Evaluation of the information content of long-term wastewater characteristics data in relation to activated sludge model parameters

Jamal Alikhani, Imre Takacs, Ahmed Al-Omari, Sudhir Murthy and Arash Massoudieh

ABSTRACT

A parameter estimation framework was used to evaluate the ability of observed data from a full-scale nitrification–denitrification bioreactor to reduce the uncertainty associated with the bio-kinetic and stoichiometric parameters of an activated sludge model (ASM). Samples collected over a period of 150 days from the effluent as well as from the reactor tanks were used. A hybrid genetic algorithm and Bayesian inference were used to perform deterministic and parameter estimations, respectively. The main goal was to assess the ability of the data to obtain reliable parameter estimates for a modified version of the ASM. The modified ASM model includes methylotrophic processes which play the main role in methanol-fed denitrification. Sensitivity analysis was also used to explain the ability of the data to provide information about each of the parameters. The results showed that the uncertainty in the estimates of the most sensitive parameters (including growth rate, decay rate, and yield coefficients) decreased with respect to the prior information.

INTRODUCTION

The activated sludge process is one of the most widely applied systems for the removal of natural organic matter and nutrients in municipal wastewater treatment plants (WWTPs) (Gernaey et al. 2004). The International Water Association’s activated sludge model number 1 (ASM1) was one of the first modeling concepts to explicitly account for the interactions of multiple constituents and functional groups of microorganisms (Henze et al. 1987). International Water Association task groups later developed more complex versions of the ASM1: ASM2, ASM2d, and ASM3 (Henze et al. 2000). Other models that have been developed include the Barker and Dold model (Barker & Dold 1997), the TUDP model (Meijer et al. 2001), and the EAWAG Bio-P module for ASM3 (Rieger et al. 2001). A comprehensive review of the different ASM models can be found in Gernaey et al. (2004).

The ASMs provide a systematic way to represent the interaction between state variables, process rates, and stoichiometric coefficients in bioreactions, making them flexible tools for incorporating various processes in both full-scale and pilot systems (Comas et al. 2008; Sun et al. 2009; Fang et al. 2010; Fall et al. 2011; Busch et al. 2015). Modified versions of the ASMs have been developed to address different levels of complexity in representing processes, emerging contaminants, particulate matter forms, and functional groups of microorganisms (Ekama & Wentzel 2004; Stare et al. 2006; Bolong et al. 2009; Sun et al. 2009; Hao et al. 2011; Yang et al. 2013).

One of the challenges faced in using any ASM is the difficulty of calibrating a large number of bio-kinetic and stoichiometric parameters (Smets et al. 2003). Modeling can be performed without calibration and just by using parameter values obtained from the literature based on independent batches or pilot investigations (Henze et al. 2000; Cox 2004; Hauduc et al. 2011; Rahman et al. 2016a, 2016b). However, the complexities of biological reaction systems, variations in raw wastewater characteristics, design configurations, and different experimental or operating
conditions can make it difficult to find suitable parameter values from the literature for a specific plant. Furthermore, wide ranges of values have been reported for ASM parameters in the literature, which makes choosing a single parameter set representing the condition in a specific reactor difficult. Cox (2004) reviewed a number of articles and created a database containing different ASM parameter values. As an example, 31 different values have been reported for the maximum heterotrophic growth rate in 22 references, ranging between 0.51 and 11 day\(^{-1}\).

In many cases, an estimation of the ASM parameters through calibration using measured data provides more realistic values for existing biological treatment systems or design of a new system compared to determining the parameter values from the literature (Sharifi et al. 2014). Manual and automatic calibration methods have been widely used to estimate ASM parameters using measured data (Meijer et al. 2001; Fall et al. 2011). These methods are typically based on minimizing an objective function defined as the misfit between model predictions and observed data (Vanrolleghem et al. 1999; Machado et al. 2009; Kim et al. 2010; Keskitalo & Leiviskä 2012). Keskitalo & Leiviskä (2012) applied an evolutionary optimizer to minimize the sum of squared errors between modeled and measured data obtained from a municipal WWTP and a pulp mill WWTP. Kim et al. (2010) used the response surface method as a surrogate for the mechanistic ASM to estimate model parameters in an effort to reduce computational costs. Although deterministic (maximum likelihood) parameter estimation can lead to more realistic values for a particular plant than the use of default values from the literature, there remain uncertainties associated with the estimated parameters due to measurement error, model structural error, the errors associated with the representativeness of sampling, influent’s flowrate and composition, and environmental factors such as temperature (Sharifi et al. 2014). The effect of these uncertainties on the deterministically estimated parameters needs to be quantified to assess the reliability of ASM models (Neumann & Gujer 2008; Hauduc et al. 2010; Mannina et al. 2012; Sharifi et al. 2014). Uncertainty analysis can help in determining the appropriate safety factors to control the capital investment and effluent quality (Bixio et al. 2002; Flores-Alsina et al. 2008; Busch et al. 2013). A common method to conduct uncertainty analysis is the Monte Carlo simulation – by generating random parameter sets within the preconceived range for each parameter – aiming to find the model effluent’s confidence intervals. Bixio et al. (2002) applied Monte Carlo simulations to avoid large safety factors in the design stage of a conventional WWTP, which resulted in a 21% reduction in sizing and 43% reduction in capital investment equivalent to 1.2 million euros. Flores-Alsina et al. (2008) evaluated different control strategies with input uncertainty using Monte Carlo simulations. Sin et al. (2011) and Sin et al. (2009) performed Monte Carlo simulations by considering uncertainties associated with ASM1’s parameters, partitioning coefficients (e.g., chemical oxygen demand (COD) content of particulate constituents), and hydraulic parameters to obtain the confidence intervals for model predictions. Alikhani et al. (2015) applied autoregressive moving-average model to generate random realizations of influent flowrates and composition into a biological nutrient removal system aiming at evaluating the influent uncertainty.

Different approaches have been used in different areas of water resources to perform backward uncertainty propagation. These approaches can be classified into those based on multiple linear regression (Hill & Tiedeman 2007; Foglia et al. 2009), global methods such as the generalized likelihood uncertainty estimator (Beven & Freer 2001; Mannina et al. 2012; Sathyamoorthy et al. 2014), and Bayesian inference (Dotto et al. 2011; Albrecht 2013; Sharifi et al. 2014; Alikhani et al. 2016a). Albrecht (2013) compared four different parameter estimation techniques for a collection of reaction models with artificially added noise to experimental data and concluded that for highly nonlinear models, the Markov Chain Monte Carlo (MCMC) algorithm based on Bayesian inference has better accuracy than local sensitivity-based approaches. However, the computational cost of MCMC simulations is significantly higher than local sensitivity-based methods, as they require a relatively larger number of model runs. For a discussion of the number of runs used in our work please see ‘Parameter estimation using MCMC’ section. The Bayesian approach provides a joint probability distribution representing the degree of confidence regarding parameters after the application of observed data (referred to as the posterior distribution). The posterior range is the parameter’s degree of credibility dictated by the observed measured points through the Bayesian inference (Kaplan 1997). In the Bayesian approach, prior knowledge about parameter values, based on the literature or independent laboratory experiments, can be incorporated into estimations via prior distributions (Sharifi et al. 2014). The magnitude of reduction in the spread of the posterior distribution of parameters relative to their prior distributions can be used as an indicator for the amount of information gained regarding each parameter. The posterior joint probability distribution of
parameters obtained from the Bayesian inference can be applied in a Monte Carlo stochastic simulation technique for stochastic simulation of the biological treatment systems (Sin et al. 2011, 2009; Mannina et al. 2012).

