Environmental Modelling & Software 77 (2016) 32–49

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Influent generator for probabilistic modeling of nutrient removal wastewater treatment plants

Mansour Talebizadeh [a,*], Evangelia Belia [b], Peter A. Vanrolleghem [a]

[a] modelEAU, Département de génie civil et de génie des eaux, Université Laval, 1065 av. de La Médecine, Québec, QC G1V 0A6, Canada
[b] Primodal Inc., 145 Aberdeen, Québec, QC G1R 2C9, Canada

A R T I C L E   I N F O

Article history:
Received 22 November 2014
Received in revised form
27 August 2015
Accepted 16 November 2015
Available online xxx

Keywords:
Bayesian estimation
Probabilistic design
Uncertainty analysis
Urban hydrology
Wastewater composition
Weather generator

A B S T R A C T

The availability of influent wastewater time series is crucial when using models to assess the performance of a wastewater treatment plant (WWTP) under dynamic flow and loading conditions. Given the difficulty of collecting sufficient data, synthetic generation could be the only option. In this paper a hybrid of statistical (a Markov chain-gamma model for stochastic generation of rainfall and two different multivariate autoregressive models for stochastic generation of air temperature and influent time series in dry conditions) and conceptual modeling techniques is proposed for synthetic generation of influent time series. The time series of rainfall and influent in dry weather conditions are generated using two types of statistical models. These two time series serve as inputs to a conceptual sewer model for generation of influent time series. The application of the proposed influent generator to the Eindhoven WWTP shows that it is a powerful tool for realistic generation of influent time series and is well-suited for probabilistic design of WWTPs as it considers both the effect of input variability and total model uncertainty.

Software availability

Name of the software: WWTP influent advisor
Developer: Mansour Talebizadeh, Evangelia Belia, Peter A. Vanrolleghem
Programing language: Matlab 2012
Availability: The software can be obtained upon request by contacting Evangelia Belia, Primodal Inc., 145 Aberdeen, Québec, QC G1R 2C9, Canada. Email:belia@primodal.com

1. Introduction

One of the major sources of uncertainty/variability that both plant designers and operators must deal with is the dynamics of the influent (Belia et al., 2009). The recent advances in mathematical modeling and improved computational power have enabled researchers to better understand the performance of different WWTP design alternatives (Hao et al., 2001; Salem et al., 2002; Hyland et al., 2012) and/or evaluate control strategies under dynamic flow and loading conditions. However, the application of mathematical models used for simulating the performance of a WWTP could be misleading unless, among other reasons, models are provided with representative influent time series. One of the problems that arise in this regard is the scarcity or even lack of long-term influent data. To remedy this problem, some researchers have proposed models for synthetic dynamic influent time series scenarios (Bechmann et al., 1999; Gernaey et al., 2011). The development of a tool capable of generating dynamic influent time series that is representative of the climate and characteristics of the sewershed could have several applications. Synthetically-generated influent time series can serve as input to a dynamic model of a plant for checking the performance of different configurations, sizings, as well as devising an optimum control strategy regarding the treatment of wastewater (Benedetti et al., 2006; Devisscher et al., 2006; Guerrero et al., 2011; Ciggin et al., 2012). In addition, realistic generation of different realizations of influent time series is one of the most important component of studies that take into account the issue of uncertainty in design and operation of WWTPs (Rousseau et al., 2001; Bixio et al., 2002; Martin et al., 2012; Talebizadeh et al., 2014).

Various approaches have been adopted by different researchers

* Corresponding author.
E-mail addresses: mansour.talebizadehsardari.1@ulaval.ca (M. Talebizadeh), belia@primodal.com (E. Belia), Peter.Vanrolleghem@gci.ulaval.ca (P.A. Vanrolleghem).

http://dx.doi.org/10.1016/j.envsoft.2015.11.005
1364-8152/

for influent generation (for a review see Martin and Vanrolleghem (2014)). One of the simplest approaches in synthetic generation of influent time series is the application of stochastic or regression models with or without periodic components (Capodaglio et al., 1990; Martin et al., 2007; Rodríguez et al., 2013). However, these models may have a poor performance especially during wet weather flow conditions as different complex processes affect the dynamics of the influent. Indeed, such statistical models do not consider the underlying processes that govern the generation and the dynamics of the influent. Langeveld et al. (2014) proposed an empirical model for predicting the influent pollutant time series as a function of influent flow for both dry and wet weather flow conditions (simulation of pollutant time series as a function of flow was also adopted by Rousseau et al. (2001) and Bixio et al. (2002)). Although the proposed model could be used for prediction of pollutant influent time series, it requires a stochastic input (i.e. influent flow) generator if it is to be used for generating different realizations of influent time series.

To consider the underlying phenomena, some researchers have advocated the use of detailed conceptual and/or physically-based models (Hernebring et al., 2002; Temprano et al., 2007). The application of these complex models might be useful for certain purposes, e.g. evaluating the performance of different operating strategies in a sewer system. However, in cases in which the overall behavior of the influent time series is of interest (i.e. the overall variation of influent time series, not all the different phenomena that have resulted in that time series), they might not be very useful as they require very detailed information on the sewer system and running them for a large number of times could be computationally expensive. Besides, even if a detailed sewershed model proves to have a good performance regarding the simulation of the influent time series under a given set of inputs, it cannot be called an influent generator unless a procedure is available for the generation of different realizations of stochastic inputs (e.g. rainfall time series, wastewater time series in dry weather flow (DWF) conditions).

Some researchers have proposed parsimonious conceptual models as an alternative to the complex mathematical models that require detailed information and data (Gernaey et al., 2011). In these models a conceptual view of the main phenomena and interactive processes are formulated in terms of mathematical equations. Despite successful application of these models (at least in giving an overall view of the system), the performance of these models to a great extent depends on the proper choice of model parameters. Since some of the parameters may not have a clear physical meaning they are usually estimated through model calibration. In cases in which there is no measured data available for model calibration, only a rough estimate or a range of values could be inferred from the values reported in literature. In addition, it is almost impossible to have a complete similarity between the model output(s) and the observed values owing to the inextricable uncertainties (e.g. input data uncertainty and/or model structure uncertainty) in any modeling exercise (Belia et al., 2009; Freni and Mannina, 2010).

Given the importance of the issue of uncertainty, several studies have been conducted that consider its effect on both water quality and quantity prediction in urban drainage modeling (Freni et al., 2009; Dotto et al., 2012). However, in these studies, only the effect of model uncertainty under a set of historical rain events (wet weather flow, WWF, conditions) has been considered (i.e. the time series of rainfall and also the contribution of wastewater in DWF conditions were assumed known a priori). In this study not only are we interested in the effects of model uncertainty, but also in the variability of rainfall and influent time series in DWF conditions which significantly affect both the amount and the dynamics of the influent loads.

Considering the shortcomings of the previous studies, this study aims to develop an influent generator which is capable of producing dynamic influent time series of flow and traditional wastewater component concentrations (TSS, COD, TN, TP, NH_4) with 15-min temporal resolution (15-min temporal resolution was assumed to be enough for capturing sub-daily time variations of the influent which could affect the operating parameters and the performance of WWTPs). The proposed methodology will enable users to generate dynamic influent time series that have the same statistical properties as the observed ones using a set of statistical and conceptual modeling tools that only require basic information on climate and characteristics of the sewershed under study. It should be noted that the proposed influent generator is capable of considering the effect of uncertainty in model parameters on the generated influent time series whether the uncertainty can be reduced using observed data (e.g. for the current study) or not (uncertainty in model parameters is characterized by a range of values, determined through expert elicitation or the data from similar sewersheds). In the current study, the variability in inputs (captured by generating different realizations of rainfall and influent time series in DWF conditions, explained in Section 2.1 and Section 2.2, respectively) as well as the uncertainty in model parameters (explained in Section 2.4) on the generated dynamic influent time series are other important issues that will be covered.

2. Methodology

In this paper, a hybrid of statistical and conceptual modeling tools is proposed for synthetic generation of influent time series considering both model parameter uncertainty and input variability. A two-state Markov chain-gamma model (Richardson, 1981) in conjunction with two time series disaggregation methods were used for the stochastic generation of rainfall time series with a high temporal resolution (i.e. 15-min). To generate the influent time series during DWF conditions, taking into account the daily periodic variation, auto-correlation, and cross-correlation in time, a multivariate time series models was developed and its parameters were estimated using the methodology proposed by Neumaier and Schneider (2001). The proposed stochastic model is expected to be superior compared to previous attempts in the generation of influent, as in previous studies the diurnal variation of the influent under DWF conditions was modeled using only univariate time series models (Martin et al., 2007), or by multiplying the daily average influent values to a set of coefficients representing the normalized dynamics of the influent at different times of a day with or without addition of a noise term to the generated time series (Achleitner et al., 2007; Langergraber et al., 2008; Gernaey et al., 2011). The problem resulting from the application of univariate time series models is that the cross-correlation structure that exists among different wastewater constituents may not be respected, as the different constituents are generated independently from the others.

In DWF conditions, the influent time series is generated using multivariate time series models. Conversely in WWF conditions, the outputs of the two statistical models used for the generation of the rainfall and influent time series in DWF conditions are input to a conceptual model for modeling the influent time series in WWF conditions (Fig. 1). In this study the CITYDRAIN model (Achleitner et al., 2007) was selected as the conceptual model owing to its flexibility and parsimony. The CITYDRAIN model of the sewershed under study is calibrated using measured influent data through a Bayesian calibration procedure to account for the total model uncertainty (uncertainty stemming from both model parameters and the distribution of error, i.e. the difference between the observed and simulated time series).

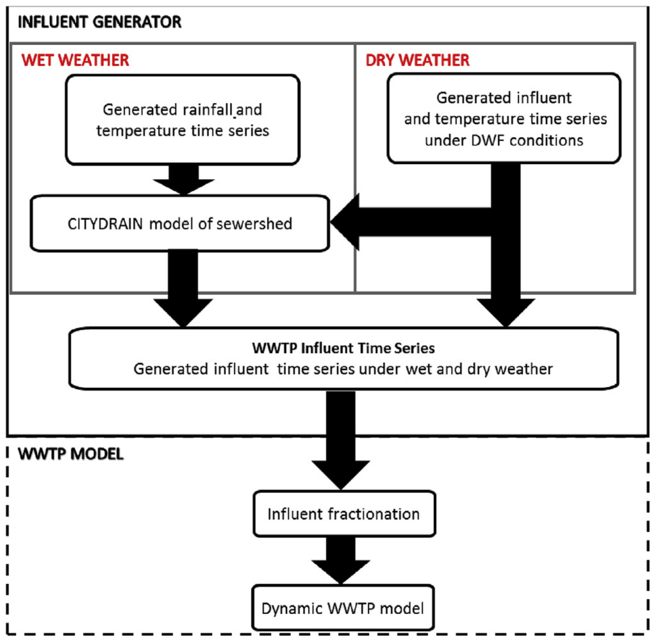

Fig. 1. Schematic of the proposed influent generator.

Finally, different realizations of the influent time series can be generated by running the calibrated CITYDRAIN model using a realization of a generated time series of rainfall and a realization of influent under DWF conditions (i.e. the two stochastic input time series).

Depending on the model to be used for modeling the processes inside the WWTP, an influent fractionation block must be added to convert the generated traditional wastewater composition (COD, TSS, etc) into the state variables of the selected biological models, e.g. the ASM models. Therefore, influent fractionation should be one of the components of WWTP modeling, not a component of the influent generator as different WWTP models may have different state variables.

2.1. Weather and air temperature generators

2.1.1. Daily weather generator

Realistic generation of rainfall time series is crucial as it is one of the most important factors that affect the dynamics of the influent. In this study a stochastic model proposed by Richardson (1981) was used for the synthetic generation of daily rainfall and air temperature time series. According to this method the sequence of dry and wet days is generated using a two-state Markov chain model with parameters $P(W|W)$ and $P(W|D)$ which represent the probability of having a wet day at day t given a wet day at day $t-1$ and the probability of having a wet day at time t given a dry day at time $t-1$ respectively (Fig. 2).

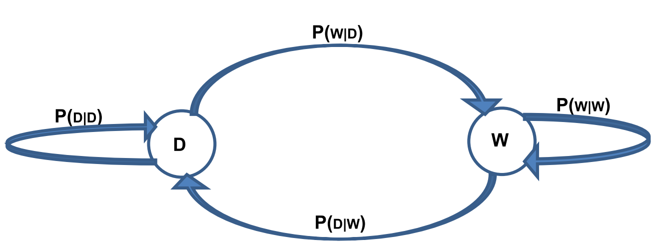

Fig. 2. Schematic of a two-state Markov chain with the two states being wet (W) or dry (D) and four transitions between them.

