties σ_e compared with the uncertainty of the model states (σ_1, σ_2). This result demonstrates, that we consider the information content in the flow observations to be high and update the model to stay close to these observations. The different forecast uncertainties apparent in Figs. 4 and 5 are a result of the different state uncertainties σ_1 and σ_2, which are shown in Table 1.

Table 1 Parameter estimates for the two catchments obtained with different estimation approaches

	a_0 (m^3/h)	K (h)	A (ha)	σ_1	σ_2	σ_e
Ballerup						
A	393	7.52	206	1.61E+00	1.28E−02	6.88E−06
B	400	5.28	55	1.26E+00	5.54E−02	6.76E−07
C	372	3.03	74	4.35E−01	1.82E−02	4.35E−09
D	307	3.63	78	3.62E−01	1.09E−02	6.03E−06
Damhusåen						
A	841	1.95	94	1.16E+00	6.79E−03	1.24E−08
B	1,678	2.49	270	7.70E+00	1.29E−01	1.07E−10
C	997	1.88	207	1.34E+00	7.34E−03	6.72E−10
D	933	2.51	122	7.64E−01	5.92E−03	7.24E−10

3.1.2 Evaluation of the predictive distributions

The first step in this analysis is to study the overall quality of the predictive distributions. The *CRPS* was used to compare the forecasts and the observations. Note that, other than in the criterion derived for the model estimation in Sect. 2.3.4, we based the analysis on the predicted runoff volumes for a given horizon.

The estimation based on the volume prediction errors (model B) clearly gives the worst *CRPS* values. For the other models, we cannot easily identify the differences based on this criterion. In both catchments, the volume forecasts generated by models A, C, and D are very similar with respect to the *CRPS*.

Table 3 shows the *ARIL* values for the 95 % prediction intervals of the runoff volumes for different forecast horizons. Model B yields very wide prediction intervals because it considers only the point prediction in the model estimation. Large state uncertainties facilitate the state updating and, if the quality of the observations is acceptable, lead to better point predictions. The predicted uncertainties, however, are too large.

There is less difference between the forecasts generated by models A and C with respect to *ARIL*. The forecasts generated by model D are clearly sharper than those obtained with the other models. This tendency of the *CRPS*-based estimation was also noted by Gneiting et al. (2005).

Although we assumed that the simple lumped reservoir model is much less suited to the prediction of the runoff in the bigger and more complex Damhusåen catchment than in the Ballerup catchment, we cannot identify a trend toward relatively larger forecast uncertainties for this catchment.

We continued this analysis by evaluating the distribution of the predicted runoff volumes. Figure 6 shows the reliability biases *RB* of the runoff volume predictions considering different levels of significance β and prediction horizons. A significance level of $\beta = 0.05$ corresponds to a $1 - 0.05 = 95$ % prediction interval.

Table 2 CRPS for volume predictions in m^3 considering different prediction horizons (in time steps, step length = 10 min) and different estimation approaches for the two catchments

Horizon	Ballerup				Damhusåen			
	A	B	C	D	A	B	C	D
1	2	3	2	2	5	25	5	5
2	5	7	5	5	14	62	14	15
3	8	12	8	8	28	110	29	30
4	12	17	12	11	48	169	48	51
5	16	23	16	15	73	239	73	76
6	21	30	21	20	103	319	103	106
7	27	38	26	25	137	410	137	141
8	33	46	32	30	176	511	176	181
9	39	54	38	36	219	623	218	224
10	46	63	45	42	267	745	264	271
Mean	21	29	20	19	107	321	107	110

Table 3 ARIL for 95 % volume prediction intervals considering different prediction horizons (in time steps, step length = 10 min) and different estimation approaches for the two catchments

Horizon	Ballerup				Damhusåen			
	A	B	C	D	A	B	C	D
1	0.21	0.67	0.22	0.19	0.10	1.65	0.12	0.08
2	0.22	0.75	0.24	0.19	0.14	2.15	0.16	0.10
3	0.24	0.85	0.27	0.20	0.18	2.73	0.20	0.12
4	0.27	0.95	0.30	0.22	0.22	3.38	0.25	0.14
5	0.30	1.04	0.33	0.23	0.26	4.13	0.29	0.16
6	0.33	1.14	0.35	0.25	0.29	4.97	0.34	0.19
7	0.36	1.23	0.38	0.26	0.33	5.93	0.39	0.21
8	0.39	1.31	0.40	0.27	0.37	7.00	0.43	0.23
9	0.42	1.40	0.43	0.29	0.40	8.19	0.48	0.25
10	0.45	1.48	0.45	0.30	0.44	9.50	0.52	0.27
Mean	0.32	1.08	0.34	0.24	0.27	4.96	0.32	0.18