Only a small number of studies have focused on the parameter estimation of the nitrification and denitrification (NDN) processes using real data from full-scale bioreactors (Andreottola et al. 1997; Koch et al. 2001; Choubert et al. 2009; Kaelin et al. 2009; Mannina et al. 2012). The main contribution of the presented study is to systematically quantify the information content of long-term field data from a methanol-feed NDN system in the context of the estimation of the modified ASM’s bio-kinetics and stoichiometric parameters. The objectives of this study can be summarized as follows: (1) finding the level of confidence regarding parameter values and model prediction by applying Bayesian inference over the long-term wastewater characteristic data, (2) explaining the ability of the observed data to inform about each parameter through sensitivity analysis and evaluating parameter correlation, and (3) evaluating the effect of sampling intervals on the information content of the observed data with respect to model parameters.

MATERIALS AND METHODS

Modeling biological treatment processes

The plant model consists of a series of continuously stirred tank reactors (CSTRs) connected to a single clarifier (Figure 1). Transient mass balance equations for all the state variables, consisting of the concentration of each of the model constituents in each CSTR, can be expressed as a system of ordinary differential equations (ODEs):

\[
\frac{d(V_kC_{i,k})}{dt} = \omega_{r,k}Q_kC_{i,r} + \omega_{k}Q_mC_{i,m} + \sum_{k'=k}^m H(Q_{k',k})Q_{k',k}C_{i,k'} + \left(\sum_{k'=k}^m H(-Q_{k',k})Q_{k',k}\right)C_{i,k} + V_k \sum_{i=1}^n \Phi_{i,k}r_i + F_i,kV_k(C_i^* - C_i,k) + \dot{m}_{i,k}
\]

(1)

where \(V\), \(C\), \(Q\), and \(\omega\) indicate the volume, concentration, flow, and flow’s fraction factor (determining how much of the inflow and return flow enters each stage in a step-feed system), respectively, \(H\) is the Heaviside (unit step) function, \(R\) is the reaction rate, \(\Phi\) is the stoichiometric coefficient, \(F\) is the mass transfer rate, \(C^*\) is the saturated concentration, and \(\dot{m}\) is the external mass flowrate. The indices are: \(i\) indicates a constituent, \(k\) represents a tank, \(r\) indicates a return flow, ‘in’ shows conditions in influent, and \(j\) indicates a process (or reaction). Moreover, \(m\) is the total number of tanks, and \(n\) is the total number of reactions. \(Q_{k',k}\) is the flowrate from tank \(k'\) to tank \(k\) either due to sequential stages or through feed-back or bypass (a positive \(Q_{k',k}\) value means flow into stage \(k\)). In Equation (1), the term on the left side is the rate of change in the total mass of the constituents \(i\) in tank \(k\); the first term on the right-hand side of Equation (1) is the mass inflow due to return flow; the second term is the mass inflow of the influent; the third and fourth terms represent inner-connection flow between tanks \(k\) and \(k'\); the fifth term is the production or disappearance of constituents due to reactions; the sixth term is the effect of rate-limited mass transfer (e.g., aeration); and the last term is the direct addition of the constituents (e.g., the addition of a carbon source for denitrification). To obtain solid particle
concentrations in the return flow \( (C_{i,r}) \), a dynamic clarifier model (Takács et al. 1991) or a quasi steady-state approximation (i.e., performing mass balance while ignoring the solid storage changes in the clarifier) can be applied.

**Bayesian parameter estimation**

Equation (1) can be expressed in vector form as (Sharifi et al. 2014):

\[
d\mathbf{C}/dt = f(\mathbf{C}(t), \mathbf{U}(t), \Theta)
\]

where \( \mathbf{C}(t) \) is the state variables vector containing the concentrations of all the constituents in the model. The length of the \( \mathbf{C} \) vector is equal to the number of constituents times the number of CSTR tanks. \( \mathbf{U}(t) \) is the external forcing input representing the influent flowrate \( Q_{in} \), influent concentrations \( C_{i,in} \), return activated sludge (RAS) flowrate \( Q_r \), waste activated sludge (WAS) flowrate \( Q_w \), external chemical loading rate \( m_i \), and the temperature. \( \Theta \) is the vector of the model’s parameters. Function \( f \) in Equation (2) represents the bioreactor model structure, which is represented in Equation (1).

Estimation of the parameters can be performed by comparing modeled and observed concentrations. To obtain the proper model concentration consistent with the measured concentrations (e.g., COD, volatile suspended solids (VSS), and total Kjeldahl nitrogen (TKN)), the following transformation can be used:

\[
\mathbf{Y} = \mathbf{ZC}(t)
\]

\[
\hat{\mathbf{Y}}(t) = g^{-1}(g(\mathbf{Y}(t)) + \varepsilon)
\]

where \( \mathbf{Y} \) is the vector containing the model’s observed constituents, which is considered a linear combination of the model’s individual concentrations. \( \mathbf{Z} \) is the partitioning coefficients matrix, \( \hat{\mathbf{Y}}(t) \) represents the observed data vector, \( g \) represents the error structure function (e.g., for log-normal error structure, \( g \) is the natural logarithm function, while for Gaussian error structure, \( g \) is the identity function), and \( \varepsilon \) is a random vector containing measurement, structural, and external forcing errors, which is assumed to collectively follow a multivariate normal distribution.

The Bayes’ theorem (Kaipio & Somersalo 2005) can be used to obtain the joint probability distribution of parameters given the observed data, namely the posterior distribution of the parameters as:

\[
p(\boldsymbol{\phi}|\mathbf{Y}) = p(\mathbf{Y}|\boldsymbol{\phi})p(\boldsymbol{\phi}) / p(\mathbf{Y})
\]

where \( p(\mathbf{Y}|\boldsymbol{\phi}) \) is the likelihood function that is the probability of observing the measured data \( \mathbf{Y} \) given a parameter set \( \boldsymbol{\phi} \), \( p(\boldsymbol{\phi}) \) represents the prior knowledge about the parameters and error structure that can be obtained from literature reviews, experts’ experience, or independently performed experimental results. The denominator \( p(\mathbf{Y}) \) is a normalizing factor (Lu et al. 2014a).