The other two parameters of the transition matrix needed for the generation of dry and wet days (i.e.$P(D|D)$ the probability of having a dry day at day t given a dry day at day $t-1$ and $P(D|W)$ the probability of having a dry day at day t given a wet day at day $t-1$) can be calculated using Equation (1) and Equation (2).

$$P(D|D) = 1 - P(W|D) \tag{1}$$

$$P(D|W) = 1 - P(W|W) \tag{2}$$

Once the transition probabilities have been estimated from climate data, the sequence of wet and dry days can be generated and the amount of rainfall in each wet day is generated by sampling from a gamma probability distribution (Equation (3)) where x is the depth of daily rainfall, α and β are the two parameters of the distribution (estimated from the measured rainfall time series), and $\Gamma(\alpha)$ represents the gamma function evaluated at α.

$$f(x) = \frac{(x/\beta)^{\alpha-1}\exp(-x/\beta)}{\beta\,\Gamma(\alpha)} \tag{3}$$

It should be noted that the seasonal variation of the daily rainfall generator parameters (Markov chain transition probabilities, i.e. $P(W|D)$ and $P(W|W)$) as well as the parameters of the gamma distribution (i.e. α and β) were taken into account by fitting different Fourier series models on the parameter values derived from rainfall records. To do so, each year with rainfall records was divided into 26 two-week time spans and then the transition probabilities were estimated by dividing the number of wet days preceded by a dry day by the total number of days (for estimating$P(W|D)$) and also dividing the number of wet days preceded by a wet day by the total number of days. Moreover, the parameters of the gamma distributions were calculated for each two-week time span using the maximum likelihood method. Once the parameters of the Markov chain-gamma model are estimated for each two-week time span, different Fourier series are fitted on the estimated parameters to provide a value for each day of the year.

The parameters of the daily rainfall model can be calculated for different regions using regional rainfall records. In addition, different values of parameters reflecting future climate change scenarios could be elicited from experts and used for generating daily rainfall time series.

The time series of minimum and maximum air temperature are generated conditioned on the state of the day (i.e. wet or dry) using a multivariate linear first-order time series model (Matalas, 1967). Starting points are a time series of daily maximum and minimum temperature values.

The seasonal variation in mean and standard deviation of maximum and minimum daily temperature values for dry and wet days are captured (in two Fourier series models) and subtracted from the data to derive the residual time series of maximum and minimum temperature (Equation (4) and Equation (5)).

$$Y_i^d(j) = \frac{X_i^d(j) - \overline{X}_i^d(j)}{\sigma_i^d(j)} \quad \textit{for dry days} \tag{4}$$

$$Y_i^w(j) = \frac{X_i^w(j) - \overline{X}_i^w(j)}{\sigma_i^w(j)} \quad \textit{for wet days} \tag{5}$$

In the above equations $\overline{X}_i^d(j)$ and $\sigma_i^d(j)$ are the mean and standard deviation for a dry day, $\overline{X}_i^w(j)$ and $\sigma_i^w(j)$ are the mean and standard deviation for a wet day, and $Y_i(j)$ is the residual component for transformed variables (i.e. j=1 for maximum temperature, and j=2 for minimum temperature). Once the residual time series are derived, a multivariate time series model as proposed by Matalas (1967) is fitted on the residual time series of maximum and minimum temperatures (Equation (6)).

$$Y_i(j) = AY_{i-1}(j) + B\varepsilon_i(j) \tag{6}$$

In the above equation $Y_i(j)$ is a 2 × 1 matrix for day i whose elements are residuals of maximum temperature (j=1) and minimum temperature (j=2). $Y_{i-1}(j)$ is a 2 × 1 matrix of the previous day's residuals, ε_i is a 2 × 1 matrix of independent random components (the noise term is assumed to be a normal, independent, and identically distributed (i.i.d) variable with zero mean and unit variance), and A and B are 2 × 2 matrices whose elements are derived according to (Equation (7)) and (Equation (8)):

$$A = M_1 \cdot M_0^{-1} \tag{7}$$

$$B \cdot B^T = M_0 - M_1 \cdot M_0^{-1} \cdot M_1^T \tag{8}$$

where the subscript −1 denotes the inverse of matrix and M_0 and M_1 are defined as:

$$M_0 = \begin{bmatrix} 1 & \rho_0(1,2) \\ \rho_0(1,2) & 1 \end{bmatrix} \tag{9}$$

$$M_1 = \begin{bmatrix} \rho_1(1,1) & \rho_1(1,2) \\ \rho_1(2,1) & \rho_1(2,2) \end{bmatrix} \tag{10}$$

where $\rho_0(j,k)$ is the correlation coefficient between variables j and k on the same day where j and k may be set to 1 (maximum temperature) or 2 (minimum temperature). $\rho_1(j,k)$ is the correlation coefficient between variable j and k lagged one day with respect to variable j.

The above weather generator is suited for random generation of daily rainfall and temperature. However, in this study we need to generate rainfall time series with a finer temporal resolution than daily resolution (15-min temporal resolution, comparable to the temporal resolution of rainfall in the BSM influent model (Gernaey et al., 2011)). Some methodologies have been proposed for random generation of hourly rainfall time series based on historical hourly rainfall data (Pattison, 1965; Rodriguez-Iturbe et al., 1987). However, long-term hourly rainfall data may not be available in every region and using a limited hourly rainfall record for random generation of long-term hourly rainfall time series may result in misrepresentation of the inter-annual variability in rainfall.

In this study the proposed Richardson-based weather generator (used for daily rainfall generation) was combined with two time series disaggregation techniques. Daily rainfall time series is first generated using the Richardson (1981) method and then two time series disaggregation models, including a daily-to-hourly model (Koutsoyiannis and Onof, 2001) and an hourly-to-15-min model (Ormsbee, 1989) are applied for generation of long-term rainfall time series with 15-min temporal resolution.

2.1.2. Daily to hourly rainfall time series disaggregation

The time series disaggregation method proposed by Koutsoyiannis and Onof (2001) is used in the proposed methodology to disaggregate the synthetic daily rainfall time series (i.e. generated using the previously explained Richardson-based weather generator) into hourly rainfall. The proposed disaggregation method combines the Bartlett–Lewis stochastic rainfall model (Rodriguez-Iturbe et al., 1987) with an adjusting algorithm so that the total amount of hourly rainfall in each day becomes consistent with its corresponding daily value. A general description of the Bartlett–Lewis model can be summarized as follows (Koutsoyiannis and Onof, 2001):

1) Storm origins (t_1, t_2, t_3 in Fig. 3) occur according to a Poisson process with rateλ.
2) Arrival times of the contributing cells in a storm (t_{21}, t_{22}, t_{23} for Storm2 in Fig. 3) occur according to a Poisson process with rateβ.
3) Cell arrival terminates after time v_i (v_2 for Storm2 in Fig. 3) that is exponentially distributed with parameterγ.
4) Each cell has a duration that is exponentially distributed with parameterη.
5) Each cell has a uniform intensity (R_{21}, R_{22}, R_{23}, R_{24} for Storm2 in Fig. 5) coming from an exponential distributionμ.

The parameters of the Bartlett–Lewis rainfall model can be calculated from hourly rainfall records (Rodriguez-Iturbe et al., 1987; Koutsoyiannis and Onof, 2001) and then the model can be used for synthetic generation of hourly rainfall time series. However, the hourly rainfall time series generated using the Bartlett–Lewis model should be adjusted so that the sum of hourly rainfall time series in each day becomes consistent with its corresponding daily value.

In the proposed methodology, a simple adjusting procedure known as the proportional adjusting procedure (Koutsoyiannis and Onof, 200) is used. According to this procedure the initially generated hourly rainfall values ($\widetilde{X}_s$) are adjusted to new values (X_s) using Equation 11

$$X_s = \widetilde{X}_s \left(\frac{Z}{\sum_{i=1}^{k} \widetilde{X}j} \right) \quad (s = 1, 2, 3, \ldots k) \tag{11}$$

where Z is the amount of daily rainfall (generated using the Richardson-based daily rainfall generator), and k is the number of hourly rainfall values within a day.

2.1.3. Hourly to 15-min rainfall time series disaggregation

The disaggregated hourly rainfall time series in Section 2.1.2 is further disaggregated to 15-min rainfall time series using the empirical time series disaggregation procedure proposed by Ormsbee (1989). According to this empirical model, four types of rainfall patterns are identified (Fig. 4) for each 3-h sequence of hourly rainfall. After determining the type of sequence, the amount of rainfall at the central hour of each 3-h rainfall sequences is disaggregated into a time series with a desired temporal resolution (15 min here).

The probability distribution function, *F(t)* of sub-hourly rainfall

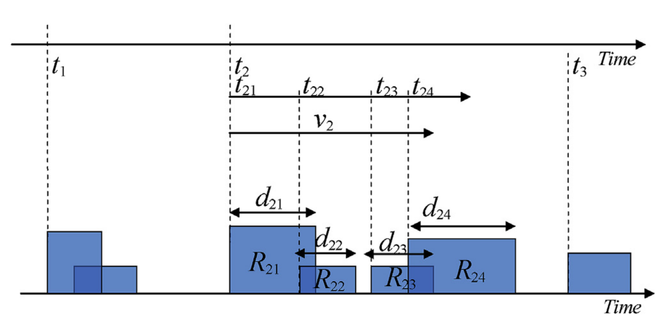

Fig. 3. Schematic of Bartlett–Lewis rainfall model.

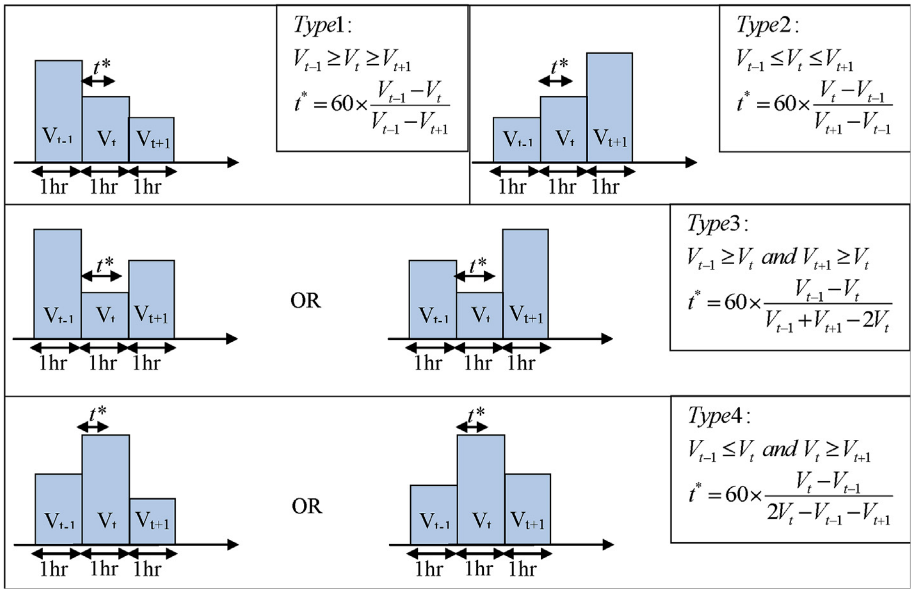

Fig. 4. Four types of rainfall patterns for a 3-h rainfall sequence (Ormsbee (1989).

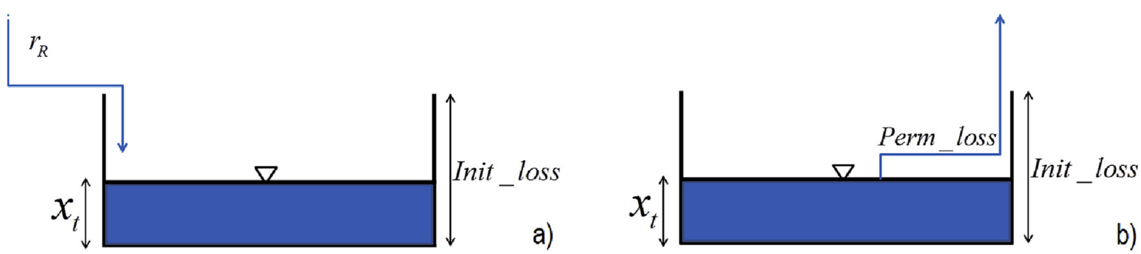

Fig. 5. Schematic of virtual basin. a) filling of the basin in wet periods where the amount of spilled water is multiplied by the runoff coefficient for calculating the amount of effective rainfall, b) emptying process with a fixed rate (permanent loss) in dry periods (see Equation (14)).

intervals of the central hour rainfall (V_t) is calculated using the time parameter t^* and the amounts of rainfall in the first and third hour (V_{t-1} and V_{t+1}) of each 3-h rainfall sequence (Equation (12)).

$$F(t) = \begin{cases} \frac{V_{t-1}t}{V_t^*} - \frac{(V_{t-1} - V_t)t^2}{2V_t^* t^*} & \text{for } 0 \leq t < t^* \\ \frac{(V_t + V_{t-1})t^*}{2V_t^*} + \frac{V_t(t - t^*)t^2}{V_t^*} - \frac{(V_t - V_{t+1})(t - t^*)^2}{2V_t^*(60 - t^*)} & \text{for } t^* \leq t \leq 60 \end{cases} \quad (12)$$

For the proposed methodology the central hour rainfall (V_t) is disaggregated into 4 intervals and the portion of each interval is calculated by multiplying the probability of each interval to the central hour rainfall.

2.1.4. Bioreactor temperature

The explained Richardson-based weather generator is suited for the generation of daily air temperature which could serve as an input for modeling the temperature effect of the influent or in the bioreactors of a WWTP. Bioreactor temperature is of particular interest as it affects the rate of many biological processes taking place in the bioreactors (Antoniou et al., 1990). Whereas, based on the influent temperature a heat balance could be constructed around the bioreactors of the WWTP to calculate the bioreactor temperature (see Gillot and Vanrolleghem (2003) for details), preference is often given to directly input the bioreactor temperature in the WWTP model (e.g. Gernaey et al. (2014)).