For model A, we observed small reliability biases for high coverage rates, i.e., at the tails of the distribution. For smaller coverage rates, however, we overestimated the forecast uncertainties, which led to positive reliability biases. This problem becomes more pronounced for longer forecast horizons.

As indicated previously, model B clearly overestimates the forecast uncertainties and yields strongly positive reliability bias values. Model C gives results that are similar to those of model A, but generates smaller reliability biases at longer horizons.

Model D yields a slight underestimation of the forecast uncertainties for high coverage rates. Compared with

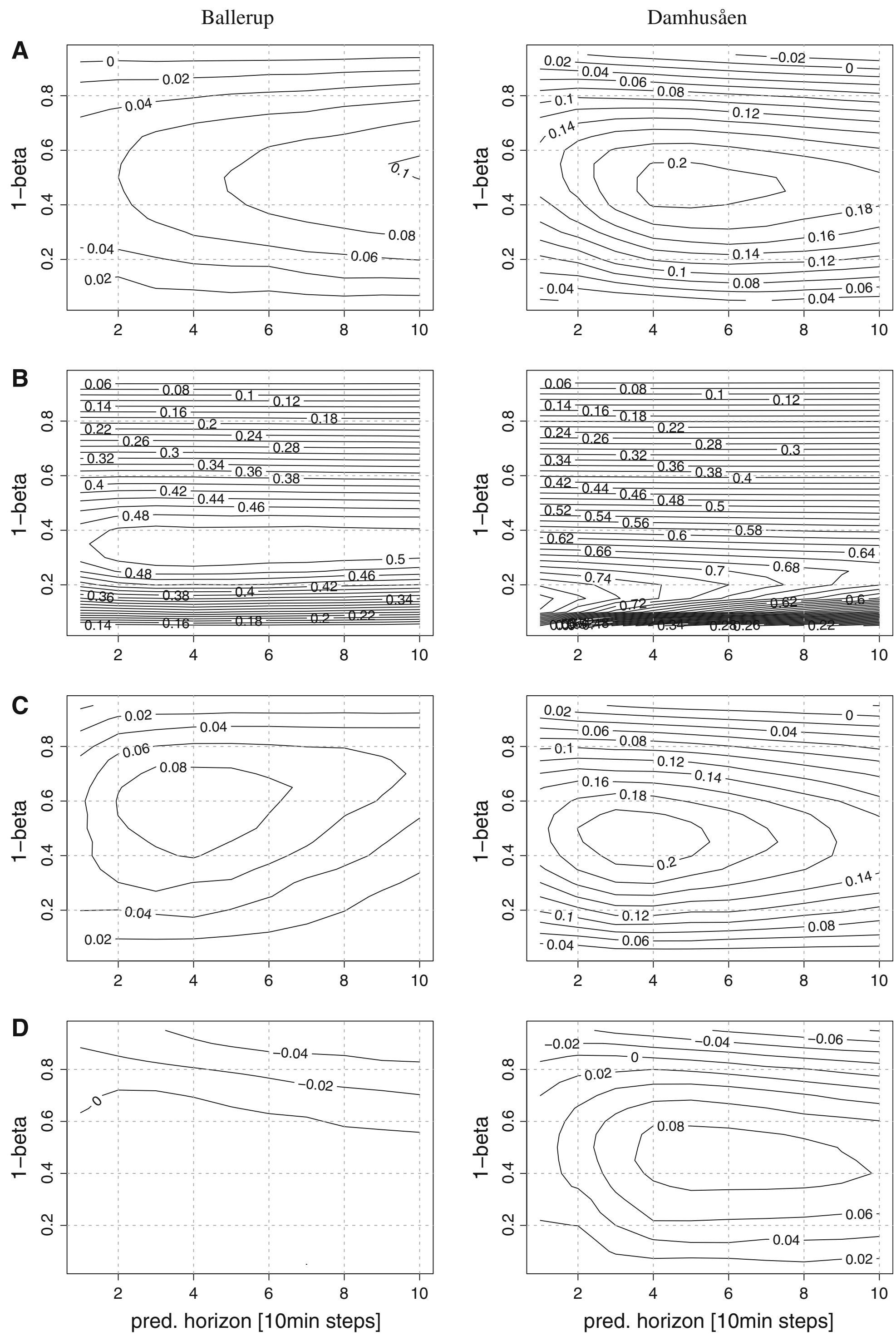

Fig. 6 Reliability bias for models A (*top*) through D (*bottom*) considering the prediction intervals for different levels of significance β and different forecast horizons for the Ballerup (*left*) and Damhusåen (*right*) catchments

models A and C, however, the overestimation of the uncertainties in the center of the distribution is also reduced. Similar to model C, we observed smaller reliability bias values at longer horizons with model D compared with model A.

Models C and D account for multistep predictions in model estimation. In both cases, this results in reduced reliability bias values at longer horizons compared to model A. The parameter estimation in model C focuses on 95 % prediction intervals. This model consequently provides the best fit at the tails of the distribution.

For model D, a more balanced behavior can be observed with a reduced overestimation of the uncertainties at the center of the distribution but an underestimation at the tails.

The latter leads to slightly worse *CRPS* values of the forecasts of runoff volume compared with model A.

In general, all of the models result in either an overestimation of the forecast uncertainties at the center of the distribution or an underestimation at the tails. This finding indicates that the normality assumption used in the multivariate sampling approach may not hold.

There is a noticeable difference between the two catchments. Although we obtained somewhat reliable (or calibrated) forecasts for the Ballerup catchment, we tended to overestimate the uncertainties in the Damhusåen catchment because the applied model is less suitable to the description of the behavior of this system.