In Equation (5) \( \boldsymbol{\phi} \) contains all the parameters that are intended to be estimated using the observed data set \( \mathbf{Y} \). Vector of \( \boldsymbol{\phi} \) includes the bio-kinetics and stoichiometric parameters \( \theta_i \in \Theta \), the elements of the partitioning coefficient \( z_i \in \mathbf{Z} \), and the elements of the variance-covariance matrix for the random error term \( \mathbf{v} \) \( (\sigma_j \in \Gamma) \). Therefore, \( \boldsymbol{\phi} = \{\theta_1, \theta_2, \ldots, \theta_n, z_1, z_2, \ldots, z_n, \sigma_1, \sigma_2, \ldots, \sigma_n\} \), where \( n_\theta \) is the total number of unknown ASM model parameters in Equation (2), \( n_z \) is the total number of unknown partitioning coefficients in Equation (5), and \( n_\sigma \) is the total number of elements in the variance-covariance matrix in Equation (7). Applying Bayesian inference enhances our degree of confidence about each unknown parameter any time a new evidence point (observed data) applies. In fact, each measured data point contains a potential level of information that can be determined by applying the Bayesian inference.

If the errors due to the uncertainties associated with the external forcing are considered part of the observation error, then the external forcing can be considered deterministically and thus the model outputs become a function of only the bio-kinetic and stoichiometric parameters. In this case, the error function can be rewritten as \( p(\hat{\mathbf{Y}}|\Theta, \mathbf{Z}, \Gamma) \). If it is assumed that errors of consecutive observations of the same observed constituent are independent of each other, then the likelihood function of the observed vector can be computed as a function of the model parameters (Freni et al. 2009):

\[
P(\hat{\mathbf{Y}}|\boldsymbol{\phi}) = \prod_{i=1}^{n_\theta} \prod_{j=1}^{n_z} g'(\tilde{y}_{ij}) \\
\times \exp \left( -\frac{1}{2} \sum_{i=1}^{n_\theta} \sum_{j=1}^{n_z} \left[ g(\tilde{y}_{ij}) - g(\tilde{y}_{ij}) \right] \Gamma^{-1} \left[ g(\tilde{y}_{ij}) - g(\tilde{y}_{ij}) \right]^T \right)
\]

where \( \tilde{y}_{ij} \in \hat{\mathbf{Y}} \) is the observed concentration of constituent \( i \), \( n_\theta \) is the number of different observed constituents, \( j \) indicates the time and location of the measurement, \( n_z \) is the number of
total samples of observed constituent $i$. The mapping function $g$ in the likelihood function is the transformation depending on the error distribution, and $g'$ is its derivative. For example, considering $g$ being identity function implies a normally distributed and additive error structure, while assuming $g$ to be logarithm function results in a log-normally distributed and multiplicative error structure.

In theory, each element of the variance-covariance matrix $\Gamma$ should be estimated using the MCMC approach. However, this makes the total number of error standard deviations ($\sigma_{ij} \in \Gamma$) to be estimated very large and imposes a large computational burden. Therefore, observation errors of different constituents are typically assumed to be independent, and the correlations between observed concentration errors of different constituents are neglected (Walsh & Whitney 2012). This assumption is justified when the main source of the error is the measurement error. However, when the main error component is the model structural error, there is a possibility of correlation between errors of different constituents. By making this assumption, the variance-covariance matrix becomes diagonal and Equation (6) can be simplified to:

$$P(Y|\phi) = \frac{\prod_{i=1}^{n} \prod_{j=1}^{m} g'(\bar{y}_{ij})}{(2\pi \prod_{i=1}^{n} n\sigma_{i}^{2})^\sum_{i=1}^{n}/2} \exp \left( -\sum_{i=1}^{n} \sum_{j=1}^{m} \left( \frac{g(\bar{y}_{ij}) - g(y_{ij})}{2\sigma_{i}^{2}} \right)^{2} \right)$$

(7)

where $y_{ij} \in Y$ is the predicted concentration of constituent $i$ at time/location $j$, and $\sigma_{i}$ is the standard deviation of observed constituent $i$.

**MCMC sampling and numerical solution scheme**

Due to a large number of parameters and heterogeneity of the observed data, solving Equation (5) by using the analytical methods is not possible. Therefore, an MCMC approach (Gamerman & Lopes 2006; Meyn & Tweedie 2012) was used to generate random samples according to the posterior distribution. In this study, the Metropolis–Hasting (M–H) algorithm (Metropolis et al. 1953) was used to obtain a sequence of random numbers from the posterior probability distribution. In the deterministic approach, the values of parameters that maximize the likelihood function in Equation (7) were obtained by applying the hybrid genetic algorithm (GA).

To evaluate the convergence of the MCMC algorithm, a fake parameter with no influence on the model prediction is defined and its posterior distribution is monitored during the Markov Chain sampling. It is expected that the posterior and prior distribution of the fake parameter be identical after a sufficient number of MCMC sampling. The practice of throwing away the initial portion of Markov Chain samples in order to avoid the effect of potentially mis-represented samples on the construction of the posterior distribution is referred to as burn-in (Meyn & Tweedie 2012). The fake parameter was also used to estimate the number of burn-in samples.

To calculate the likelihood function of Equation (6), for each set of Markov Chain samples, the system of ODEs comprising Equation (2) should be solved. In addition to the non-linearity of these ODEs, a wide range of biochemical reaction rate scales ranging from seconds (e.g., for oxygen transfer rate) to days (e.g., for microbial growth rates) results in a stiff system of ODEs (Lindberg & Carlsson 1996). The stiffness of the system of ODEs and daily fluctuations in external forcing (influent, RAS, WAS, loadings, and temperature changes) requires small computational time-steps, which in turn can result in a long simulation runtime. This is particularly a challenge due to the fact that the MCMC algorithm requires a large number of model simulations to obtain the final solution. To overcome this problem, an adaptive time-step backward differentiation formula (BDF) algorithm was developed to solve the system of ODEs. A detailed explanation of the proposed adaptive BDF algorithm for solving the system of ODEs comprising the ASM is presented in Alikhani et al. (2016b) and Alikhani et al. (2016c).

The total suspended solids (TSS) and VSS values in the effluent of the NDN system during the time of simulation show average values of 3.8 and 2.6 mg L$^{-1}$, respectively, indicating the very close to the ideal settling performance of the clarifier. Therefore, a quasi steady-state approximation is assumed to obtain the solid particle concentrations in the return flow ($C_{r}$) in Equation (1). In addition, in the biological reaction network used for modeling the reactors, the program developed for this study allows any reaction rate expression $R_{l}$ and stoichiometric constant $\phi_{lj}$ to be specified by the user as a function of concentrations of components and/or model parameters.