To estimate the bioreactor temperature a simple linear regression model between the concurrently measured daily air and bioreactor temperatures is used. Once the parameters of the linear regression model (which calculates the daily bioreactor temperature as a function of daily air temperature) have been estimated, it can be used to convert the generated daily air temperatures (generated using the Richardson-based weather generator) to daily bioreactor temperatures. It should be noted that the variation in bioreactor temperature is not solely function of air temperature (Gillot and Vanrolleghem, 2003). However, calculating the daily bioreactor temperature as a function of daily air temperature would capture the seasonal variation of bioreactor temperature time series.

To further disaggregate the daily bioreactor temperature time series into a time series with 15-min temporal resolution, the average normalized pattern representing the diurnal variation of bioreactor temperature is estimated by a Fourier series and multiplied to the daily bioreactor temperature values to obtain a bioreactor temperature time series with 15-min temporal resolution.

2.2. Influent generation in DWF conditions

The influent time series in DWF conditions usually shows specific periodic patterns which can be mainly attributed to the socio-economic fabric of society and also to the physical characteristics of the wastewater collection system. To mimic these variations in time, it is common practice to estimate representative values (e.g. multiplying flow per person to the total population for estimating flow) for flow and loads and then multiply them to a set of normalized coefficients reflecting diurnal, weekly and seasonal time variation of the influent time series (Gernaey et al., 2011; Flores-Alsina et al., 2014). Moreover, Gernaey et al. (2011) proposed to add a noise term to the deterministic influent profile in order to avoid generating the same influent time series in subsequent days. In this study, the effect of rainfall on the contribution of infiltration (rainfall induced infiltration) is not considered explicitly. Rather, the application of a multivariate auto-regressive model (Neumaier and Schneider, 2001) with periodic components is proposed.

To estimate the parameters of the proposed time series model, an influent time series measured during DWF conditions is to be extracted and analyzed for estimating the parameters of the multivariate auto-regressive model. First, the seasonal (e.g. associated to groundwater infiltration) and diurnal periodic components of flow and other wastewater constituents are to be estimated using different Fourier series approximations (depending on the underlying expected periodic patterns, e.g. a bimodal periodic pattern for flow in urban sewersheds) and removed from the original influent time series to calculate the residual time series. The zero-mean residual time series of influent flow and composition is then further standardized to have an influent time series with a zero mean and unit standard deviation. The parameters of the multivariate autoregressive model in Equation (13) (i.e.p, A_l, C) are then estimated through a stepwise least square algorithm proposed by Neumaier and Schneider (2001).

$$v_t = \sum_{l=1}^{p} A_l \times v_{t-l} + \varepsilon_t \tag{13}$$

In Equation (13), v_t is an m-dimensional vector (i.e. for our application m = 5 which corresponds to the flow and the four wastewater compositions) containing the generated influent component at time t, p is the order of the auto-regressive model (p is to be selected based on Schwarz's (1978) Bayesian Criterion SBC, based on the fitting results). More details can be found in Neumaier and Schneider (2001)), $A_1,...,A_p$ are the coefficient matrices of the auto-regressive model, and ε_t is a noise term generated from an uncorrelated zero-mean multivariate normal distribution with covariance matrix C (i.e. $\varepsilon_t \sim N(0,C)$). Different realizations of the residual influent time series can be generated using this time series model and converted to the original scale depending on the mean and standard deviation of the original influent time series.

2.3. Influent generation in WWF conditions

Synthetic generation of the influent time series during WWF conditions is relatively more complicated than the generation of the influent time series during DWF conditions. Difficulties arise as various phenomena are occurring during WWF conditions and as the availability of measured data is usually scarce for these periods. Hence, using a purely statistical model may result in significant discrepancies between simulated and observed time series. Therefore, we used a combination of statistical modeling techniques and a conceptual model to generate the time series of the influent during WWF conditions. The CITYDRAIN model (Achleitner et al., 2007) was selected as the conceptual model as it is open source (inside Matlab) and it takes into account the basic phenomena (see 2.3.1 and 2.3.2) that govern the amount and dynamics of the influent. Also, it requires the estimation of only a small number of parameters whose values or ranges of values can be inferred from the basic information of a sewershed.

2.3.1. Flow

CITYDRAIN calculates the amount of effective rainfall by adopting the concept of a virtual basin (Achleitner et al., 2007). According to this concept (Fig. 5), effective rainfall is calculated by subtracting the initial loss from total rainfall and then multiplying it with the runoff coefficient. Permanent losses like evapotranspiration are considered only in dry periods to mimic an emptying process of the virtual basin (Equation (14)).

$$\begin{cases} h_e = Max(r_R - (Int\_loss - x_t) \times Runoff\_coeff, 0) & \textit{Wet periods} \\ \dfrac{dx}{dt} = -Perm\_loss & \textit{Dry periods} \end{cases} \tag{14}$$

In the above equation, h_e(mm/t) represents the effective rainfall, r_R(mm/t) the total rainfall, Int_loss(mm) the initial loss, $Runoff\_coeff$(−) the runoff coefficient, and $Perm\_loss$(mm/t) the permanent loss (overall, three parameters).

The height of the effective rainfall is then multiplied by the fraction of sewershed area which contributes to the generation of runoff to calculate flow. The routing method, proposed by (Motiee et al. (1997) that is based on a simplified form of the Muskingum flow routing equations (Roberson et al., 1995) is then used for routing flow and pollutants inside the sewer system.

2.3.2. Composition

For the generation of pollutant time series in WWF conditions, CITYDRAIN uses a rather simplistic approach in which a fixed pollutant concentration is imposed to the system (Equation (15)):

$$\begin{cases} C(t) = C & \text{if } h_e > 0 \\ C(t) = 0 & \text{if } h_e = 0 \end{cases} \tag{15}$$

where, $C(t)$ is the generated pollutant concentration in time, C is a model parameter representing the concentration in WWF conditions, and h_e is the effective rainfall. Given the importance of the influent time series in WWF conditions, a more appropriate conceptual model was used for simulating the accumulation-wash off processes corresponding to the particulate concentrations. To this aim, a new block was developed and implemented in CITYDRAIN to generate the pollutant concentration time series in WWF conditions. Equation (16) shows the mathematical formulation of the selected accumulation-wash off model (Kanso et al., 2005).

$$\begin{cases} \text{Accumulation model}: \dfrac{dM_{(t)}}{dt} = K_a \times \left(m_{\lim} \times S_{imp} - M_{(t)}\right) \\ \text{Wash off model}: \dfrac{dM_{(t)}}{dt} = -W_e \times I_{(t)}{}^{w} \times M_{(t)} \end{cases} \tag{16}$$

where, $M_{(t)}$ is: the available pollutant mass on the sewershed surface at time t (kg), K_a is the accumulation coefficient (1/day), m_{lim} is the maximum accumulated mass (kg/ha), S_{imp} is the impervious area (ha), $I_{(t)}$ is the rainfall intensity (mm/hr), W_e, and w are calibration parameters to be estimated using observed rainfall and influent data. In case of the availability of influent data, the CITYDRAIN model parameters can be calibrated (i.e. their uncertainty reduced through Bayesian calibration (see Section 2.4)). However, if there is no measured influent data, the uncertainty in the parameters (i.e. parameters in Equation (14) and Equation (16)) will remain and the uncertainty on the generated influent time series could be larger. Other parameters like the sewershed area or maximum conveyance capacity of the sewer system that have physical meaning can be obtained from general information on the sewershed.

2.3.3. Model setup

In the CITYDRAIN model, the components of a sewershed system are modeled by a set of sewershed blocks and depending on the availability of data and level of heterogeneity in the sewershed, users may choose different numbers of blocks for modeling the entire sewershed. However, it should be noted that increasing the number of blocks will results in an increase in the number of model parameters which in turn could cause difficulties in parameter estimation (e.g. an unrealistic number of simulations (Martin and Ayesa, 2010)) when the model is to be calibrated using the measured flow and water quality data. In the modeling step, different CITYDRAIN configuration should be tested and a decision made on the best one.

2.4. Bayesian model calibration of the CITYDRAIN sewer model

As explained in the previous section, the dynamics of the influent time series under WWF conditions is modeled using the CITYDRAIN model. However, one should be aware of the fact that modeling the influent time series under WWF conditions using a conceptual model may not lead to reliable results unless the model is calibrated and the effect of different sources of uncertainties on the model outputs (i.e. flow and other pollutants) are taken into account. To this aim, a Bayesian estimation framework was used to update the initial ranges of values (i.e. in a Bayesian parameter estimation method, the initial probability distributions or prior distributions reflect the initial knowledge on the value of uncertain model parameters) that were assigned to the parameters of the CITYDRAIN model (i.e. estimating the posterior distribution of parameters using their prior distribution and the measured data on flow and pollutant concentrations (see Table 5 for the uncertain model parameters)). In general, the posterior distribution of parameters using Bayes' theorem can be formulated by Equation (17).

$$h(\theta|Data) = \frac{f(Data|\theta)p(\theta)}{f(Data)} \tag{17}$$

where $h(\theta|Data)$ is the posterior distribution, $p(\theta)$ is the prior distribution, $f(Data)$ is merely a proportionality constant so that $\int h(\theta|Data) = 1$, and $f(Data|\theta)$ constitutes the likelihood function which measures the likelihood that the data correspond to the model outputs with parameter set $\boldsymbol{\theta}$. Assuming homoscedastic uncorrelated Gaussian error (i.e. having normal distribution with the same variance and no correlation in time), the likelihood function can be formulated according to Equation (18) (Bates and Campbell, 2001; Marshall et al., 2004).

$$f(Data|\boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-n/2} \prod_{t}^{n} exp\left\{-\frac{[Data_t - R(x_t;\boldsymbol{\theta})]^2}{2\sigma^2}\right\} \tag{18}$$

where n is the number of observations, σ^2 is the variance of the residual error (i.e. the difference between model predictions and observed values which equals the measurement error if it is assumed that the model is perfectly representing reality), $Data_t$ is the observed variable at time t, x_t is the set of inputs at time t, $\boldsymbol{\theta}$ is the set of model parameters and $R(x_t;\boldsymbol{\theta})$ represents the model output as a function of x_t and $\boldsymbol{\theta}$.

A specific form of Markov chain Monte Carlo (MCMC) sampler known as differential evolution adaptive Metropolis or DREAM (Vrugt et al., 2008) is used to efficiently estimate the posterior distribution of the CITYDRAIN model parameters that are involved in the generation and routing of flow and pollutants in WWF conditions, given the time series of flow and influent composition. It should be noted that the proposed Bayesian approach is not only capable of capturing the effect of model parameter uncertainty, but also of capturing the effect of other sources of uncertainties (model structure, input, etc) that could result in some discrepancies between the simulated influent time series and the observed series.

2.5. Synthetic influent generation

Once the uncertainty ranges of the CITYDRAIN model parameters are updated, synthetic influent time series for a desired number of years considering the variability in the inputs of the CITYDRAIN model (i.e. rainfall and influent time series in DWF conditions) and also the total uncertainty can be obtained as follows:

1. Synthetic generation of the 15-min time series of rainfall for one year
2. Synthetic generation of the 15-min time series of the influent under DWF conditions for one year
3. Sampling a point from the posterior distribution of the CITYDRAIN model parameters
4. Inputting the generated time series 1) and 2) and the parameters sampled in 3) and running the CITYDRAIN model for one year
5. Repeating 1) to 4) for a desired number of years

In this study, the contribution of the noise term (i.e. characterized using a normal distribution with zero mean and standard deviation of Sigma) to the output is treated as a source of variability. This decision is based on the assumption that the main part of the difference between the simulated and observed signals is due to an actual fluctuation of the influent time series (an instance of variability) or some random measurement error (an instance of uncertainty).

Obviously during the DWF conditions, the influent time series is generated using the statistical model that is explained in Section 2.2 and the CITYDRAIN model has no effect on the generated influent time series.

3. Data and case study

The Eindhoven WWTP with a design capacity of 750,000

Table 1
Summary of the data type and their applications.

Types of data	Application
Rainfall (1951–2013)	✓ Estimation of the rainfall generator parameters (i.e. $P(W\|D)$, $P(W\|W)$, α, β ✓ Calibration of the CITYDRAIN model
Minimum and maximum air temperature	✓ Calibration of the air temperature generator
Bioreactor temperature	✓ Calibration of the regression model for the generation of bioreactor temperature as a function of air temperature
Influent data in DWF conditions	✓ Calibration of the DWF generator
Influent data in WWF conditions	✓ Calibration of the CITYDRAIN model

population equivalent (PE) is the third largest WWTP in the Netherlands. The sewershed served by the Eindhoven WWTP has a total area of approximately 600 km^2 and comprises of three main sub-sewersheds called Nuenen/Son, Eindhoven Stad, and Riool-Zuid. The influent data used in this study comprised of sensor data of flow, ammonia (measured using an ion-selective sensor) soluble COD, total COD, and TSS (the latter three measured using an UV/VIS-based sensor) in the period of September 2011 to September 2012 at the outlet of the Nuenen/Son, Eindhoven Stad, and Riool-Zuid sub-sewersheds (entrance point to the treatment plant).

The long-term daily rainfall data and also rainfall data with finer temporal resolution provided by KNMI (Royal Netherlands Meteorological Institute De Bilt, The Netherlands) and Waterschap De Dommel (Boxtel, The Netherlands) were used for estimating the parameters of the weather generator proposed in this paper. Table 1 shows a summary of the data types and also their applications in the development of the proposed influent generator.

