Shape Constrained Splines as Transparent Black-Box Models for Bioprocess Modeling

Alma Mašić[a], Sriniketh Srinivasan[b], Julien Billeter[b], Dominique Bonvin[b], Kris Villez[a,*]

[a] *Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland*
[b] *Laboratoire d'Automatique, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

Abstract

Empirical model identification for biological systems is a challenging task due to the combined effects of complex interactions, nonlinear effects and lack of specific measurements. In this context, several researchers have provided tools for experimental design, model structure selection, and optimal parameter estimation, often packaged together in iterative model identification schemes. Still, one often has to rely on a limited number of candidate rate laws such as Contois, Haldane, Monod, Moser, and Tessier. In this work, we propose to use shape-constrained spline functions as a way to reduce the number of candidate rate laws to be considered in a model identification study, while retaining or even expanding the explanatory power in comparison to conventional sets of candidate rate laws. The shape-constrained rate laws exhibit the flexibility of typical black-box models, while offering a transparent interpretation akin to conventionally applied rate laws such as Monod and Haldane. In addition, the shape-constrained spline models lead to limited extrapolation errors despite the large number of parameters.

Keywords: mathematical models, microbial growth-rate kinetics, Monod equation, shape-constrained spline function, wastewater treatment

1. Introduction

Despite major advances in computational tools, the task of building reliable models for process design, monitoring, operation, and automation remains difficult (e.g., Mašić and Eberl, 2014). Quite often, modeling is challenged by the complexity and nonlinearity of the process at hand. In the case of biological systems, especially mixed cultures, a large number of key variables cannot be measured. This typically includes the concentrations of active organisms and their internal metabolites.

*Corresponding author
Email addresses: alma.masic@eawag.ch (Alma Mašić), sriniketh.srinivasan@epfl.ch (Sriniketh Srinivasan), julien.billeter@epfl.ch (Julien Billeter), dominique.bonvin@epfl.ch (Dominique Bonvin), kris.villez@eawag.ch (Kris Villez)

The lack of completeness of experimental data has led to the formulation of the activated sludge model (ASM) family in the case of biological wastewater treatment systems with suspended biomass. These models represent mixed-culture biological systems in a simplified way by identifying the most important groups of bacteria and a macroscopic description of the growth and decay processes associated with them. In these models, one makes use of *switching functions* to describe the most important effects of substrates, products, and inhibiting compounds on the growth and decay processes (Henze et al., 2008). The Monod function (Monod, 1949) is most popular to describe substrate affinity. However, the Monod model is not considered a universal representation of all bacterial behaviors (Moser, 1985). Less popular alternatives include models by Moser (1958), Tessier (1942), and Contois (1959). Importantly, this approach is necessarily empirical. In other words, these switching functions describe empirically established relationships rather than laws derived from first principles. As a result, extrapolation errors can easily be observed when a model is used to optimize process controls (e.g., Sin et al., 2006).

Avoiding extrapolation errors can in part be solved by designing experiments carefully (e.g., Donckels et al., 2009). In addition, frequent model updating might help account for stochastic changes in the process. However, modifying both the model structure and its parameters on a frequent basis leads to large computational efforts for experimental design, model structure selection, and parameter estimation. With the methods proposed and applied in this work, we aim to reduce such efforts and thereby facilitate faster model identification procedures.

Our method relies on the observation that many switching functions have the same shape despite being different functions. This is the case for the affinity switching functions discussed above. Indeed, the Contois, Monod, Moser, and Tessier switching functions exhibit the same increasing and concave shape with respect to the substrate concentration. The Monod function is often used by default, mainly to avoid large computational efforts related to the selection among the list of candidates. However, this can lead to severe extrapolation errors during process design, as is also demonstrated in Neumann and Gujer (2008). Alternatively, one can consider several candidates in a library of rate laws and select via an iterative process of experimental design, parameter estimation, and model structure selection (e.g., Sin et al., 2005). The power of such an approach increases with the number of candidate rate laws, which however results in a larger computational effort. Even if the computational requirements can be satisfied, such an approach can still fail as a library cannot be guaranteed to be universal, that is, to encompass all feasible behaviors (Refsgaard et al., 2006).