**Global sensitivity analysis**

Local and global sensitivities can be used to rank the model parameters in terms of the significance of their impact on important state variables (Saltelli et al. 2008; Kalyanaraman et al. 2015). Sensitivity analysis can also be applied to explain parameters’ identifiability (Weijers et al. 1996; Brun et al. 2002; Sin et al. 2005). A large sensitivity value of model output with respect to each parameter is a necessary but
not sufficient condition for the identifiability of the parameter due to the fact that the model output can change more significantly as a result of changing that parameter (Freni et al. 2009). Another factor that can influence the identifiability of parameters is parameter correlation (Eberly & Casella 2003; Ngo et al. 2015). Several approaches for estimating local and global sensitivity have been used in the literature (Saltelli et al. 2008; Zhan et al. 2013).

In this study, an aggregate local sensitivity for each time-series output is defined as the root mean square of sensitivity over the time-series:

$$s_{ij} = \sqrt{\frac{1}{(t_2 - t_1)} \int_{t_1}^{t_2} \left( \frac{\partial \ln y_j(t, \theta)}{\partial \ln \theta_i} \right)^2 dt}$$  \hspace{1cm} (8)

where $s_{ij}$ represents the sensitivity of the observed $j^{th}$ constituent in the model with respect to the $i^{th}$ parameter, and $t_1$ and $t_2$ represent the simulation period. Since in the Bayesian parameter estimation approach, the value of the parameters is expressed as a probability density, the expected value of sensitivity ($s_{ij}$) can be expressed as:

$$s_{ij} = \int s_{ij}(\theta_i)p(\theta_i|\tilde{Y})d\theta_i$$  \hspace{1cm} (9)

where $p(\theta_i|\tilde{Y})$ is the posterior distribution of parameter $i$ obtained from the Bayesian inference application. The global sensitivities in this study were calculated by sampling the parameter sets provided by the MCMC analysis and performing local sensitivity analyses. The expected values of the sensitivities can then be estimated algebraically.

Equation (9) results in $n_i$ sensitivity values for each parameter associated with the $n_j$ measured constituents, which makes it difficult to gain an overall insight into the sensitivities associated with each parameter. One way would be to add up the values of sensitivity for each assigned parameter to a single value. However, the ability of data on a particular measured constituent to inform about a certain parameter depends on $s_{ij}$ as well as the standard deviations of error $\sigma_i$ for that constituent, as in effect $1/\sigma_i$ acts as a weighting factor on the observed data. Because the standard deviations of errors are treated as unknown parameters in Equation (7), the Bayesian inference provides a way to objectively assign weights to each group of measured data. For this reason, a scaled sensitivity factor (SSF) is defined as:

$$SSF_i = \sum_{j} \frac{s_{ij}}{\sigma_i}$$  \hspace{1cm} (10)

where $SSF_i$ is the sum of the sensitivities of all the observed constituents with respect to the parameter $i$, weighted by the reciprocal of the observation error standard deviation, $\sigma_i$, obtained by applying Equation (7). In fact, SSF is an indicator of the overall sensitivity of the model output with respect to each parameter.

**MODEL APPLICATION**

**Biological nitrogen removal process configuration**

The described parameter estimation framework was applied to estimate the bio-kinetic, stoichiometric, and aeration parameters in the NDN phase of the Blue Plains Advanced WWTP in Washington, DC, USA. The system can be described as a post-denitrification with methanol addition as an external carbon source. The NDN reactor consists of 12 parallel reactors, each having an active volume of 17,500 m$^3$. Each reactor is compartmented into eight tanks in series, including two large and six smaller compartments (Figure 1). Nitrification occurs mainly at tanks 1A, 1B, and 2. Tank 3A is not aerated and serves as the deoxygenation stage. Methanol is added at tank 3B, and denitrification occurs in tanks 3B, 4, and 5A under anoxic conditions. The last tank is aerated again to strip nitrogen gas, which can otherwise hinder the settling of activated sludge in the clarifier. Fine bubble diffusers in the aerated tanks and hyperboloid mixers in the un-aerated tanks are used to create completely mixed conditions.

**Observed data**

The flowrates of influent, effluent, WAS, and RAS, and wastewater characteristics including TSS, VSS, biochemical oxygen demand (BOD), soluble chemical oxygen demand (sCOD), TKN, ammonium (NH$_3$), nitrite and nitrate (NO$_x$), dissolved oxygen, and methanol concentrations data were collected from different sampling locations (Figure 1) during February to June of 2010. Three sets of measured data were used in this study, including the following:

1. External forcing data referring to the flowrate and characteristics of the influent, WAS, RAS, and temperature (Figure 2) that were used as given inputs ($U(t)$ in Equation (2)) to the model.
2. Daily-average flowrate and characteristics of effluent over 150 days of simulation, used as observed data.
3. Spatial profile data referring to the measured concentrations of different constituents in the bioreactors in different sampling locations used as observed data. These observed data were gathered on a weekly to bi-weekly basis as grab samples from points 1 to 9 in tanks 3A through 5A (Figure 1) to take into account the effect of spatial heterogeneity as a source of uncertainty in measured constituents. Approximately 7,000 measured data points taken from different times and locations were used to calculate the likelihood functions. In the Blue Plains advanced WWTP, the influent to the denitrification stage is the effluent of the BOD removal phase (the phase prior to NDN), and thus contains high levels of ammonia and low sCOD (roughly 15 mg N L$^{-1}$ of ammonia and 25 mg L$^{-1}$ of sCOD). Inflows typically varied between 113,000 and 133,000 m$^3$ d$^{-1}$ per reactor depending on the influent wastewater flowrate and the number of reactors in service. A higher inflow fluctuation occurred in the period prior to early April (Figure 2) because the Blue Plains WWTP receives combined domestic wastewater and stormwater runoff and thus faces high flow-rate fluctuations, particularly in early spring. Methanol is applied at a rate of 45 liters per 1,000 m$^3$ of the influent and serves as the carbon source for methylotrophic denitrification. The return flowrate (RAS) ratio to the influent flowrate ($Q_{R}/Q$) varied from 0.5 to 1.3.