4. Results and discussion

This section presents the outputs and some discussion on the results of different components of the proposed influent generator. As explained in the methodology section, the parameters of the statistical models used for synthetic generation of rainfall, air and bioreactor temperature, as well as the multivariate auto-regressive time series models (used for the generation of influent time series in DWF conditions) were calibrated using the historical weather data and observed influent time series in DWF conditions. The parameters of the CITYDRAIN model (the conceptual model for modeling influent time series under WWF conditions) were estimated using the Bayesian estimation framework. Once the parameters of the statistical and the conceptual models were calibrated, different realizations of influent time series would be generated by running the CITYDRAIN model with different realizations of the rainfall and influent time series under DWF conditions (stochastic inputs) and different sets of parameters sampled from the posterior distribution of the CITYDRAIN model parameters.

The performance of the weather generator and the influent generator under DWF conditions were evaluated by comparing the statistical properties of the generated time series with those of the historical time series. The results corresponding to the Bayesian calibration of the CITYDRAIN model are explained and at the end a 7-day snapshot of a generated one year influent time series is presented and discussed.

4.1. Synthetic generation of rainfall

The parameters of the statistical Markov-gamma model were estimated using different Fourier series models fitted on parameter values derived from the recorded rainfall data (Fig. 6) in the studied

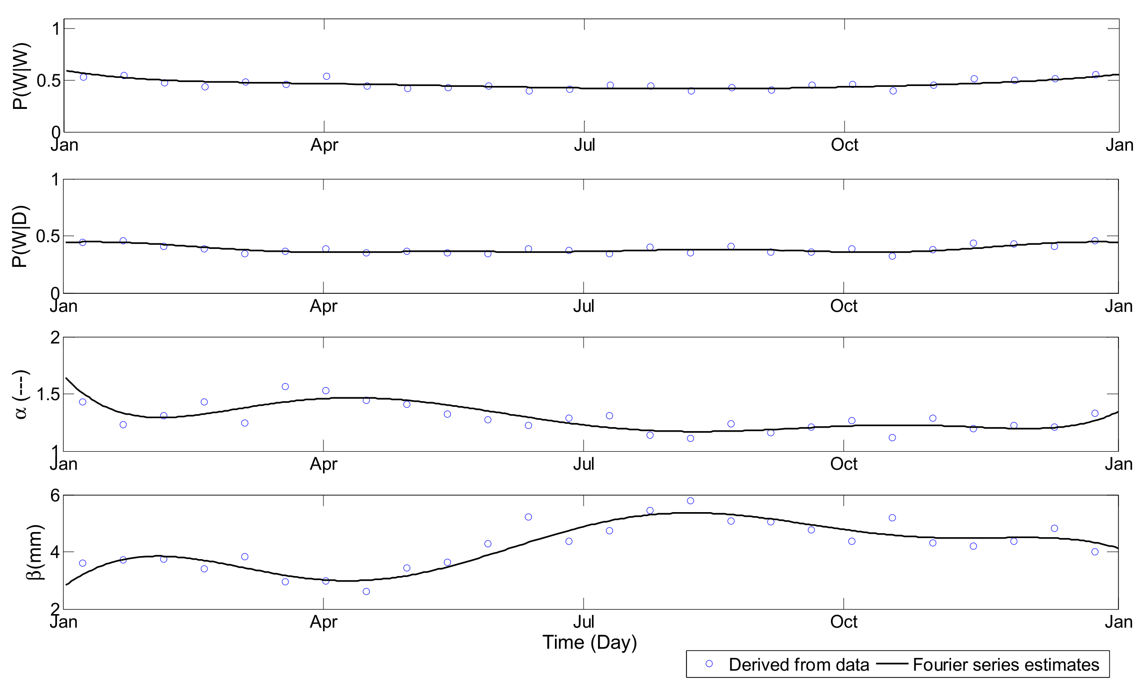

Fig. 6. Seasonal variation in the Markov chain-gamma model parameters.

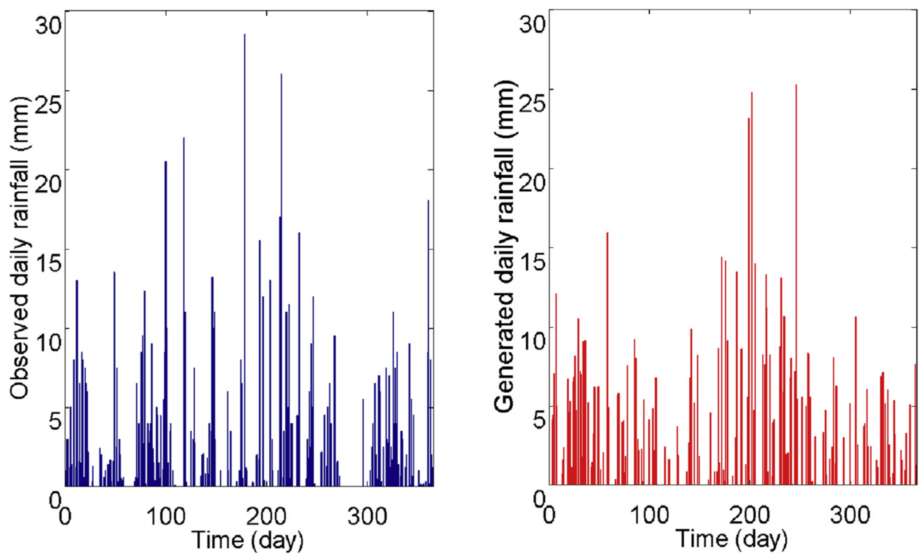

Fig. 7. A year-long realization of daily rainfall (right) versus an observed one (left).

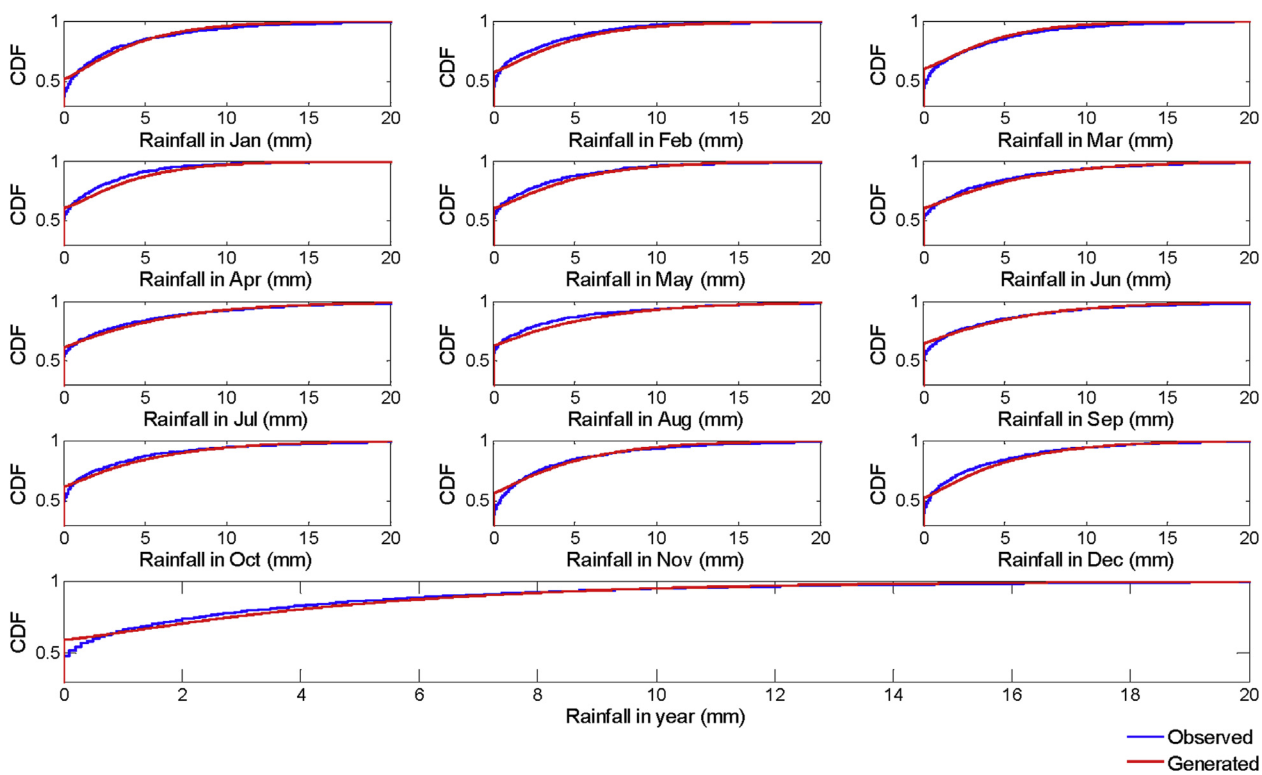

Fig. 8. Cumulative distribution function of daily rainfall in the studied Eindhoven catchment.

Eindhoven catchment and then different realizations of rainfall time series (with a random seed used for each year long generation of rainfall time series) were generated (Fig. 7).

To evaluate the performance of the Markov-gamma model for realistic generation of rainfall time series, the CDF of the observed rainfall time series (CDF curves include both wet and dry days) corresponding to different seasons were compared with those of the generated rainfall time series (Fig. 8 for the CDF comparison and Fig. A1 for a q–q plot between the simulated and observed rainfall time series).

The observed CDFs in Fig. 8 are constructed using the daily rainfall records between 1951 and 2013 and the generated CDFs correspond to 1000 years of synthetic rainfall time series, generated using the explained Markov chain-gamma model whose parameters (depicted in Fig. 6) were estimated from the daily rainfall records (i.e. from 1951 to 2013).

The results indicate that not only are the basic yearly statistics (i.e. average and variance) of the generated rainfall time series consistent with the recorded rainfall time series, but also the seasonal variations in rainfall intensity and frequency of wet days are

Table 2
Average rainfall amount and number of wet days for the Eindhoven catchment.

Month	Amount of rainfall[a] (mm)		Average number of wet days[b]	
	Observed	Generated	Observed	Generated
January	72.3	67.0	16	14
February	52.0	57.0	12	11
March	63.4	54.4	13	12
April	44.1	51.9	12	11
May	58.3	60.9	12	12
June	68.0	68.4	12	11
July	74.7	73.5	12	11
August	64.6	71.0	11	11
September	67.9	62.1	12	10
October	62.0	65.0	12	11
November	71.1	66.4	15	12
December	70.0	74.0	14	14
Annual	768	772	152	141

[a] The average amount of total rainfall in different months for observed (i.e. rainfall data from 1951 to 2013) and generated rainfall time series (i.e. 1000 years of rainfall data, generated using the proposed rainfall generator).
[b] The average number of wet days in different months for observed (i.e. rainfall data from 1951 to 2013) and generated rainfall time series (i.e. 1000 years of rainfall data, generated using the proposed rainfall generator).

Table 3
Basic statistics of hourly rainfall data for the Eindhoven catchment.

Statistics	Unit	Observed Value	Simulated Value
Mean	mm	0.08	0.08
Standard deviation	mm	0.60	0.60
Lag 1 auto-correlation[a]	–	0.33	0.36
Fraction of dry hours	–	0.92	0.94

[a] Correlation between the amount of rainfall at time t and $t-1$.

respected. As indicated in Table 2, the differences between the observed average rainfall and the simulated average rainfall (based on the rainfall data from 1951 to 2013 and 1000 years of synthetic rainfall time series, respectively) in the different months are below 10%, except for the months March and April in which the differences are −16.5% and 15% respectively. Nevertheless, all the differences between the average simulated and average observed rainfall values are below 20% which is acceptable according to the work of Richardson (1981) in which the differences between average simulated and observed values in some months are above 20%. The discrepancies between certain percentiles of the simulated and observed rainfall distributions (Fig. 8 or Fig. A1) could be associated to the difference between rainfall generator parameters derived from observed data (sample parameters, i.e. blue circles in Fig. 6) and those that are estimated using different Fourier time series and used for the rainfall generator, i.e. solid lines in Fig. 6). In addition, the difference between the length of observed and generated rainfall time series (i.e. 62 years of observed data compared to 1000 years of generated rainfall data) could be another reason for the difference between the extreme percentiles (see Fig. A1).

Moreover, Table 3 shows that the hourly time series of rainfall which was generated using the time disaggregation method (i.e. disaggregation of daily to hourly time series) has the same statistical characteristics as the observed one. Overall, the synthetic generation of rainfall in which the statistical properties of the time series is respected across different time scales is a significant improvement compared to the rainfall generation in for instance the BSM influent generator (Gernaey et al., 2011) in which there is no clear way for extracting and incorporating the statistical properties of available recorded rainfall data into synthetic rainfall time series generation. In addition, the flexibility of the proposed rainfall generator allows users to define different scenarios reflecting future changes in precipitation regime (e.g. due to climate change (Chen et al., 2010)) and its effect on the influent time series (e.g. what would happen if the amount of rainfall or the number of wet/dry days increases by 20% in specific seasons, a feature that is not available in previous rainfall generators, e.g. the rainfall generator proposed by Gernaey et al. (2011)).