To accommodate for the lack of universality discussed above, we propose shape-constrained spline func-

tions (SCS, Villez et al., 2013) as an alternative way to formulate rate laws. Instead of evaluating multiple candidate rate laws with approximately the same shape, we propose to use a single shape-constrained spline function for each considered shape. In other words, we replace all candidate rate laws with a given shape with a single generic rate law. Initial results obtained with this approach were presented at the DYCOPS-CAB2016 conference (Mašić et al., 2016a). The present work expands and completes this study. In Mašić et al. (2016a), simplified biological processes were simulated by assuming that the net growth is zero at all time. This led to the analysis of univariate processes. In this work, this assumption is removed, thus leading to a more general, multivariate approach. In addition, while Mašić et al. (2016a) only dealt with the increasing-concave case described above, we consider here rate laws that include inhibition effects as well. Furthermore, the simulation study in this work includes *(i)* more realism, *(ii)* a single improved parameter estimation method for parameter estimation in practical conditions, *(iii)* a validation test demonstrating that extrapolation errors are limited and *(iv)* a more detailed interpretation and discussion of the results.

The considered spline functions are flexible thanks to the use of a large number of parameters. As a result, they can describe a wide range of kinetic behaviors, akin to black-box modeling approaches (Guay et al., 2004). Note that the application of shape constraints ensures the identifiability and straightforward interpretation of the resulting models, as will be shown below. Shape restrictions are commonly applied for fitting hazard models (Meyer, 2008). More recently, SCS functions were adopted for fault detection and fault diagnosis in a qualitative trend analysis framework (Villez et al., 2013; Villez and Habermacher, 2016). The main difference with these previous studies is that the SCS functions now appear inside a set of nonlinear differential equations.

2. Differences with prior work

The differences with the previous DYCOPS-CAB study (Mašić et al., 2016a) are:

- The substrate and biomass concentrations are considered as state variables, as opposed to the DYCOPS-CAB case, where the biomass concentration was assumed constant. As a consequence, the estimated parameters are associated with the stoichiometry, the growth rate, and the decay rate, whereas the DYCOPS-CAB study only considered the growth rate.

- The estimation of additional parameters in the multivariate case called for the development of a new parameter estimation method (see Section 3.4.2 below).

- The simulated experiments have been modified to appear more realistic. In particular, the sampling frequency used in the DYCOPS-CAB paper has been reduced significantly.
- In this work, only one parameter estimation procedure is used for indirect model fitting of every model (Section 3.4.2). In contrast, the DYCOPS-CAB study used different parameter estimation procedures for the conventional rate laws and the SCS-based rate laws.
- This work includes a validation test, in which the identified models are tested for their extrapolative capability. Such a test was not part of the DYCOPS-CAB study.
- All figures in this paper are new. Although Figures 4-6b bear similarity with figures in the DYCOPS-CAB study, the data and their interpretation have been modified according to the changes made in the simulations. Furthermore, Figures 1-3b and 7a-8, which describe new ideas and results, were not in the DYCOPS-CAB study.
- The discussion and conclusion sections were modified and expanded significantly.

3. Mathematical model & methods

3.1. Model description

In this study, simple models describing bacterial growth and decay are used. The models are similar in structure to the activated sludge models discussed in Henze et al. (2008). Let $S(t)$ and $X(t)$ denote the substrate and biomass concentrations over time t. The change in these concentrations with respect to time can be expressed as

$$\frac{dS}{dt} = -\frac{r_g(S)}{Y}X, \qquad S(0) = S^0 \tag{1}$$

$$\frac{dX}{dt} = r_g(S)X - r_d(X), \qquad X(0) = X^0 \tag{2}$$

where $r_g(S)$ and $r_d(X)$ are rate laws expressing the bacterial growth and decay as a function of S and X, respectively. The metabolic product concentration $P(t)$ can be computed as

$$P(t) = S^0 - S(t), \quad P(0) = P^0. \tag{3}$$

The initial concentrations are S^0, P^0, and X^0. This model describes growth and decay as distinct processes in contrast to Mašić et al. (2016a) which implicitly assumed the two process rates are the same at all times.