**Reaction network**

In the United States, methanol is often used as an external substrate for heterotrophic denitrification because of its availability and cost (Mokhayeri et al. 2008). When the methanol concentration is adequate, a new functional group of specialist heterotrophic microorganisms, collectively referred to as methylotrophs, are able to thrive (Lu et al. 2014b). Methylotrophs only utilize methanol as an electron donor in anoxic conditions and have different growth, yield, and decay rates than generalist heterotrophic organisms that use natural organic matter (Lee & Welander 1996; Hallin et al. 2006; Baytshtok et al. 2008; Dold et al. 2008; Mustakhimov et al. 2013; Alikhani et al. in press). Therefore, a modified version of ASM1 (referred to as M-ASM1), to take into account the anoxic growth and decay of methylotrophs plus the aerobic growth of heterotrophs on methanol, has been developed (Alikhani et al. 2014). The M-ASM1 reaction table is shown in Table 1 along with the process description rate expressions ($R_l$) and stoichiometric coefficients ($\Phi_{l,i}$), which are respectively introduced in Equation (1). In this modification, processes 2, 4 and 7 are added to ASM1. The process 2 was included because the unused methanol in the anoxic zone can be utilized by heterotrophs in the aerobic zone. There are also two new constituents, methanol ($S_{m}$) and methylotrophic biomass ($X_{B,M}$), added to the original 13 constituents of ASM1. Reaction rates 1 to 5 were modified by considering ammonia as a limiting factor to all the biomass growth rates using Monod kinetics (Hauduc et al. 2010). Processes 6 to 7 were modified by implementing the adjustable decay rates under anoxic and aerobic conditions.

**Parameter estimation using MCMC**

All model parameters, including the 95% prior ranges used to construct the prior distributions and temperature correction factor ‘$\theta_T$’ are shown in Table 2. The prior
<table>
<thead>
<tr>
<th>J</th>
<th>Process Description</th>
<th>Reaction</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aerobic growth of heterotrophs</td>
<td>( \frac{1}{Y_H}S_s + \frac{1}{Y_H}Y_HS_o + i_{XB}S_{NH} \rightarrow X_{B,H} )</td>
<td>( \mu_H ) ( S_s ) ( S_o ) ( S_{NH} ) ( X_{B,H} ) ( K_s + S_sK_{O,H} + S_oK_{NH} + S_{NH} ) ( X_{B,H} )</td>
</tr>
<tr>
<td>2</td>
<td>Aerobic growth of heterotrophs on methanol</td>
<td>( \frac{1}{Y_{H,M}}S_m + \frac{1}{Y_H}Y_HS_o + i_{XB}S_{NH} \rightarrow X_{B,H} )</td>
<td>( \mu_H ) ( K_{O,M} + S_mK_{O,H} + S_oK_{NH} + S_{NH} ) ( X_{B,H} )</td>
</tr>
<tr>
<td>3</td>
<td>Anoxic growth of heterotrophs</td>
<td>( \frac{1}{Y_H}S_s + \frac{1}{2.86Y_H}S_{NO} + i_{XB}S_{NH} \rightarrow X_{B,H} )</td>
<td>( \mu_H ) ( K_s + S_sK_{O,H} + S_oK_{NO,H} + S_{NO}K_{NH} + S_{NH} \eta X_{B,H} )</td>
</tr>
<tr>
<td>4</td>
<td>Anoxic growth of methylotrophs</td>
<td>( \frac{1}{Y_M}S_m + \frac{1}{2.86Y_M}S_{NO} + i_{XB}S_{NH} \rightarrow X_{B,M} )</td>
<td>( \mu_M ) ( K_{O,M} + S_mK_{O,M} + S_oK_{NO,M} + S_{NO}K_{NH} + S_{NH} ) ( X_{B,M} )</td>
</tr>
<tr>
<td>5</td>
<td>Aerobic growth of autotrophs</td>
<td>( \frac{4.57 - Y_A}{Y_A}S_o + \left( i_{XB} + \frac{1}{Y_A} \right)S_{NH} \rightarrow X_{B,A} + \frac{1}{Y_A}S_{NO} )</td>
<td>( \mu_A ) ( S_o ) ( S_{NH} ) ( K_{O,A} + S_oK_{NH,A} + S_{NH} ) ( X_{B,A} )</td>
</tr>
<tr>
<td>6</td>
<td>Decay of heterotrophs</td>
<td>( X_{B,H} \rightarrow (1 - f_p)pX_s + f_pX_p + (1 - f_p)i_{XB}X_{ND} )</td>
<td>( b_H \left( \frac{S_o}{K_{O,H} + S_o} + \frac{S_{NO}}{K_{O,H} + S_o} \right) ) ( X_{B,H} )</td>
</tr>
<tr>
<td>7</td>
<td>Decay of methylotrophs</td>
<td>( X_{B,M} \rightarrow (1 - f_p)pX_s + f_pX_p + (1 - f_p)i_{XB}X_{ND} )</td>
<td>( b_M \left( \frac{S_o}{K_{O,M} + S_o} + \frac{S_{NO}}{K_{O,M} + S_o} \right) ) ( X_{B,M} )</td>
</tr>
<tr>
<td>8</td>
<td>Decay of autotrophs</td>
<td>( X_{B,A} \rightarrow (1 - f_p)pX_s + f_pX_p + (1 - f_p)i_{XB}X_{ND} )</td>
<td>( b_A \left( \frac{S_o}{K_{O,A} + S_o} + \frac{S_{NO}}{K_{O,A} + S_o} \right) ) ( X_{B,A} )</td>
</tr>
<tr>
<td>9</td>
<td>Ammonification of soluble organic nitrogen</td>
<td>( S_{ND} \rightarrow S_{NH} )</td>
<td>( k_aS_{ND}X_{Bio} \cdot \left[ X_{Bio} = X_{B,H} + X_{B,M} + X_{B,A} \right] )</td>
</tr>
<tr>
<td>10</td>
<td>Hydrolysis of entrapped organics</td>
<td>( X_S \rightarrow S_S )</td>
<td>( k_H ) ( X_S/X_{Bio} ) ( \left( \frac{S_o}{K_{O,H} + S_o} + \frac{S_{NO}}{K_{O,H} + S_o} \right) ) ( X_{Bio} )</td>
</tr>
<tr>
<td>11</td>
<td>Hydrolysis of entrapped organic nitrogen</td>
<td>( X_{NO} \rightarrow S_{ND} )</td>
<td>( \frac{X_{NO}}{X_S}R_{10} )</td>
</tr>
</tbody>
</table>
Table 2  | M-ASM1 parameter ranges including the posterior ranges

<table>
<thead>
<tr>
<th>Type</th>
<th>Symbol</th>
<th>Parameter Description</th>
<th>Unit</th>
<th>Factor $\theta$</th>
<th>Prior Range</th>
<th>Posterior Range</th>
<th>Expected Value</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.072</td>
<td>2</td>
<td>10</td>
<td>1.742</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.09</td>
<td>0.8</td>
<td>2</td>
<td>0.720</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.072</td>
<td>0.7</td>
<td>1.2</td>
<td>1.059</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>1</td>
<td>10</td>
<td>8.065</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.1</td>
<td>0.5</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.01</td>
<td>0.1</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.04</td>
<td>0.1</td>
<td>0.098</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.01</td>
<td>0.2</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.01</td>
<td>0.2</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.6</td>
<td>0.25</td>
<td>0.185</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.6</td>
<td>0.7</td>
<td>0.611</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.4</td>
<td>0.4</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.24</td>
<td>0.24</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.3</td>
<td>0.5</td>
<td>0.462</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.08</td>
<td>0.08</td>
<td>–</td>
</tr>
</tbody>
</table>

(continued)
distributions of each parameter are assumed to follow a log-normal distribution, as suggested by Cox (2004). The given range for parameters was obtained from various sources (Bullock et al. 1996; Lee & Welander 1996; Cox 2004; Baytshtok et al. 2008; Dold et al. 2008; Hauduc et al. 2011; Alikhani et al. in press). Although the partition coefficient (components of Z vector in Equation (3)) can contribute to the uncertainty, in this research the main focus was on estimating the bio-kinetics and stoichiometric parameters, so the partitioning coefficients were assumed to be known with an acceptable level of certainty.