4.2. Synthetic generation of air and bioreactor temperature

The seasonal variation in the mean maximum air temperature, the mean minimum air temperature, the standard deviation of the maximum air temperature, and the standard deviation of the minimum air temperature for dry and wet days, captured using Fourier series models are illustrated in Fig. 9 a to Fig. 9d. As indicated, the mean values for both maximum and minimum air temperatures have an upward trend from the winter until the midst of summer (when they reach their maximum values), followed by a downward trend until they reach their minimum values in the winter again. However, comparing Fig. 9a and b, representing the seasonal variations for wet and dry days suggest that there is no significant difference between the seasonal variation when the state of day (wet or dry) is taken into account. In other words, it can be concluded that for the case study of this research, the variation in mean maximum and mean minimum temperatures is mostly a function of the seasons of the year rather than the state of the day.

As explained in Section 2.1, a multivariate linear first-order model was fitted on the residual time series of maximum and minimum air temperatures for synthetic generation of maximum and minimum air temperatures and in the end the generated air time series were converted to their original values through Equation (4) and Equation (5) using the seasonal mean and standard deviation values illustrated in Fig. 9a to d.

The daily temperature of the bioreactor was generated through a linear regression model which relates the daily average bioreactor temperature to the daily average air temperature. Fig. 9e and g show random generation of an air and bioreactor temperature time series for one year. The linear model in Fig. 9f which was developed using the concurrently measured air and bioreactor temperature for one year (i.e. September 2011 to September 2012, illustrated in Fig. 9i) shows that the average bioreactor temperature can be estimated reasonably (R^2=0.70) as a linear function of air temperature. It should be noted that the effect of the state of day (i.e. dry or wet) on air temperature (although for the current case study it was not significant) which in turn affects the bioreactor temperature has, been taken into account in random generation of the air temperature. Moreover, an attempt to use two different regression models depending on the state of day (i.e. one regression model for dry days and another one for wet days) did not result in any improvement in the prediction of bioreactor temperature as a function of air temperature.

The average diurnal variation of bioreactor temperature in Fig. 9h was extracted by fitting a first order Fourier series estimate to the normalized bioreactor temperature variations which in turn was used for converting the daily bioreactor temperature time series into a time series with 15-min temporal resolution.

Despite the fact that the diurnal variation pattern in Fig. 9f clearly shows a periodic behavior in time (which corresponds to the diurnal variation of bioreactor temperature), there is no significant difference between the highest and lowest temperature throughout a day (i.e. the highest temperature is only around 1.001 times the daily average bioreactor temperature and the lowest temperature is around 0.9985 times the daily average bioreactor temperature). Therefore, in practical applications (at least for the

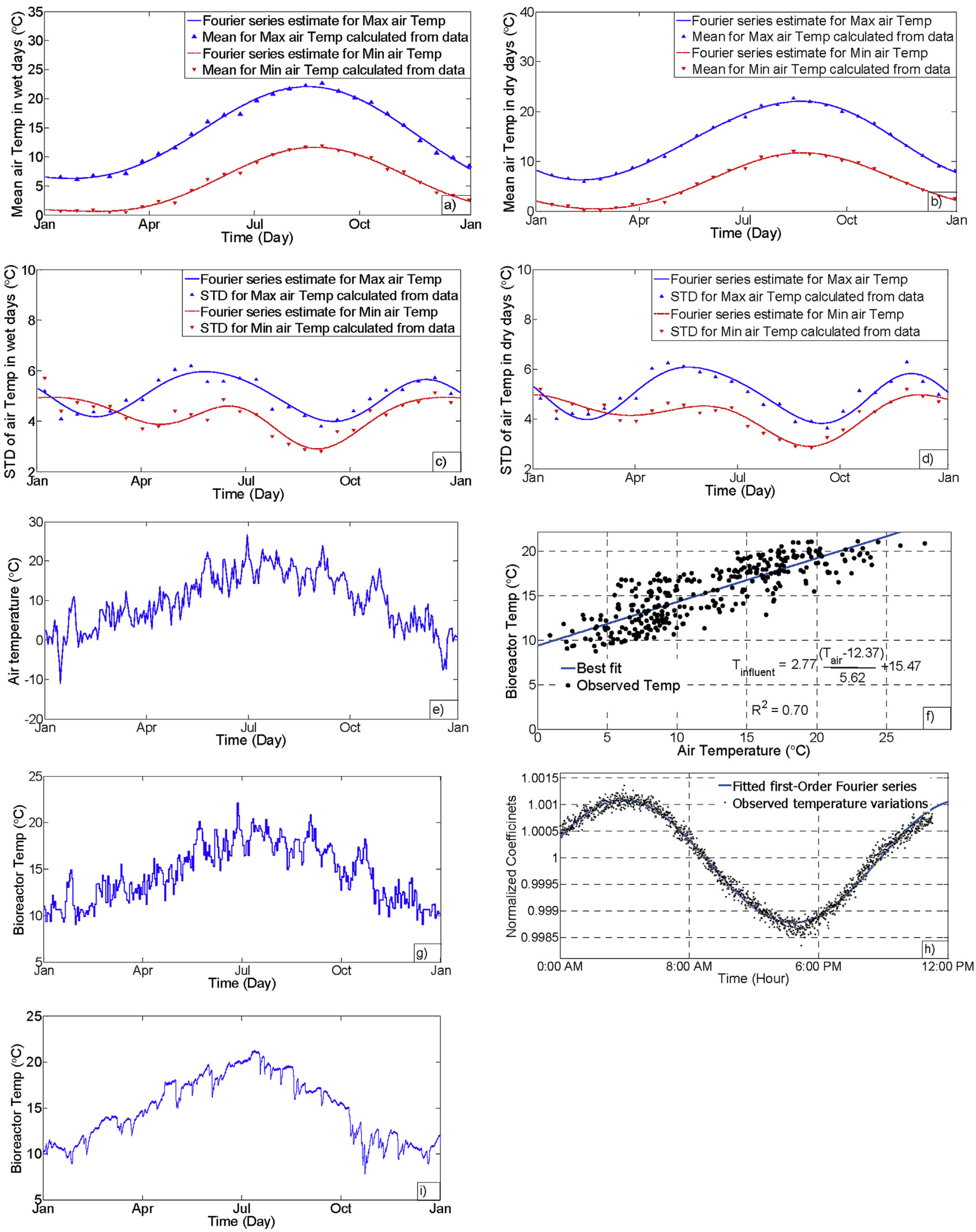

Fig. 9. Random generation of air and bioreactor temperature for one year for the Eindhoven WWTP: Seasonal variation in mean maximum and mean minimum air temperatures in wet days (a), in dry days (b), Seasonal variation in standard deviations of maximum and minimum air temperatures in wet days (c), in dry days (d), Randomly generated daily air temperature (e), Linear regression model between daily air and bioreactor temperatures (f), Bioreactor temperature time series with 15-min temporal resolution (g), Average diurnal variation of bioreactor temperature (h), Observed historical bioreactor temperatures used in the analysis (i).

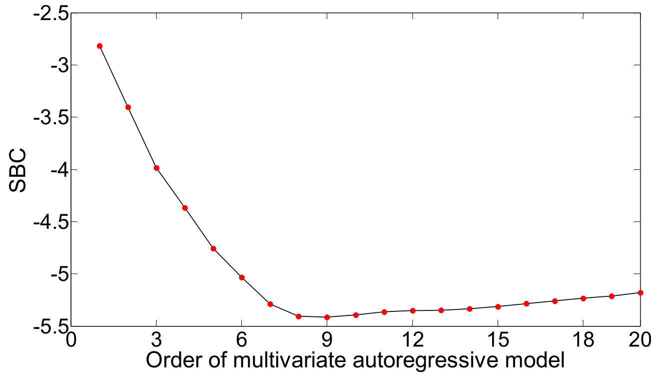

Fig. 10. Variation of SBC with order of multivariate time series model.

case study in this research), the diurnal temperature variation can be ignored.

4.3. Multivariate auto-regressive model for DWF generation

Influent data corresponding to 82 dry days were analyzed for estimating the parameters of the multivariate auto-regressive time series model for DWF generation (Table A2). As explained, the order of the multivariate auto-regressive model was determined based on the SBC criterion and the parameters were estimated according to a specific least square algorithm proposed by Neumaier and Schneider (2001). Fig. 10 shows the variation of the SBC criterion for different model orders, ranging from 1 to 20 (p in Equation (13)). As indicated, the SBC criterion reaches its minimum value at 9, which was thus selected as the order of the multivariate autoregressive model. This means that the value of the influent time series at time t is simulated as a function of the last 9 influent values antecedent to time t.

Fig. 11 shows a continuous 3-day DWF influent time series with the results corresponding to the most likely simulated multivariate auto-regressive model. The uncertainty band was generated through random generation of the noise term (i.e. p, A_l in Equation (13) were fixed and the noise term was generated from $\varepsilon_t \sim N(0,C)$). It should be noted that the water quality data, except for ammonia, did not exhibit the strong diurnal variation that is typically observed in other catchments (Martin and Vanrolleghem, 2014).

One of the main advantages of the proposed multivariate time series model over univariate time series models (Martin et al., 2007) or the DWF generator in the BSM influent generator (Gernaey et al., 2011) is that not only are the auto-correlation

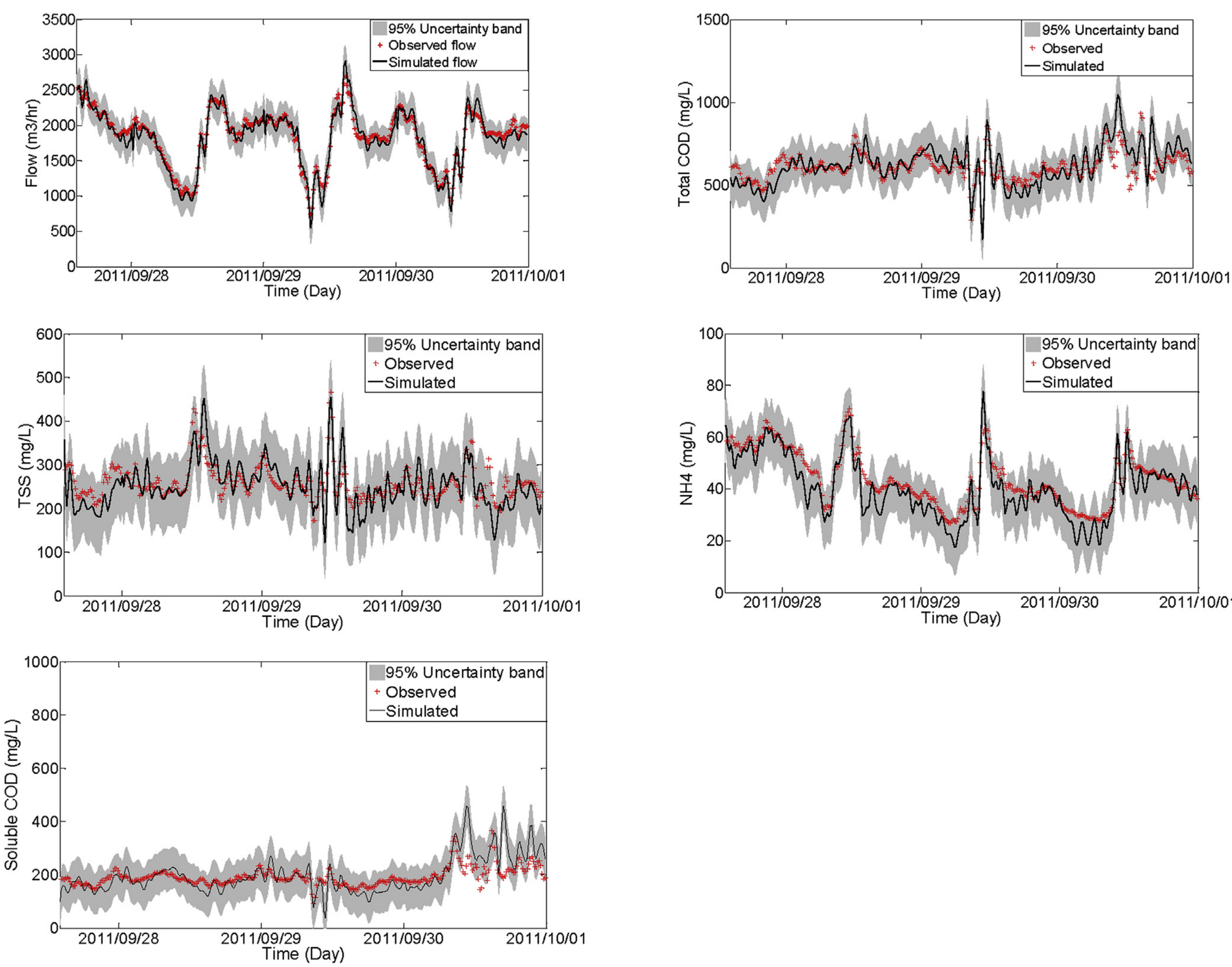

Fig. 11. Observed and most likely simulated influent time series under DWF conditions. Note: The continuous 3-day influent time series belongs to 82 days of DWF data used for estimating the parameters of the multivariate auto-regressive model.

Table 4
Correlation matrix for the generated and observed influent time series in DWF.

Generated influent time series						Observed influent time series					
	Flow	Soluble COD	Total COD	TSS	NH4		Flow	Soluble COD	Total COD	TSS	NH4
Flow	1.00					Flow	1.00				
Soluble COD	−0.11	1.00				Soluble COD	−0.12	1.00			
Total COD	−0.04	0.77	1.00			Total COD	−0.06	0.77	1.00		
TSS	0.06	0.32	0.80	1.00		TSS	0.05	0.33	0.81	1.00	
NH4	−0.43	−0.04	−0.06	−0.04	1.00	NH4	−0.46	0.00	−0.02	−0.03	1.00

Table 5
Prior distribution of parameters and the values for the maximum likelihood function.