In both catchments, the forecast uncertainties during rain events are clearly underestimated by models A, C, and D (data not shown). This finding indicates that the applied model structure is not able to properly distinguish between dry and wet weather uncertainties.

3.2 Predicted overflow cost for fictive basins

Table 4 shows the overflow cost predicted using the simplified approach described in Sect. 2.6. The values for the true observations and the runoff predictions generated by the different models over a horizon of 10 time steps are shown. The values shown are integrated over the whole time period of 3 months. To compare the results with the state-of-the-art method, we included two additional cases:

- Gamma—uses the point prediction from model A and derives the forecast uncertainty for the runoff volumes from a Gamma distribution, the shape parameters of which depend on the point value (Vezzaro and Grum 2012).
- Model A Point—derives the predicted overflow volumes using the point forecast of model A without considering the forecast uncertainties.

We found that model A produces values close to the true overflow volumes in the Ballerup catchment and underestimates the true cost in the Damhusåen catchment. When ignoring the forecast uncertainties provided by model A, we obtained almost the same results (Model A Point), whereas the description of the forecast uncertainties with a Gamma distribution results in a clear overestimation of the overflow volumes. This finding indicates that a correct point forecast is essential for a good estimation of the overflow volumes in the simplified setup, whereas too small or no estimates of forecast uncertainties hardly affect the estimation of overflow volumes. In contrast, a too large estimate of the runoff forecast uncertainties, as obtained from the Gamma distribution, will lead to an overestimation of the overflow risk.

Table 4 Predicted overflow cost for 10-step horizon in m^3 accumulated over all time steps

	Ballerup	Damhusåen
True	3.1E+05	6.5E+05
A	3.2E+05	4.2E+05
B	3.8E+05	2.2E+06
C	2.5E+05	4.4E+05
D	2.6E+05	4.6E+05
A Γ	4.5E+05	8.2E+05
A Point	3.1E+05	4.1E+05

Table 5 RMSE between the true overflow cost and the prediction in m^3 from different models for a 10-step horizon

	A	B	C	D	A Γ	A Point
Ballerup	37	43	45	40	62	37
Damhusåen	190	672	177	167	194	195

A similar result was obtained with model B, which generated reasonable results for the Ballerup catchment but strongly overestimated the overflow risk in the Damhusåen catchment as a result of the very high estimates of the forecast uncertainty.