Using the M–H algorithm, 500,000 samples in 10 chains were drawn from the posterior distribution of parameters. To expedite the convergence of the MCMC algorithm, the deterministically estimated parameters using a hybrid GA with 500 generations, and a population of 50 were used as the starting point of each MCMC chain. This expedites the convergence of the M–H algorithm by decreasing the number of burn-in samples. The posterior probability of each parameter was evaluated by the Bayesian inference introduced in Equation (5), assuming a log-normal error structure. To reach stationary distributions, the estimated parameters in the model, shown with solid lines, along with the prior distribution of parameters, shown as dashed curves. The expected value (mean) and standard deviation of the posterior distributions of parameters are also provided in the last two columns of Table 2. The posterior distribution represents our degree of confidence about the parameter value. By and large, the ranges of values of parameters are consistent with previous studies. For example, the estimated range for parameter $b_H$ (0.5–0.97 d$^{-1}$) is consistent with the deterministically estimated value of 0.926 d$^{-1}$ by Keskitalo & Leiviskä (2012) using the ASM1 model. The estimated 95% range for $b_A$ (0.17–0.21 d$^{-1}$) is larger than the value of 0.12 d$^{-1}$ obtained by Keskitalo & Leiviskä (2012) but consistent with the value of 0.17 d$^{-1}$ suggested by Fall et al. (2011) and the value of 0.18 d$^{-1}$ suggested by Mannina et al. (2012). The estimated range for $K_i$ (4.53–11.22 g COD m$^{-3}$) is smaller than the value of 14.48 g COD m$^{-3}$ found by Keskitalo & Leiviskä (2012). The ranges of $K_{NH, A}$ (0.8–1.06 g N m$^{-3}$) and $K_{NH}$ (0.01–0.03 g N m$^{-3}$) are both smaller than the value of

### RESULTS AND DISCUSSION

#### Bayesian credible interval of parameters

Figure 3 shows the model’s ability to reproduce observed data. The 95% bands have been obtained by performing 1,000 realizations of Monte Carlo simulations using the posterior distributions of parameters and error standard deviations for each observed constituent. As shown in Figure 3, the model performed well in terms of capturing the trends and magnitudes of the observed data. The spread of data points at identical times in stages 3B and 4 is an indicator of the heterogeneity of concentrations inside a single stage, which is represented as homogeneous CSTR in the model. This variability can be due to actual variations in the concentrations at small temporal and spatial scales, or errors due to sampling methods and chemical analysis. It appears that trends in effluent concentrations generally reproduced better results compared to the spatial profile data.

Figure 4 illustrates the posterior distributions for all the estimated parameters in the model, shown with solid lines, along with the prior distribution of parameters, shown as dashed curves. The expected value (mean) and standard deviation of the posterior distributions of parameters are also provided in the last two columns of Table 2. The posterior distribution represents our degree of confidence about the parameter value. By and large, the ranges of values of parameters are consistent with previous studies. For example, the estimated range for parameter $b_H$ (0.5–0.97 d$^{-1}$) is consistent with the deterministically estimated value of 0.926 d$^{-1}$ by Keskitalo & Leiviskä (2012) using the ASM1 model. The estimated 95% range for $b_A$ (0.17–0.21 d$^{-1}$) is larger than the value of 0.12 d$^{-1}$ obtained by Keskitalo & Leiviskä (2012) but consistent with the value of 0.17 d$^{-1}$ suggested by Fall et al. (2011) and the value of 0.18 d$^{-1}$ suggested by Mannina et al. (2012). The estimated range for $K_i$ (4.53–11.22 g COD m$^{-3}$) is smaller than the value of 14.48 g COD m$^{-3}$ found by Keskitalo & Leiviskä (2012). The ranges of $K_{NH, A}$ (0.8–1.06 g N m$^{-3}$) and $K_{NH}$ (0.01–0.03 g N m$^{-3}$) are both smaller than the value of

### Table 2 | continued

<table>
<thead>
<tr>
<th>Type</th>
<th>Symbol</th>
<th>Parameter</th>
<th>Unit</th>
<th>Factor $\theta^a$</th>
<th>Prior Range</th>
<th>Posterior Range</th>
<th>Expected Value</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partitioning coefficients</td>
<td>$i_{XB}$</td>
<td>Nitrogen fraction in biomass</td>
<td>g N g COD$^{-1}$</td>
<td>1</td>
<td>0.05 0.1</td>
<td>0.058 0.005</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$i_{VSS,b}$</td>
<td>COD/VSS ratio of biomass</td>
<td>g COD g VSS$^{-1}$</td>
<td>1</td>
<td>1.42 1.42</td>
<td>– –</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$i_{VSS,i}$</td>
<td>COD/VSS ratio of $X_i$</td>
<td>g COD g VSS$^{-1}$</td>
<td>1</td>
<td>1.5 2</td>
<td>– –</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$i_{VSS,s}$</td>
<td>COD/VSS ratio of $X_s$</td>
<td>g COD g VSS$^{-1}$</td>
<td>1</td>
<td>1.8 1.8</td>
<td>– –</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$i_{VSS,P}$</td>
<td>COD/VSS ratio of $X_P$</td>
<td>g COD g VSS$^{-1}$</td>
<td>1</td>
<td>1.42 1.42</td>
<td>– –</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$i_{MeOH}$</td>
<td>COD ratio of methanol</td>
<td>g COD$^{-1}$</td>
<td>1</td>
<td>1.5 1.5</td>
<td>– –</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass transfer</td>
<td>$k_{LO2}$</td>
<td>Oxygen mass transfer rate</td>
<td>day$^{-1}$</td>
<td>1</td>
<td>50 250</td>
<td>190.446 7.780</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$Temperature dependent parameters: $P(T) = P_0 e^{T-20}$. 