Parameter	Unit	Lower limit	Upper limit	Values corresponding to the maximum likelihood
Runoff coeff	–	0.6	0.9	0.69
Init loss	mm	0	2	0.3
Perm loss	mm/day	0	2	0.57
K (Muskingum coeff)	second	8000	20,000	16 869
X (Muskingum coeff)	–	0.1	0.4	0.12
Sigma (for flow)	m^3/hr	0.05	2	0.22
Ka	1/day	0.001	2	0.08
m lim	Kg/ha	0.001	120	100
We	–	0.0004	0.002	0.0005
w	–	1.5	2	1.667
Sigma (for TSS)	g/m^3	20	70	38.9

structures in time respected but also the cross-correlation structures. Table 4 shows the correlation matrix for the randomly generated and observed influent time series under DWF conditions.

Depending on the type of the model that is to be used for modeling the treatment processes inside a WWTP (e.g. ASM models (Henze et al., 2000)), the wastewater constituents in Table 4 can be further converted to WWTP model state variables. However, as illustrated in Fig. 1, influent fractionation should be considered part of WWTP modeling as different WWTP models may have different state variables (Martin and Vanrolleghem, 2014).

4.4. CITYDRAIN model calibration and synthetic influent generation

As explained in the methodology section, the CITYDRAIN model with three catchment blocks representing the main sub-sewersheds in the Eindhoven sewershed was used for modeling

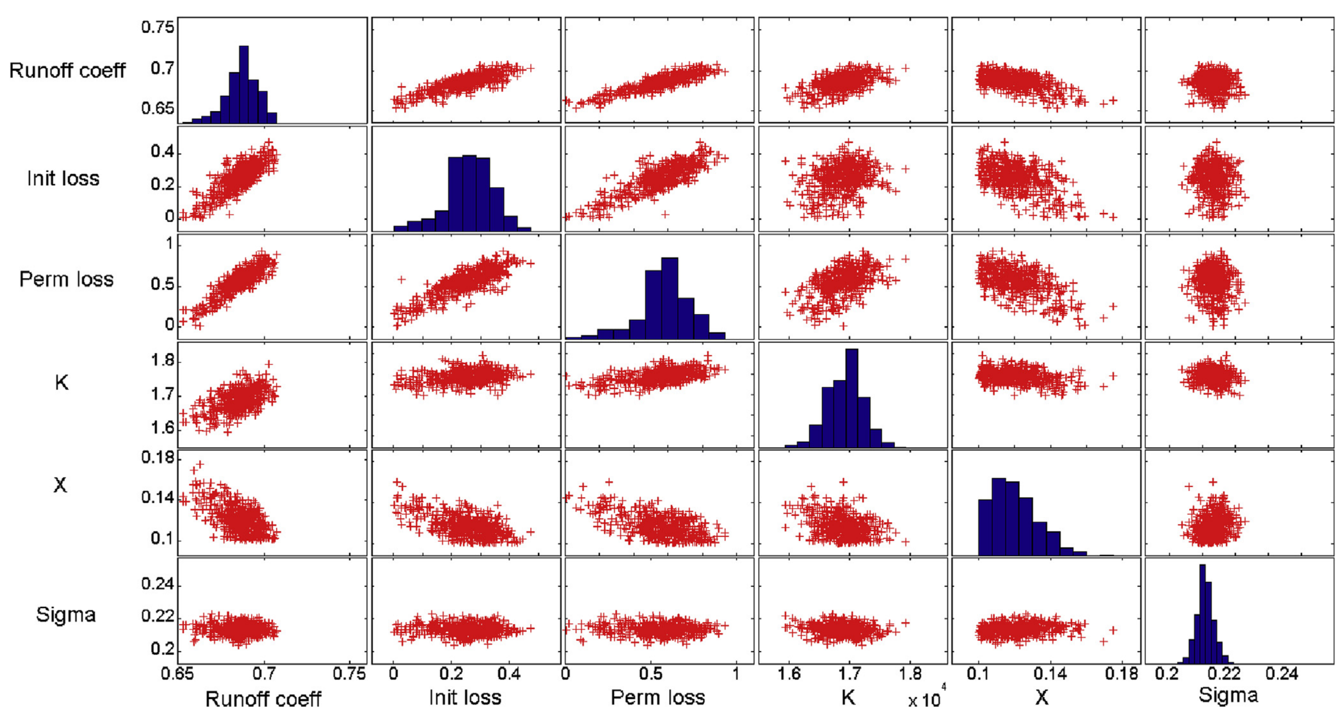

Fig. 12. Posterior distribution of parameters for flow calibration where, runoff coeff, init loss, and perm loss are respectively the runoff coefficient, initial loss (mm), permanent loss (mm/day) parmeters in the virtual basins model that is used in the CITYDRAIN model, K (sec) and X are the routing parameters used in the Muskingum method, and Sigma is the standard deviation of the residual error. The blue histograms represent the marginal posterior distributions of the individual parameters and the red scatter plots represent the relationships corresponding to various combinations of parameters (Equation (14)).

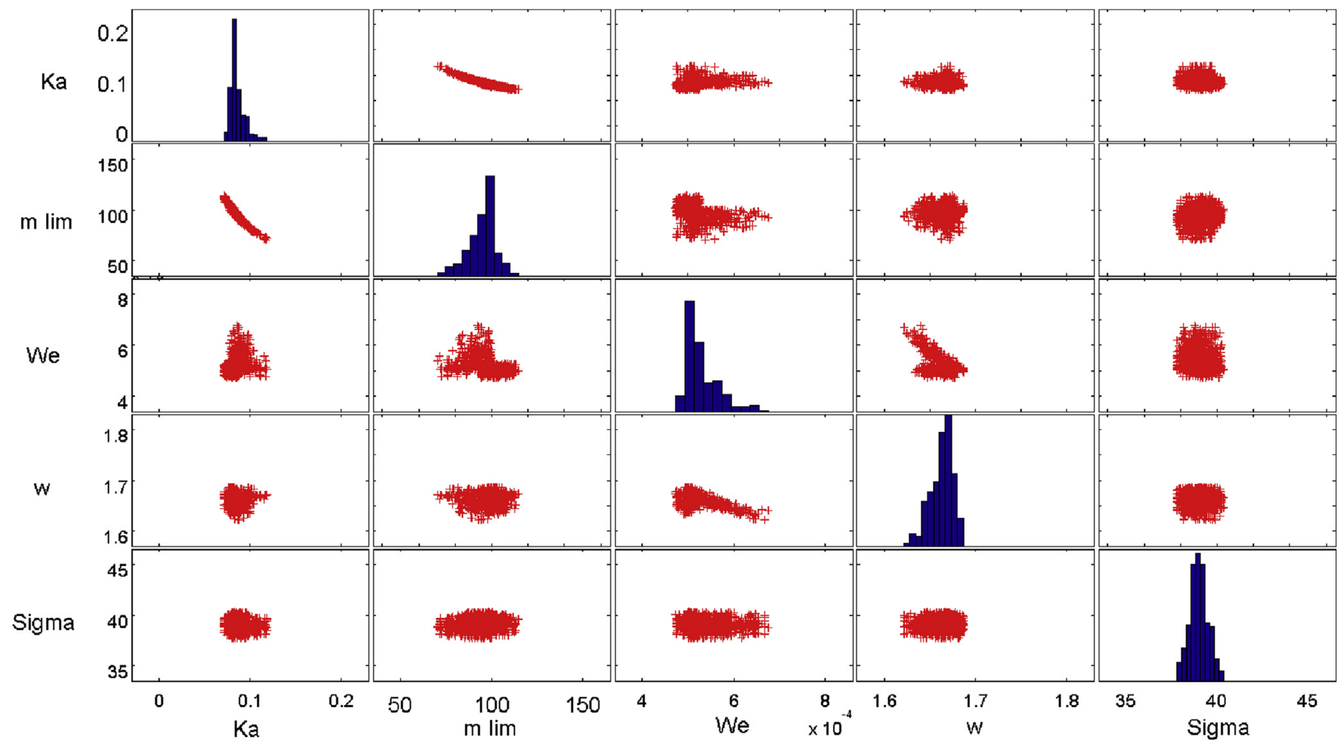

Fig. 13. Posterior distribution of parameters used for TSS calibration where Ka is the accumulation coefficient (1/day), m_lim is the maximum accumulated mass (kg/ha), We, and w are the calibration parameters (Equation (16)).

the dynamics of the influent time series during WWF conditions. The decision on the number of catchment blocks was made based on the information obtained from previous studies as well as the measured influent data that were used for model calibration (Schilperoort, 2011).

Uniform distributions representing the initial knowledge on parameters were selected as prior distributions (Table 5) and their corresponding posterior distributions were estimated by sampling from Equation (18) (i.e. 12,000 samples which required around 10 h of computation) using the DREAM sampler (as indicated in Table 5, 11 parameters were estimated). Fig. 12 and Fig. 13 show the posterior distributions of the CITYDRAIN model after calibrating the model for flow and TSS time series in WWF conditions (three days of dry weather simulation were used as the warm-up period to set the initial conditions of the system).

As indicated in Fig. 12 and Fig. 13, some correlation among the parameters of the CITYDRAIN model exists. For example, the parameters that affect the generation of effective rainfall (i.e. runoff coefficient, initial loss, and permanent loss) are correlated, meaning that different combinations of these parameters could result in approximately the same amount of effective rainfall given the same inputs and values for other parameters. However, given the narrow ranges of values obtained for the marginal posterior distribution of the parameters that affect the amount and dynamics of flow (i.e. Runoff coeff, Init loss, Perm loss, K, and X in Fig. 12), the uncertainty band for flow relating to the *total model uncertainty* is mainly affected by the standard deviation of the residual error (i.e. Sigma in Fig. 12) and not by the uncertainty of the CITYDRAIN model parameters.

The parameters that affect the accumulation of a pollutant (i.e. m_lim, and Ka) and those that affect the wash-off of TSS are also correlated (Fig. 13). Given the different correlation structures that exist among some parameters it is very important to sample from the joint posterior distribution of parameters to properly propagate the effect of parameter uncertainties to the outputs.

To consider the effect of *total model uncertainty* (i.e. including model parameter uncertainty and the standard deviation of noise (i.e. Sigma for flow and TSS in Fig. 12 and Fig. 13) on the outputs of

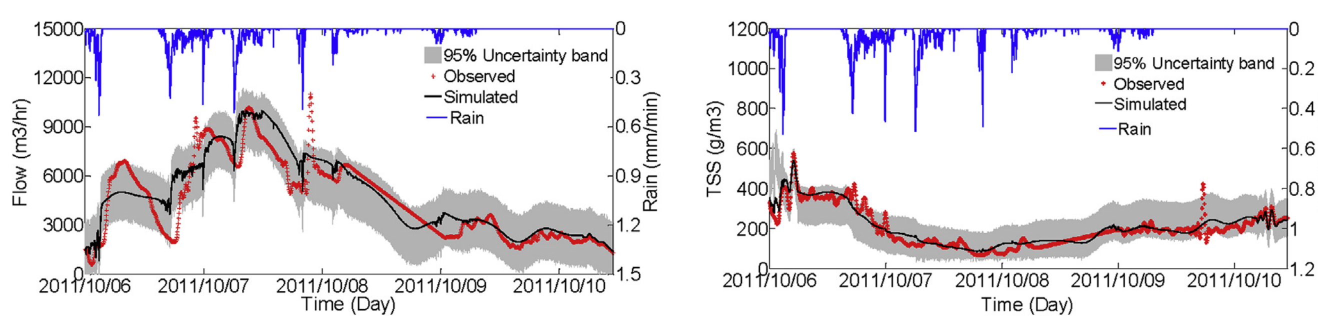

Fig. 14. Uncertainty bands for flow (left) and TSS concentration (right) in a 4-day wet weather period (Rain series in blue and maximum likelihood simulation in black). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

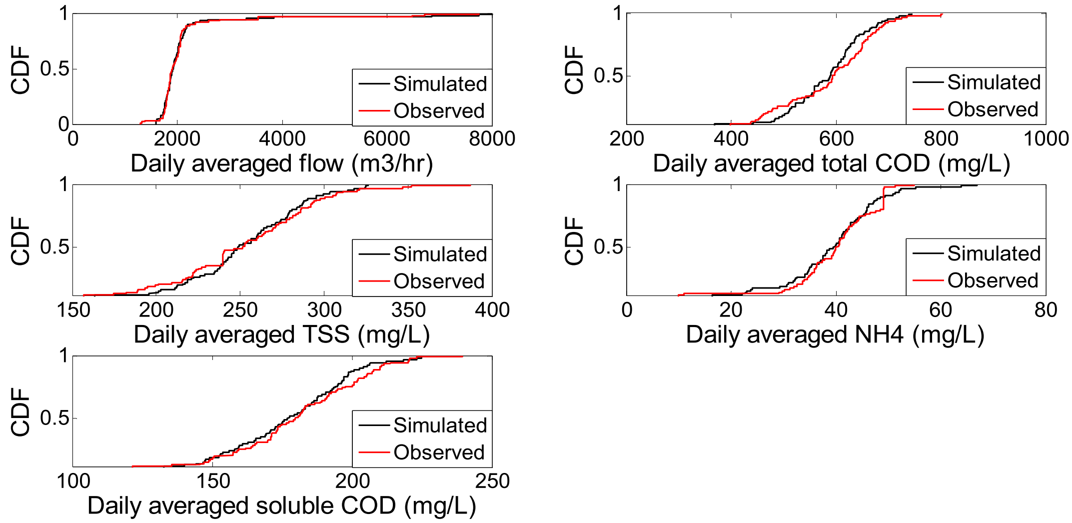

Fig. 15. CDFs of daily-averaged influent flow and concentration of influent pollutants.