Models C and D exhibit a tendency to underestimate the overflow volumes. As in model A, this underestimation is the result of the underestimation of the runoffs by the point prediction, as observed in Fig. 5.

Considering the *RMSE* between the true overflow cost for a 10-step horizon and the predicted overflow cost values derived from the different models, we obtained a similar picture. A clear overestimation of the forecast uncertainties also results in increased *RMSE* values for the overflow risk (model A Γ and model B), whereas neglecting the forecast uncertainties hardly affects the estimated overflow cost values (model A Point). Models C and D provide better point forecasts during the overflow events in the Damhusåen catchment, which results in smaller *RMSE* values for the overflow cost.

In the authors' view, the most interesting outcome of this analysis is that no difference was found between the deterministic prediction of the overflow risk (model A Point) and the use of forecast uncertainties (model A). Two possible reasons can be suggested for this result. First, a linear relationship between the overflow volume and the overflow cost was used in this simplified analysis. With a nonlinear relationship that punishes (for example) the start of an overflow event, forecasts of the overflow risk will profit more from a proper quantification of the uncertainty of the runoff predictions.

Second, the analysis performed here was static in the basin layout. Using the considered dataset, we obtained either only one event with a small overflow volume or several events with rather large overflow volumes by choosing different basin dimensions or outlet capacities. The effect of considering the forecast uncertainties is most visible in those events where either the basin is close to being completely filled or where only small overflow volumes are observed. In a predictive RTC system, the simulated basin outlet is varied in the optimization routine, which results in strong variations in the simulated basin filling. A proper description of the forecast uncertainties is more important in those cases.

4 Conclusions

We have evaluated the quality of probabilistic multistep runoff volume forecasts generated by stochastic greybox models and compared the effect of different estimation methods on the forecast quality. Four methods were compared: a maximum likelihood estimation based on one-step-ahead predictions (model A), a deterministic method that minimizes the error of the 10-step-ahead predictions (model B), and two methods that minimize the interval score for the 95 % intervals of the multistep flow predictions (model C) or the continuous ranked probability score (*CRPS*, model D).

We concluded that, although it focuses on the whole prediction horizon, the deterministic estimation method (model B) is unsuitable for estimating the stochastic models. The quality of the predictive uncertainty is not a criterion in the objective function for this method. In the cases considered here, this model results in too large estimates of the uncertainty for the states.

Models A, C, and D provided reasonable estimation results and multistep forecasts of the runoff volume with similar skill values. Overfitting by model A was not observed as a result of the high quality of the considered flow observations. More noisy measurements will make the parameter estimation using one-step predictions more difficult and favor approaches focusing on multistep predictions.

However, the use of multistep predictions in parameter estimation (models C and D) clearly reduces the overestimation of the uncertainties at longer forecast horizons. Using the interval score for the parameter estimation (model C) results in forecasts that are suitable for the 95 % prediction interval and overestimate the uncertainties in the center of the distribution. Applying the CRPS as the objective function (model D) allows the balance of this effect and gives forecasts that are more evenly calibrated over the whole distribution.

In the prediction of the overflow risk in a simplified setup, it was demonstrated that a significant overestimation of the runoff forecast uncertainties leads to a strong overestimation of the overflow risk. Consequently, models A, C, and D all outperform the reference model, which describes the forecast uncertainties with a simple gamma distribution.

In the applied setup, it is not possible to show that the risk of basin overflow can be predicted better through the dynamic modeling of the uncertainties of the runoff forecasts compared to the application of a simple point forecast. However, the analysis applied here is linear and static in the basin layout. It is expected that the forecast uncertainties will be relevant in a more realistic control setting that exhibits nonlinear relationships between the forecast values and the risk and where the basin outflows are dynamically modified as part of an optimization routine.

In addition, all of the models clearly underestimate the forecast uncertainties during rain events. This finding suggests that the model structure should be modified to allow a proper separation of the dry and the wet weather uncertainties.

We need to be aware that this study focuses strongly on the correct prediction of the overflow risk to improve the RTC of sewer systems. The methods suggested for the prediction of these risks, however, are also applicable in other contexts of the urban drainage system, such as the prediction of the critical operational states at a wastewater treatment plant, the risk of flooding induced by overloading of the sewer system, and the risk of microbial pollution as a result of sewer overflows close to bathing areas.