---

Downloaded from https://iwaponline.com/wst/article-pdf/75/6/1370/455020/wst075061370.pdf by guest
Figure 3 | Modeled 95% intervals and measured constituent concentrations at the effluent and inside the bioreactors. The solid lines represent the 95% intervals of model predictions only considering parameter uncertainty, and the gray areas represent the 95% interval considering parameter uncertainty as well as other uncertainties, including measurement errors and model structural error. Dots show the measured data points.
1.41 g N m$^{-3}$ provided by Mannina et al. (2012). The inconsistency between the $K_{NH}$ ranges obtained in this study with previous studies is due to the consideration of two different parameters for ammonia half saturation for autotrophs and heterotrophs, which was not done in ASM1. It worth noting that the presented approach does not capture the possible temporal changes in the values of the parameters that may occur as a result of aging or seasonal effects. However, the presence of such temporal variations in the values of model parameters is implicitly laid within the posterior interval. In order to explicitly capture these temporal changes, the processes leading to them should be incorporated into the model.

The capability of the data to provide information for each parameter can be assessed by comparing the entropy of posterior distribution compared to the prior distribution (Shannon 2001; Lewis et al. in press):

$$I_i = \int p(\theta_i | \mathbf{Y}) \ln[p(\theta_i | \mathbf{Y})] d\theta_i - \int p(\theta_i) \ln[p(\theta_i)] d\theta_i$$  \hspace{1cm} (11)

where $I_i$ is the information content gained for parameter $i$ (referred to as entropy improvement), the first term on the right-hand side is the entropy of posterior distribution and the second term on the right-hand side is the entropy of the prior distribution of parameter $i$. A positive value for $I_i$ shows a degree of information gain for parameter $i$. The higher the value of $I_i$ the higher the level of information is obtained by applying the observed data in the Bayesian inference.

Figure 4 | Posterior distribution (solid lines) and prior distribution (dashed lines) of M-ASM1 parameters for 500,000 MCMC samples.

Downloaded from https://iwaponline.com/wst/article-pdf/75/6/1370/455020/wst075061370.pdf by guest
To evaluate the amount of information content of the observed data two scenarios have been tested, one with higher frequency time intervals of observed data and the second one a lower frequency basis. Figure 5 illustrates the reduction of the 95% credible intervals for both scenarios due to posterior distribution by mapping the 95% equal-tail credible intervals (Mukhopadhyay 2000; Casella & Berger 2002) of each parameter – by excluding 2.5% from each tail of posterior distribution – inside the 95% confidence intervals of prior distributions. Calculated values of $I_i$ for all model parameters for two scenarios are shown in Table 3. In Figure 5, a narrower 95% posterior range of a parameter relative to its prior indicates more information being provided by the observed data. Moreover, the value of $I_i$ in Table 3 quantitatively shows the level of obtained information obtained for the parameter $i$ by applying the observed data.

The values of the entropy improvement ($I$) for each parameter in Table 3 is consistent with the posterior credible intervals shown in Figure 5. In addition, the entropy improvement values ($I$) of the parameters are consistent with the sensitivity level of outputs with respect to them, with parameters having larger SSFs (Figure 6) typically showing a larger increase in the value of $I$. Table 3 shows the entropy improvement for both scenarios of high frequency and low frequency observed data. The comparison between the two scenarios clearly shows that for all parameters, except for some of the observed standard deviations, the information gained from the high frequency scenario is higher than the information obtained from the low frequency scenario. This result indicates that the information content of the observed data is a function of the number of data points and the frequency of sampling.

For the first scenario (labeled as High Frequency (daily) Obs. Data), the observed data provided a good level of a posterior confidence interval for a large number of the parameters, which can be interpreted as reducing a parameter’s uncertainty. For most of the parameters presenting higher sensitivity (Figure 6), including $\mu_H$, $K_{L, O2}$, $\mu_M$, $Y_M$, $K_{M,M}$, $K_{O,H}$, and $b_A$, the uncertainty was reduced by more than 60% in comparison with their prior range. Among the kinetics parameters, large reductions in the uncertainty are obtained for the maximum specific growth rates of heterotrophs and methylophats, $\mu_H$ and $\mu_M$, respectively, and the decay rate for autotrophs, $b_A$. Among the yield coefficients, both methylophats and heterotrophs yield coefficients, $Y_M$ and $Y_H$, showed high levels of reduction in the uncertainty. Most of the half saturation coefficients represent low levels of sensitivity in comparison with growth, decay, and yield parameters, as shown in the ‘Sensitivity analysis’ section (Figure 6), and thus for them lower levels of uncertainty reduction are obtained. The exceptions are the oxygen half saturation of heterotrophs $K_{O,H}$ and the methanol half saturation of methylophats $K_{M,M}$, which showed around 70% reduction in the 95% posterior interval relative to their prior ranges. The oxygen mass transfer rate $K_{L, O2}$ also showed a large reduction in its posterior intervals, although this could be due to the lack of prior knowledge about its value. As for the parameters related to the methylophatic process the
posterior ranges for $\mu_M$, $b_M$, $Y_M$, and $K_{MM}$ were significantly reduced with respect to the given prior range, especially for $\mu_M$ and $b_M$ where the 95% Bayesian credible interval is almost outside the prior range, indicating poor prior knowledge about these parameters that can be enhanced by applying the Bayesian parameter estimation method.

For the second scenario (labeled as Low Frequency (weekly) Obs. Data in Figure 5), the spatial profile data were reduced to one-third of their original size (from 1,650 data points to 450 data points) and the effluent observed data were reduced to weekly intervals instead of their original daily intervals. As can be seen in Figure 5, the posterior ranges of the low frequency observed data are, as might be expected, wider for almost all of the parameters, especially for the $\mu_H$, $K_{OH}$, $\mu_M$, $K_{MM}$, $Y_M$, $\eta_M$, and $K_{LO2}$. However, in some cases, the minimum to maximum ranges are slightly moved. This observation indicates that high frequency observed data carry a higher amount of information about most of the parameters.

A common practice when calibrating in an over-parameterized model such as ASM is to fix the parameters that are deemed to present lower sensitivity and only calibrate the model using the high-sensitivity parameters (Sin et al. 2005). This practice is helpful because it reduces the dimension of the search space and makes the calibration more computationally efficient and it also avoids the chance of becoming
trapped in local minima for an over-parameterized ASM model. To study the effect of this decision the analysis was also performed by only considering the most sensitive parameters (Figure 6). The specific growth rate and decay rate of all types of bacteria ($\mu_H, b_H, \mu_M, b_M, \mu_A$ and $b_A$), the heterotrophic and methylotrophic yield coefficients ($Y_H$ and $Y_M$), and the oxygen mass transfer coefficient rate ($K_{L,O_2}$) were treated as unknown parameters. The remainder of the parameters were fixed at the maximum likelihood estimation value found by the GA. The resulting 95\% intervals for this scenario are shown in Figure 5 (labeled as Most Sensitive Parameters). As shown in Figure 5, the resulting 95\% posterior intervals when only the nine most sensitive parameters are estimated mostly overlap with the results when all the parameters are estimated, with the ranges being mostly narrower when only nine parameters are considered. As expected the narrow confidence brackets obtained for $\mu_H$ and $\mu_M$ and to some degree $Y_m$ and $K_{L,O_2}$, compared to the case when all parameters are included in the analysis (in the scenario of high frequency observed data), shows that when only the most sensitive parameters are considered in the uncertainty analysis the uncertainty assigned to these parameters can be under-estimated. For a few parameters, the uncertainty increases when only the sensitive parameters are considered. This can be due to the model’s lower flexibility to reproduce the data, therefore introducing larger observed error standard deviations.