The expression $r_g(S)$ for the specific growth rate can be varied to express the effects of substrates, products, and other chemical species. In this study, we consider a classical set of rate laws describing both uninhibited and inhibited bacterial growth processes. This reflects a situation where no a priori knowledge is available about the structure of the kinetic growth-rate law. The considered growth-rate laws are described in the next section. For the decay, a rate law that is linear in the biomass concentration is adopted, as is usual for activated sludge models:

$$r_d(X) = bX \tag{4}$$

with b the specific decay rate constant.

3.2. Growth-rate models

In this section, we define eight growth-rate models that are used to simulate biological growth processes throughout this manuscript. All rate models are depicted in Figure 1 and defined in Table 1. Only the first five of these eight rate laws were considered in the DYCOPS-CAB work (Mašić et al., 2016a).

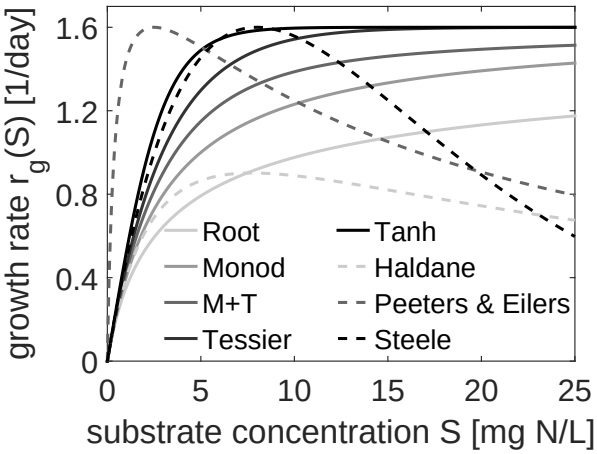


Figure 1: Growth-rate laws considered in this work, as functions of the substrate concentration. The rate laws are defined in Table 1.

The Monod rate law (6) is by far the most commonly applied rate law in biological wastewater treatment models. It has two parameters: μ^{max}, the maximum specific growth rate of the biomass, and K_S, the affinity constant. The Tessier rate law (8) is an alternative growth-rate law. Both the Monod and Tessier rate laws

Table 1: Growth-rate models.

Name	Expression		Reference
Root	$r_R(S) = \mu^{max} \frac{\sqrt{\frac{5S}{\sqrt{1.5}} + 4} - 2}{\sqrt{\frac{5S}{\sqrt{1.5}} + 4} - 2 + K_S}$	(5)	–
Monod	$r_M(S) = \mu^{max} \frac{S}{K_S + S}$	(6)	(Monod, 1949)
M+T	$r_{MT}(S) = \mu^{max} \left(\left(1 - e^{-\frac{S}{K_S}}\right) + \frac{S}{K_S + S} \right)$	(7)	–
Tessier	$r_T(S) = \mu^{max} \left(1 - e^{-S/K_S}\right)$	(8)	(Tessier, 1942)
Hyperbolic tangent	$r_{HT}(S) = \mu^{max} \tanh\left(\frac{S}{K_S}\right)$	(9)	–
Haldane	$r_H(S) = \mu^{max} \frac{S}{K_S + S + \frac{S^2}{K_I}}$	(10)	(Haldane, 1930)
Peeters & Eilers	$r_{PE}(S) = \mu^{max} \frac{2(1+\beta)\frac{S}{S_{opt}}}{\frac{S^2}{S_{opt}^2} + 2\beta\frac{S}{S_{opt}} + 1}$	(11)	(Peeters and Eilers, 1978)
Steele	$r_S(S) = \mu^{max} \frac{S}{S_{opt}} e^{\left(1 - \frac{S}{S_{opt}}\right)}$	(12)	(Steele, 1965)

(i) are increasing and concave, *(ii)* are linear in S for small values of S and *(iii)* are constant for large values of S. Each model has two parameters that need to be determined, namely μ^{max} and K_S. Still, their parameter values cannot be set so that the rate laws deliver the same value at every substrate concentration: the two rate laws intersect at $S = 0$ and in at most two more points for $S > 0$. In this study, the set of Monod and Tessier rate laws are used as the set of candidate rate laws for modeling.