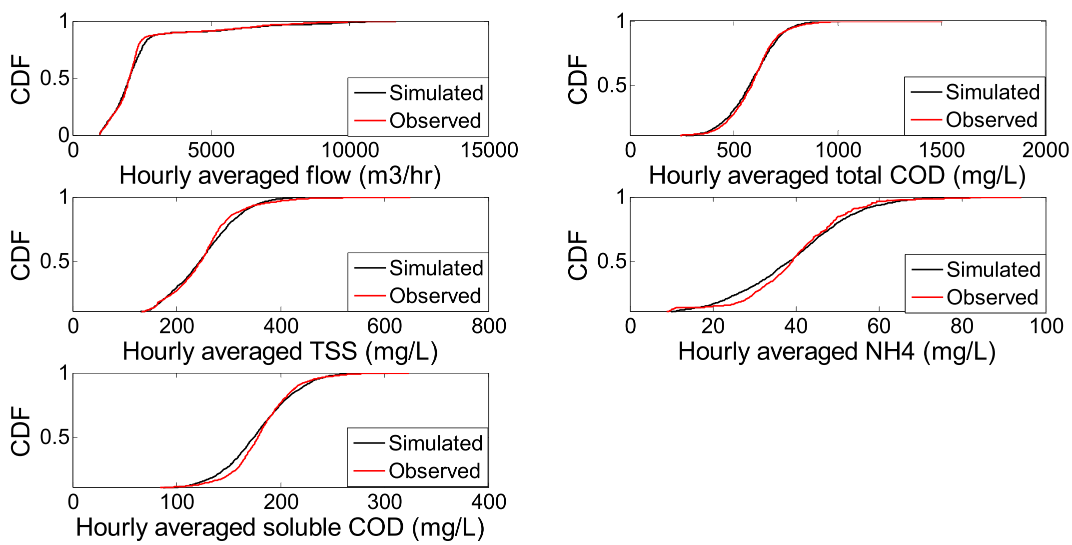

Fig. 16. CDFs of hourly-averaged influent flow and concentration of influent pollutants.

CITYDRAIN model, a Monte Carlo simulation was performed by sampling from the joint *posterior distribution* of parameters and running the model for 1000 times for the rainfall time series between 6th and 10th of October 2011 (i.e. the rainfall time series concurrent with the observed flow and TSS concentration time series used for the CITYDRAIN calibration). Fig. 14 illustrates the 95% uncertainty band for flow and TSS. This uncertainty band was constructed by selecting the 2.5 and 97.5 percentiles of the cumulative distribution of flow and TSS as the lower and upper limits of uncertainty of simulation with the rainfall time series shown in the figure. The figure also presents the observed and the best simulated time series. The latter corresponds to the simulations obtained with the set of parameters that has the highest *likelihood function* value.

To further analyze the statistical properties of the simulated influent time series during both DWF and WWF conditions, the cumulative distribution function (CDF) of the simulated and observed influent flow and pollutant concentrations were compared in Fig. 15 and Fig. 16. The simulated and observed influent time series with 15-min temporal resolution were averaged (using the flow and concentration time series) to construct the corresponding daily and hourly influent series. Fig. 15 and Fig. 16 show that the influent generator has excellent performance when it comes to predicting the daily and hourly influent flow and pollutant concentration values (the comparison between the simulated and observed load values (results not shown) also indicated an excellent performance of the influent generator).

It can be concluded from Fig. 15 and Fig. 16 that the statistical properties of the simulated time series were very similar to the properties of the observed series when the CITYDRAIN model is fed with the observed rainfall time series.

4.5. Synthetic generation of influent time series

As explained in the methodology section, synthetic generation of a one year influent time series with 15-min temporal resolution is thus possible by sampling from the joint posterior distribution of the CITYDRAIN model parameters (one vector of CITYDRAIN parameters for each year) and running the model with the synthetically-generated rainfall and DWF influent time series (both with 15-min temporal resolution). The latter two series are generated using the proposed rainfall and DWF generators respectively.

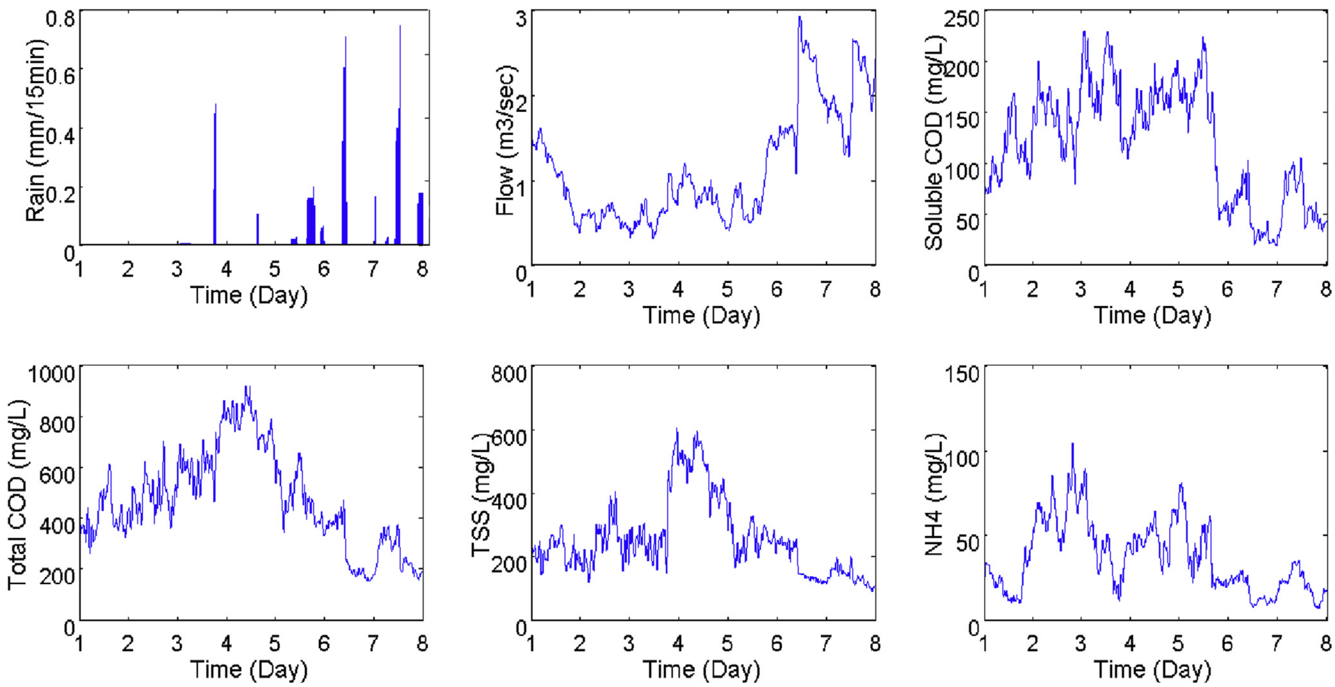

Fig. 17. A 7-day realization of rainfall and influent time series (flow and composition).

Fig. 17 shows a 7-day snapshot of a generated one year influent time series. During the hours of the first day the time series of flow has a descending trend as the runoff produced by rainfall event just before the first day (not depicted in Fig. 17) exits the sewer system and the flow time series reaches its DWF conditions with a typical periodic pattern (the second day in Fig. 17). During the last hours of the third day another rainfall event occurs and the flow time series increases while the time series of soluble COD and ammonia drop due to dilution of wastewater by runoff. However, during the same period of time there is a sudden increase in the total COD and TSS concentrations due to the wash-off of particulate material. After the wash-off of the particulates during the last hours of the fourth day, the concentrations of total COD and TSS drop due to the dilution of the wastewater with runoff.

5. Conclusion

In this paper a combination of statistical and conceptual modeling tools was proposed for synthetic generation of dynamic influent time series of flows and pollutant concentrations with 15-min temporal resolution. The rainfall generator is capable of considering observed annual and seasonal rainfall regimes and keeping the consistency of the generated rainfall time series across different temporal resolutions. For dry weather conditions, comparison between observed and simulated influent time series for the Eindhoven case study confirmed the capability of the proposed multivariate auto-regressive model in generating realistic influent time series for flow and pollutants composition. Moreover, long-term generation of influent time series under dry and wet weather conditions could be achieved by running a constructed CITYDRAIN model of the sewershed using the generated stochastic inputs (i.e. rainfall and influent time series in DWF condition). Further, uncertainty could be captured by sampling different vectors of the model parameters from the posterior distribution obtained after Bayesian parameter estimation on the basis of the case study data.

Overall, the proposed influent generator provides a clear and coherent tool to incorporate general and easy-to-obtain information on the physical characteristics of a sewershed as well as climate conditions of the region into the synthetic generation of the influent flow and composition of a treatment plant. If there are no observed data for calibrating the parameters of the proposed influent generator, a range of values should be assigned to the uncertain model parameters based on expert elicitation or transfer of information from similar sewersheds. The flexibility of the presented influent generator allows users to define different scenarios reflecting the projected change in climate and the characteristics of the sewershed (e.g. population growth, change in pervious area) and evaluate their effect on the generated influent time series and the treatment plant to be designed.

Acknowledgment

The authors would like to thank Stefan Weijers and Petra van Daal-Rombouts from Waterschap De Dommel (Eindhoven, The Netherlands) and Youri Amerlinck from BIOMATH-Ghent University (Belgium) for providing the influent and rainfall data for this study. This research work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Primodal Inc. as well as the financial support of Hampton Roads Sanitation District, Virginia. Peter Vanrolleghem holds the Canada Research Chair in Water Quality Modeling.

Appendix

Table A1
Estimated parameters of air temperature model (i.e. Equation (6))

$$\begin{pmatrix} T_{\max} \\ T_{\min} \end{pmatrix}_t = A \times \begin{pmatrix} T_{\max} \\ T_{\min} \end{pmatrix}_{t-1} + B \times N(0,1)_t$$

A		B	
0.79	0.05	0.56	−0.06
0.34	0.52	0.28	0.52

Table A2
Estimated parameters of influent model in DWF conditions (i.e. Equation (13))

$$\begin{pmatrix} Flow \\ Solube\_CO \\ Total\_COD \\ TS \\ NH_4 \end{pmatrix}_t = \sum_{l=1}^{l=9} A_l \times \begin{pmatrix} Flow \\ Solube\_COD \\ Total\_COD \\ TSS \\ NH_4 \end{pmatrix}_{t-1} + \varepsilon_t$$

A1					***A2***				
1.62	0.05	−0.09	0.03	−0.18	−0.90	−0.13	0.18	−0.09	0.23
0.00	2.06	−0.90	0.67	0.00	−0.02	−1.88	1.80	−1.33	−0.01
−0.02	0.26	1.10	0.36	−0.03	0.03	−0.51	0.05	−0.73	0.04
−0.02	0.38	−0.57	2.00	−0.01	0.04	−0.78	1.25	−1.84	0.02
−0.08	0.01	−0.04	0.04	1.58	0.14	0.00	0.06	−0.05	−0.81
A3					***A4***				
0.07	0.12	−0.16	0.10	−0.06	0.27	−0.01	0.03	−0.03	−0.04
0.02	0.81	−1.34	0.99	0.04	−0.01	0.17	0.23	−0.12	−0.05
0.00	0.37	−0.59	0.51	−0.01	−0.02	−0.05	0.40	−0.04	−0.04
0.00	0.59	−0.94	0.79	−0.01	−0.02	0.01	0.01	0.29	−0.02
−0.05	0.00	−0.06	0.04	0.13	−0.05	−0.02	0.06	−0.05	0.14
A5					***A6***				
−0.10	−0.05	0.06	−0.04	0.03	−0.13	0.05	−0.10	0.08	0.05
0.00	−0.39	0.66	−0.57	0.03	0.00	0.16	−0.65	0.55	0.02
0.00	−0.17	0.25	−0.28	0.04	0.01	0.13	−0.43	0.24	0.04
0.00	−0.40	0.58	−0.47	0.00	0.02	0.29	−0.44	0.15	0.05
0.04	0.00	−0.04	0.04	−0.01	0.01	0.03	0.00	−0.02	−0.12
A7					***A8***				
0.12	−0.01	0.04	−0.02	−0.04	−0.13	−0.11	0.13	−0.08	0.01
0.00	0.02	0.24	−0.21	−0.03	0.01	−0.21	0.12	−0.11	0.00
−0.01	−0.02	0.18	−0.06	−0.04	0.00	−0.17	0.08	−0.18	−0.01
−0.02	−0.05	0.06	0.08	−0.04	−0.01	0.01	−0.02	−0.13	0.00
−0.02	−0.02	0.02	0.00	0.06	0.01	0.14	−0.17	0.14	−0.06
A9									
0.04	0.00	0.01	−0.03	−0.03					
−0.01	0.11	−0.03	0.03	−0.02					
−0.01	0.02	0.08	0.04	−0.01					
−0.01	0.00	0.01	0.12	0.01					
0.00	−0.11	0.09	−0.07	0.05					