In all the cases, a moderate to small amount of information was obtained regarding the greater number of parameters. This is likely due to: (a) the model being

![Figure 6](https://iwaponline.com/wst/article-pdf/75/6/1370/455020/wst075061370.pdf)
over-parameterized with respect to the amount and the diversity of data and (b) the fact that the operational conditions of the full-scale bioreactors are fairly constant and the observed data are non-informative regarding some of the parameters under those conditions. The ability of the observed data to inform the values of the parameters (i.e., by providing a narrower posterior interval in comparison to the prior interval) depends on several factors including: (1) the sensitivity of the measured constituents with respect to the parameters, (2) the correlation between the parameters (collinearity), and (3) the weight of the observed constituents that are most sensitive with respect to the parameter.

Sensitivity analysis

Figure 6 shows the ranked SSF for each parameter. Among the stoichiometric and the rate parameters, $Y_M$, $Y_{HI}$, $\mu_A$, $b_A$, $\mu_M$, $b_{HI}$, and $b_{HI}$ showed a higher level of sensitivity compared to the other parameters. For all of the half saturation parameters, the model showed relatively lower sensitivities. The model also showed a relatively higher sensitivity with respect to the oxygen mass transfer rate coefficient than for other parameters. In general, the sensitivity analysis was in agreement with the uncertainty reduction, as the parameters with higher sensitivity achieved a higher level of uncertainty reduction. Nevertheless, sensitivity is not the only factor in the uncertainty analysis. A high to moderate level of information, as determined by the reduction of the 95% posterior interval relative to the prior was obtained for most high-sensitivity parameters, specifically $\mu_H$, $K_{L,O2}$ and $K_{NH,A}$. The exceptions are $K_{O2,HI}$ and $K_{M,M}$, for which a significant level of information was obtained despite a small sensitivity factor. It is speculated that this is because the prior interval attributed to these parameters was exceptionally wide due to a lack of adequate prior knowledge from the literature about their values. The small (effectively zero) measured effluent concentrations of methanol also force $K_{M,M}$ to be small, narrowing its posterior interval.

In general, a positive correlation between SSF and reduction in the credibility interval can be seen. The global sensitivity with respect to most of the parameters for which the information gain was non-significant was found to be small. Global sensitivities of the parameters estimated here represent sensitivities under operational conditions during which the data were collected. This means that as long as the operational condition of the bioreactor is within the range experienced during the data collection period, the influence of the values of parameters with small information gain (lower sensitivity) will not be substantial. However, evaluating the impact of other operational conditions on the value of these parameters requires further study.

The global sensitivity results are fairly consistent with previous studies. In other studies, the maximum growth rates, decay rates, and biomass yields have been reported as high-sensitivity parameters in the ASMs. Kim et al. (2010) reported $i_{Xp}$, $\eta_0$, $k_{NH}$, $b_A$, $\mu_A$, $k_s$, $\eta_h$, and $f_p$ as parameters with large effects on the overall output, while Sharifi et al. (2014) found $Y_A$, $b_{HI}$, $Y_{HI}$, and $\mu_A$ to be highly sensitive parameters for ASM1 based on a local sensitivity analysis at the optimal value of the parameters. Based on a local sensitivity analysis, Fang et al. (2010) reported that effluent COD and ammonia/ammonium (NH) concentrations in ASM3 are more sensitive with respect to $\mu_A$, $\mu_{HI}$, $Y_A$, and $Y_{HI}$.

Correlation between parameters

Posterior results can be used to estimate the posterior correlation between parameters (or parameter collinearity). A high correlation between two parameters means that ways exist to change both of them without seeing a change in the model output or its fitness in terms of the overall ability to reproduce the data. The correlation matrix is provided in the supplementary information for the sake of brevity (available with the online version of this paper). Significant posterior correlations exist between $\mu_{HI}$ and $b_{HI}$, $\mu_{HI}$ and $K_s$, $b_{HI}$ and $b_M$, and $\mu_M$ and $K_{M,M}$. Overall, posterior correlations were smaller than the case where synthetic (model generated) data were used as the observed data (Sharifi et al. 2014). This may be due to greater heterogeneity and structural errors associated with the real observed data, widening the range of the parameters, which reduces the correlation factor.

CONCLUSIONS

In this study, a stochastic parameter estimation method was applied to measured data collected from the NDN reactors at the Blue Plains advanced WWTP in Washington, DC. Wastewater characteristics, including sCOD, VSS, nitrite, ammonia, and methanol concentration, were measured inside the reactor in different tanks, locations, and times, as well as in the effluent wastewater. The goal of this study was to evaluate whether effluent and spatial profile data obtained from a full-scale plant could provide information about the values of model parameters and to quantify the level of confidence they can provide for each parameter relative to its prior knowledge. The following general conclusions were obtained from the presented study.
1. The model was generally able to capture the magnitude and trends of the observed data.

2. The data were able to significantly improve the level of confidence for some of the model parameters as determined by the increase in the entropy of posterior relative to the prior distribution of parameters.

3. Model prediction's uncertainty reduced by applying the posterior distributions of parameters instead of using the prior ranges.

4. Backward uncertainty assessment reveals the correlations between the parameters, representing how the combined variation of two or more parameters can result in the similar reproduction of observed data.

5. By and large, the overall sensitivity of the model output with respect to each parameter as measured by an SSF can explain the ability of data to narrow down the ranges of each parameter.

6. The posterior range of methylotrophic parameters was found to be narrowed down substantially, which reflects the poor prior knowledge about these parameters.

7. The number and the frequency of observed data can influence the level of information gained by the parameters. The results showed that lower frequency in observed data sampling (daily vs. weekly) can narrow down the posterior distribution.

ACKNOWLEDGEMENTS

This project was supported by the District of Columbia Water and Sewer Authority and partially by the DC Water Resources Research Institute.

REFERENCES


Busch, J., Elixmann, D., Kühnl, P., Gerkens, C., Schlöder, J. P., Bock, H. G. & Marquardt, W. 2013 State estimation for large-


Foglia, L., Hill, M. C., Mehl, S. W. & Burlando, P. 2009 Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. Water Resources Research 45 (6), W06427.


First received 21 August 2016; accepted in revised form 14 December 2016. Available online 2 January 2017