Table 1 and Figure 1 also include a root law r_R, a Monod+Tessier (M+T) law r_{MT}, and a hyperbolic tangent law r_{HT}, which all share the increasing-concave shape with the Monod and Tessier rate laws. The remaining rate laws (Haldane r_H, Peeters & Eilers r_{PE}, and Steele r_S) are rate laws that express both substrate affinity and substrate inhibition. These rate laws differ from the other rate laws by a decreasing trend at high substrate concentrations (substrate inhibition). In Table 1, we introduce the substrate inhibition constant K_I, the maximum growth-rate concentration $S_{opt} = \arg\max_S r(S)$, and the attenuation coefficient β.

In silico experiments are performed by simulating (1)-(4) with every growth rate law in Table 1 and using the parameter values in Table 2. The applied parameter values are different from those used in the

DYCOPS-CAB work (Mašić et al., 2016a). The simulated measurements are used to fit library rate models, namely the Monod and Tessier rate law, as well as the proposed SCS models discussed below.

Table 2: Parameter values used in the simulated calibration experiments. The same values are used for all calibration experiments (Section 3.4.1 and 3.4.2).

Parameter	Value	Unit
b	0.04	1/day
K_I	20	mg N/L
K_S	3	mg N/L
S^0	25	mg N/L
P^0	0	mg N/L
X^0	5	mg X/L
S_{opt} (Steele)	8	mg N/L
S_{opt} (P&E)	2.5	mg N/L
Y	0.1	mg X/mg N
β	3	–
μ^{max}	1.6	1/day
σ_S	0.5	mg N/L
σ_P	0.5	mg N/L
σ_X	0.1	mg X/L

3.3. Shape-constrained spline functions

3.3.1. General treatment

In this study, shape-constrained spline functions are introduced as alternative rate laws. These are B-spline functions and provide a convenient basis to use with shape constraints (Villez et al., 2013; Papp and Alizadeh, 2014; Villez and Habermacher, 2016; Mašić et al., 2016a). For detailed information regarding spline functions, we refer to the *Supplementary Information* and Ramsay and Silverman (2002). The resulting growth-rate laws are piecewise polynomial in the substrate concentration and are given as a weighted sum of spline basis functions:

$$r_{SCS}(S) = \mathbf{b}_0\left(S\right)^{\mathrm{T}}\,\boldsymbol{\theta} \tag{13}$$

with $\mathbf{b}_0(S)$ the $(n_k + n_d - 1)$-dimensional vector of spline basis functions evaluated at the substrate concentration S, and $\boldsymbol{\theta}$ the $(n_k + n_d - 1)$-dimensional vector of model parameters, where n_k is the number of knots and n_d the degree of the spline function. The piecewise behavior is controlled by the location of the n_k knots (or $n_k - 1$ segments) between S_1 and S_{n_k}. For simplicity, the location of these knots are referred to as $\{S_1, S_2, \ldots, S_{n_k}\}$.

Shape constraints on polynomial functions of any order and any nonempty interval of their domain can be specified as a finite number of semi-definite cone constraints (Nesterov, 2000). In special cases, these

inequality constraints reduce to second-order cones or even linear constraints. In previous work (Villez et al., 2013; Villez and Habermacher, 2016), the shape-constrained spline functions were fitted to data pairs consisting of values of the function input and output. Assuming Gaussian noise in the output measurements, the maximum-likelihood estimation problem is a convex optimization problem. Such problems can be solved efficiently to global optimality, even when the number of parameters is large. This property has been exploited by Villez et al. (2013), Papp and Alizadeh (2014), and Villez and Habermacher (2016) to fit SCS functions to univariate data series.

Unfortunately, one cannot expect to measure growth rates directly in practice. Instead, one relies on dynamic experiments during which concentrations of the substrate(s), product(s), and/or biomass are measured. To fit a rate model to such data, one either *(i)* differentiates the measured time series, thus leading to noise amplification, or *(ii)* integrates the rate model to predict the concentrations. The latter option is chosen here for reasons explained in Bhatt et al. (2012). However, this choice *(i)* requires integration of the rate law and *(ii)* makes the fitting problem nonlinear and possibly non-convex in the parameters of the spline function.