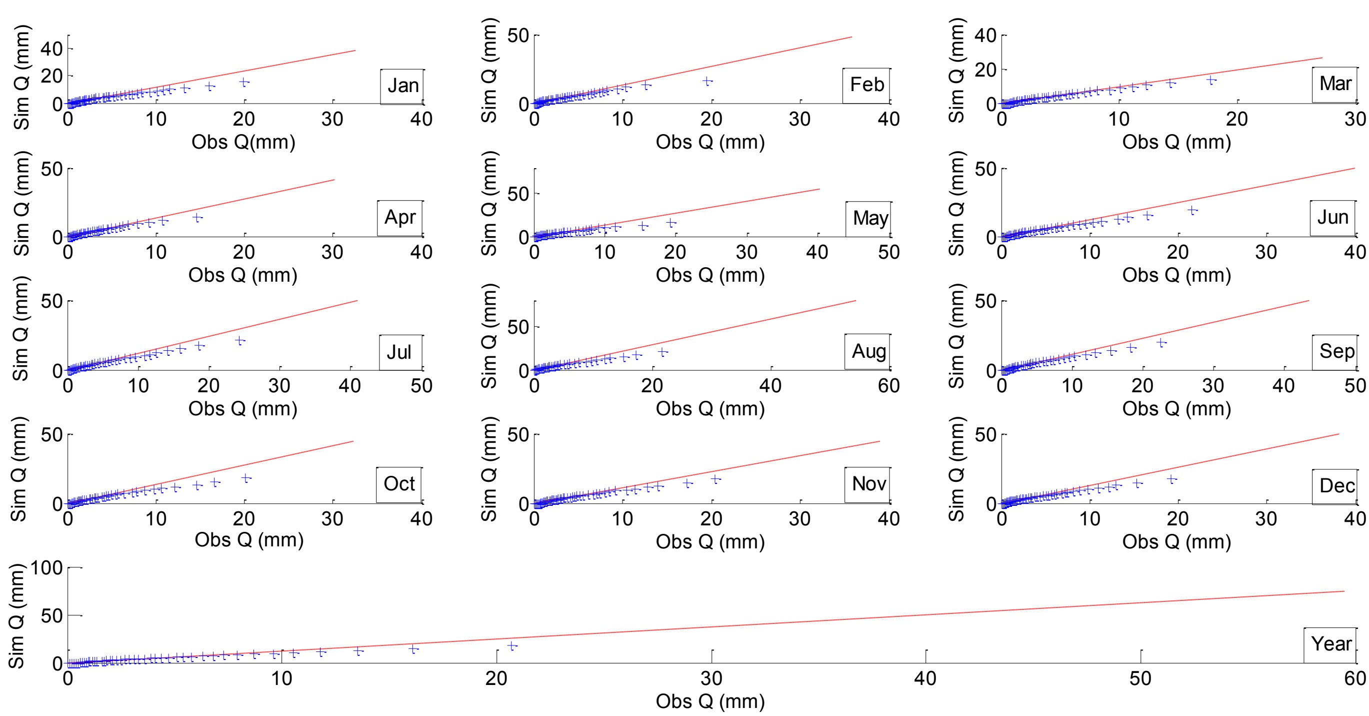

Fig. A1. q–q plot for observed and simulated rainfall quantiles in different month and the entire year. Note: the q–q plots were generated by plotting 1 to 99 quantiles of observed rainfall (Obs Q) data against their corresponding quantiles in simulated rainfall series (Sim Q). The red dotted lines represent the locations where the corresponding quantiles are equal.

References

Achleitner, S., Möderl, M., Rauch, W., 2007. CITY DRAIN © – an open source approach for simulation of integrated urban drainage systems. Environ. Model. Softw. 22 (8), 1184–1195.

Antoniou, P., Hamilton, J., Koopman, B., Jain, R., Holloway, B., Lyberatos, G., Svoronos, S.A., 1990. Effect of temperature and pH on the effective maximum specific growth rate of nitrifying bacteria. Water Res. 24 (1), 97–101.

Bates, B.C., Campbell, E.P., 2001. A Markov Chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resour. Res. 37 (4), 937–947.

Bixio, D., Parmentier, G., Rousseau, D., Verdonck, F., Meirlaen, J., Vanrolleghem, P.A., Thoeye, C., 2002. A quantitative risk analysis tool for design/simulation of wastewater treatment plants. Water Sci. Technol. 46 (4), 301–307.

Bechmann, H., Nielsen, M.K., Madsen, H., Kjølstad Poulsen, N., 1999. Grey-box modelling of pollutant loads from a sewer system. Urban Water 1 (1), 71–78.

Belia, E., Amerlinck, Y., Benedetti, L., Johnson, B., Sin, G., Vanrolleghem, P.A., Gernaey, K.V., Gillot, S., Neumann, M.B., Rieger, L., Shaw, A., Villez, K., 2009. Wastewater treatment modelling: dealing with uncertainties. Water Sci. Technol. 60 (8), 1929–1941.

Benedetti, L., Bixio, D., Vanrolleghem, P.A., 2006. Benchmarking of WWTP design by assessing costs, effluent quality and process variability. Water Sci. Technol. 54 (10), 95–102.

Capodaglio, A., Zheng, S., Novotny, V., Feng, X., 1990. Stochastic system identification of sewer-flow models. J. Environ. Eng. 116 (2), 284–298.

Chen, J., Brissette, F.P., Leconte, R., 2010. A daily stochastic weather generator for preserving low-frequency of climate variability. J. Hydrol. 388 (3–4), 480–490.

Ciggin, A.S., Rossetti, S., Majone, M., Orhon, D., 2012. Effect of feeding regime and the sludge age on the fate of acetate and the microbial composition in sequencing batch reactor. J. Environ. Sci. Health Part A 47 (2), 192–203.

Devisscher, M., Ciacci, G., Benedetti, L., Bixio, D., Thoeye, C., Gueldre, G.D., Marsili-Libelli, S., Vanrolleghem, P.A., 2006. Estimating costs and benefits of advanced control for wastewater treatment plants the MAGIC methodology. Water Sci. Technol. 53 (4–5), 215–223.

Dotto, C.B.S., Mannina, G., Kleidorfer, M., Vezzaro, L., Henrichs, M., McCarthy, D.T., Freni, G., Rauch, W., Deletic, A., 2012. Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling. Water Res. 46 (8), 2545–2558.

Flores-Alsina, X., Saagi, R., Lindblom, E., Thirsing, C., Thornberg, D., Gernaey, K.V., Jeppsson, U., 2014. Calibration and validation of a phenomenological influent pollutant disturbance scenario generator using full-scale data. Water Res. 51(), 172–185.

Freni, G., Mannina, G., 2010. Bayesian approach for uncertainty quantification in water quality modelling: the influence of prior distribution. J. Hydrol. 392 (1), 31–39.

Freni, G., Mannina, G., Viviani, G., 2009. Uncertainty assessment of an integrated urban drainage model. J. Hydrol. 373 (3), 392–404.

Gernaey, K.V., Flores-Alsina, X., Rosen, C., Benedetti, L., Jeppsson, U., 2011. Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. Environ. Model. Softw. 26 (11), 1255–1267.

Gernaey, K.V., Jeppsson, U., Vanrolleghem, P.A., Copp, J.B., 2014. Benchmarking of Control Strategies for Wastewater Treatment Plants. Scientific and Technical Report No 23. IWA Publishing, London, UK.

Gillot, S., Vanrolleghem, P.A., 2003. Equilibrium temperature in aerated basins-comparison of two prediction models. Water Res. 37(), 3742–3748.

Guerrero, J., Guisasola, A., Vilanova, R., Baeza, J.A., 2011. Improving the performance of a WWTP control system by model-based setpoint optimisation. Environ. Model. Softw. 26 (4), 492–497.

Hao, X., van Loosdrecht, M.C.M.V., Meijer, S.C.F., Qian, Y., 2001. Model-based evaluation of two BNR processes—UCT and A2N. Water Res. 35 (12), 2851–2860.

Henze, M., Gujer, W., Mino, T., van Loosdrecht, M., 2000. Activated Sludge Models ASM1, ASM2, ASM2D and ASM3. Scientific and Technical Report No. 9. IWA Publishing, London, UK.

Hernebring, C., Jönsson, L., Thorn, U., Møller, A., 2002. Dynamic online sewer modelling in Helsingborg. Water Sci. Technol. 45 (4–5), 429–436.

Hyland, K.C., Dickenson, E.R., Drewes, J.E., Higgins, C.P., 2012. Sorption of ionized and neutral emerging trace organic compounds onto activated sludge from different wastewater treatment configurations. Water Res. 46 (6), 1958–1968.

Kanso, A., Tassin, B., Chebbo, G., 2005. A benchmark methodology for managing uncertainties in urban runoff quality models. Water Sci. Technol. 51 (02), 163–170.

Koutsoyiannis, D., Onof, C., 2001. Rainfall disaggregation using adjusting procedures on a Poisson cluster model. J. Hydrol. 246 (1–4), 109–122.

Langergraber, G., Alex, J., Weissenbacher, N., Woerner, D., Ahnert, M., Frehmann, T., Halft, N., Hobus, I., Plattes, M., Spering, V., 2008. Generation of diurnal variation for influent data for dynamic simulation. Water Sci. Technol. 57 (9), 1483.

Langeveld, J.G., Schilperoort, R.P.S., Rombouts, P.M.M., Benedetti, L., Amerlinck, Y., de Jonge, J., Flameling, T., Nopens, I., Weijers, S., 2014. A new empirical sewer water quality model for the prediction of WWTP influent quality. In: Proceedings of the 13th IWA/IAHR International Conference on Urban Drainage. Sarawak, Malaysia, 7th-12th September 2014.

Marshall, L., Nott, D., Sharma, A., 2004. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. Water Resour. Res. 40 (2), W02501.

Martin, C., Eguinoa, I., McIntyre, N.R., García-Sanz, M., Ayesa, E., 2007. ARMA models for uncertainty assessment of time series data: application to Galindo-Bilbao WWTP. In: Proceedings of the Seventh International IWA Symposium on Systems Analysis and Integrated Assessment in Water Management (WATERMATEX 2007). Washington DC, USA, 7th–9th May 2007.

Martin, C., Ayesa, E., 2010. An integrated Monte Carlo Methodology for the calibration of water quality models. Ecol. Model. 221 (22), 2656–2667.

Martin, C., Neumann, M.B., Altimir, J., Vanrolleghem, P.A., 2012. A tool for optimum design of WWTPs under uncertainty: estimating the probability of compliance. In: Proceedings International Congress on Environmental Modelling and Software (IEMSs2012). Leipzig, Germany, 1st -5th July 2012.

Martin, C., Vanrolleghem, P.A., 2014. Analysing, completing, and generating influent data for WWTP modelling: a critical review. Environ. Model. Softw. 60(), 188–201. http://dx.doi.org/10.1016/j.envsoft.2014.05.026.

Matalas, N.C., 1967. Mathematical assessment of synthetic hydrology. Water Resour. Res. 3 (4), 937–945.

Motiee, H., Chocat, B., Blanpain, O., 1997. A storage model for the simulation of the hydraulic behaviour of drainage networks. Water Sci. Technol. 36 (8–9), 57–63.

Neumaier, A., Schneider, T., 2001. Estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Trans. Math. Softw. TOMS 27 (1), 27–57.

Ormsbee, L., 1989. Rainfall disaggregation model for continuous hydrologic modeling. J. Hydraul. Eng. 115 (4), 507–525.

Pattison, A., 1965. Synthesis of hourly rainfall data. Water Resour. Res. 1 (4), 489–498.

Richardson, C.W., 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour. Res. 17 (1), 182–190.

Roberson, J.A., Cassidy, J., Chaudhry, M.H., 1995. Hydraulic Engineering. John Wiley Sons, Inc, New York.

Rodríguez, J.P., McIntyre, N., Díaz-Granados, M., Achleitner, S., Hochedlinger, M., Maksimović, Č., 2013. Generating time-series of dry weather loads to sewers. Environ. Model. Softw. 43(), 133–143.

Rodriguez-Iturbe, I., Cox, D.R., Isham, V., 1987. Some models for rainfall based on stochastic point processes. Proc. R. Soc. Lond. Ser. A 410(), 269–288.

Rousseau, D., Verdonck, F., Moerman, O., Carrette, R., Thoeye, C., Meirlaen, J., Vanrolleghem, P.A., 2001. Development of a risk assessment based technique for design/retrofitting of WWTPs. Water Sci. Technol. 43 (7), 287–294.

Salem, S., Berends, D., Heijnen, J., van Loosdrecht, M.C.M., 2002. Model-based evaluation of a new upgrading concept for N-removal. Water Sci. Technol. 45 (6), 169–176.

Schilperoort, R., 2011. Monitoring as a Tool for the Assessment of Wastewater Quality Dynamics. Water Management Academic Press, Delft, the Netherlands, p. 320.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.

Talebizadeh, M., Belia, E., Vanrolleghem, P.A., 2014. Probability-based design of wastewater treatment plants. In: Proceedings International Congress on Environmental Modelling and Software (IEMSs2014). San Diego, California, USA, 15th- 19th June, 2014.

Temprano, J., Arango, Ó., Cagiao, J., Suárez, J., Tejero, I., 2007. Stormwater quality calibration by SWMM: a case study in Northern Spain. Water SA. 32 (1), 55–63.

Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour. Res. 44 (12), W00B9.