Acknowledgments This research was financially supported by the Danish Council for Strategic Research, Program of the Commission on Sustainable Energy and Environment through the Storm- and Wastewater Informatics (SWI) Project. The catchment and flow data were kindly provided by Avedøre Wastewater Services and Copenhagen Energy. We thank Luca Vezzaro for ideas on the evaluation of the overflow risk, Julija Tastu and Pierre Pinson for their input on the generation of predictive distributions and the evaluation of probabilistic forecasts and Anne Katrine Duun-Henriksen for help with the preparation of this manuscript.

References

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

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

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

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

Fuchs L, Beeneken T (2005) Development and implementation of a real-time control strategy for the sewer system of the city of Vienna. Water Sci Technol 52(5):187–194

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

Gneiting T, Raftery AE, Westveld AH III, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133(5):1098–1118

Grum M, Thornberg D, Christensen M, Shididi S, Thirsing C (2011) Full-scale real time control demonstration project in Copenhagen's largest urban drainage catchments. In: Proceedings of the 12th international conference on urban drainage, Porto Alegre

Hallam JW (2010) Genetic algorithms with shrinking population size. Comput Stat 25(4):691

Jin X, Xu CY, Zhang Q, Singh V (2010) Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. J Hydrol NZ 383(3–4): 147–155

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

Kristensen NR, Madsen H (2003) Continuous time stochastic modelling CTSM 2.3—mathematics guide. Technical report, Technical University of Denmark, http://www2.imm.dtu.dk/~ctsm/MathGuide.pdf

Kristensen NR, Madsen H, Jørgensen SB (2004) Parameter estimation in stochastic grey-box models. Automatica 40(2):225–237

Löwe R, Mikkelsen PS, Madsen H (2012a) Forecast generation for real-time control of urban drainage systems using greybox modelling and radar rainfall. In: Proceedings of the 10th international conference on hydroinformatics, Hamburg

Löwe R, Mikkelsen PS, Rasmussen MR, Madsen H (2012b) State-space calibration of radar rainfall and stochastic flow forecasting for use in real-time control of urban drainage systems. In: Proceedings of the 9th international conference on urban drainage modelling, Belgrade

Madsen H (2008) Time series analysis. Chapman Hall/CRC, Boca Raton

Nash JE, Sutcliffe J (1970) River flow forecasting through conceptual models part I—A discussion of principles. J Hydrol 10(3): 282–290

Pabst M, Alex J, Beier M, Niclas C, Ogurek M, Peikert D, Schütze M (2011) ADESBA—a new general global control system applied to the Hildesheim sewage system. In: Proceedings of the 12th international conference on urban drainage, Porto Alegre

Puig V, Cembrano GG, 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 Sci Technol 60(4):869–878

Schütze M, Campisano A, Colas H, Schilling W, Vanrolleghem PA (2004) Real time control of urban wastewater systems—where do we stand today? J hydrol 299(3):335–348

Seggelke K, Löwe R, Beeneken T, Fuchs L (2012) Implementation of an integrated real-time control system of sewer system and WWTP in the city of Wilhelmshaven. Urban Water J. doi:10.1080/1573062X.2013.820331

Spall JC (2003) Introduction to stochastic search and optimization: estimation, simulation, and control. Wiley-Interscience series in discrete mathematics. Wiley, Hoboken

Thordarson FO, Breinholt A, Møller JK, Mikkelsen PS, Grum M, Madsen H (2012) Uncertainty assessment of flow predictions in sewer systems using grey box models and skill score criterion. Stoch Environ Res Risk Assess 26(8):1151–1162

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. http://www.stats.ox.ac.uk/pub/MASS4

Vestergaard M (1998) Nonlinear filtering in stochastic volatility models. Master Thesis, IMM Eksamensprojekt 36, Technical University of Denmark, Kgs. Lyngby

Vezzaro L, Grum M (2012) A generalized dynamic overflow risk assessment (DORA) for urban drainage RTC. In: 9th international conference on urban drainage modelling, Belgrade

Weijs SV, Schoups G, van de Giesen N (2010) Why hydrological forecasts should be evaluated using information theory. Hydrol Earth Syst Sci Discuss 7:4657–4685

Whitley D (1994) A genetic algorithm tutorial. Stat Comput 4(2): 65–85