3.3.2. Application

In this work, cubic B-spline functions are used ($n_d = 3$). In all cases, the knots are placed equidistantly between $S_1 = 0$ mg/L and $S_{n_k} = 25$ mg/L. Three of the studied spline functions are given special attention in the results section and are referred to as SCS1, SCS2, and SCS3. The SCS1 function has 5 equidistant knots ($n_k = 5$), thus exhibiting $n_k + n_d - 1 = 7$ parameters and an inter-knot distance of 6.25 mg N/L. The SCS2 and SCS3 functions have 17 equidistant knots ($n_k = 17$), leading to 19 parameters and an inter-knot distance of 1.5625 mg N/L. All SCS functions are constrained to go through the origin and have either an increasing-concave shape (SCS1, SCS2) or simply a concave shape (SCS3). Note that none of the SCS functions considered here correspond to any of the SCS functions used in Mašić et al. (2016a). This is due to *(i)* the use of a different function domain ([0-25] mg/L instead of [0-26] mg/L) and *(ii)* the use of differing numbers of equidistant knots. Of practical importance in this work is that the knots are added additively and dyadically, i.e. with increasing parametric complexity new knots are added but never removed from the considered SCS functions.

For cubic polynomials, these shape constraints can be formulated as a set of conditions that are linear in the spline coefficients. More specifically, the following conditions ensure that the rate law goes through the origin and is concave:

$$\mathbf{b}_0(S_1)^{\mathrm{T}}\,\boldsymbol{\theta} = 0 \tag{14}$$

$$\mathbf{b}_2(S_k)^{\mathrm{T}}\,\boldsymbol{\theta} \leq 0 \qquad \forall k = 1, \ldots, n_k \tag{15}$$

with $\mathbf{b}_2(\cdot)$ the second derivatives of the spline basis functions. The increasing-concave shape is ensured by the following *additional* constraints:

$$\mathbf{b}_1(S_{n_k})^{\mathrm{T}}\,\boldsymbol{\theta} \geq 0 \tag{16}$$

with $\mathbf{b}_1(\cdot)$ the first derivatives of the spline basis functions. Figure 2 (top) shows 19 cubic B-spline basis functions $b_0(S)$ that are used to construct an SCS function defined with 17 knots (e.g., SCS2). Figure 2 (bottom) illustrates the SCS rate law with its basis functions multiplied by their associated spline coefficients.

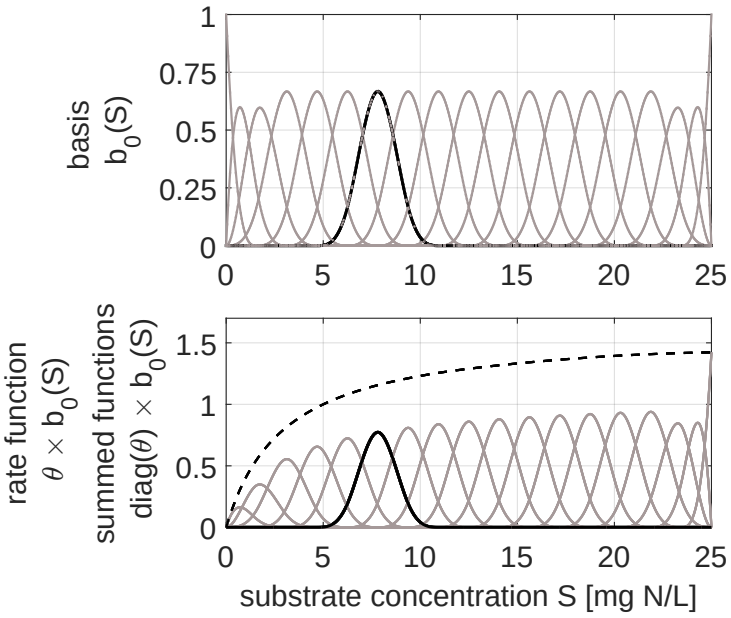
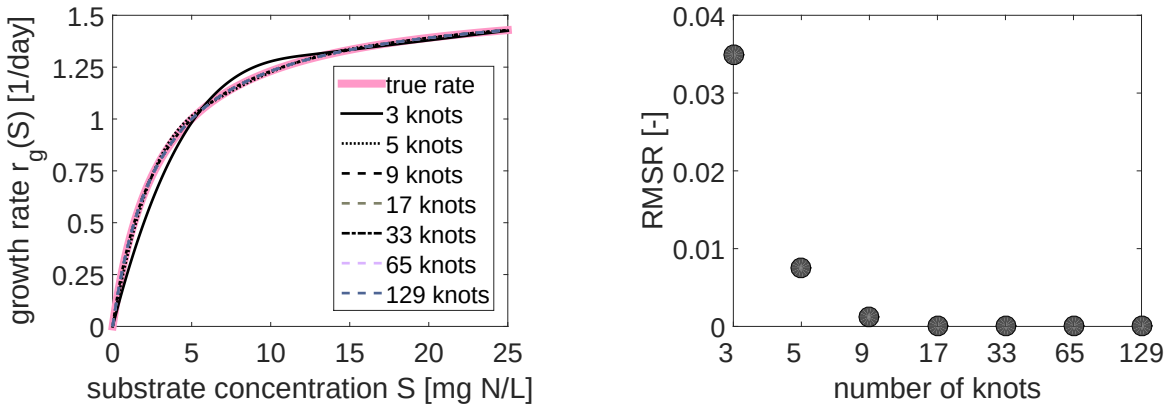


Figure 2: Visualization of the SCS2 rate law with 19 cubic B-spline basis functions with 17 equidistant knots. Top: The spline basis functions are defined over the entire domain of the SCS function but they are constrained to be non-zero in a finite segment of the domain. The black curve highlights a single basis function. The other basis functions are translated and dilated versions of the black curve, except at the domain boundaries. The basis functions are determined completely by the knot locations and the chosen degree of the splines. Bottom: The rate law (dashed line) is given as the sum of 19 functions (full lines), each of which equals a spline basis function (column of $\mathbf{b}_0\,(S)$) multiplied with the associated spline coefficient (element of $\boldsymbol{\theta}$).



(a) Approximation of the Monod rate law by seven different SCS models involving 3 to 129 knots.

(b) Quality of fit expressed by its RMSR value, as a function of the number of knots.

Figure 3: Fitting SCS models to a simulated Monod rate law. The SCS models are characterized by different numbers of knots.

3.4. Model fitting

This section describes the numerical procedure used to identify the parameters of all considered rate models. The parameter estimation procedures are the same as in Mašić et al. (2016a) for Section 3.4.1. For Section 3.4.2, a new procedure is developed to obtain adequate parameter estimates.

3.4.1. Part 1: fitting rate models to noise-free rate measurements

The first part of this study consists in fitting rate models to simulated noise-free rate measurements that were generated using a specific rate model evaluated at $N = 2501$ equidistant points (substrate concentrations). Noise-free measurements are used in this case to demonstrate the approximation properties of the SCS function. The parameters of all models are determined by nonlinear regression. For the Monod and Tessier model this is executed with the trust-region-reflective algorithm (in Matlab: `lsqnonlin`) and for the SCS models with the interior-point algorithm (in Matlab: `fmincon`).

The number of parameters is chosen by the user by selecting the number of knots between S_1 and S_{n_k}. To cover the entire range of substrate concentrations, the first and final knots are set to $S_1 = 0$ mg N/L and $S_{n_k} = 25$ mg N/L.

The optimized quality of fit is the root mean square residual

$$\mathrm{RMSR}_{j,k} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(r_j(S_i) - r_k(\hat{S}_i)\right)^2} \tag{17}$$

where, for the increasing-concave growth-rate expressions, $j \in \{R, M, MT, T, HT\}$ is the simulated rate law,

$k \in \{M, T, SCS1, SCS2\}$ is the candidate growth-rate model, S is the noise-free substrate concentration, and $\hat{S}_{k,i}$ is the modeled substrate concentration obtained with the candidate model k, at equally spaced substrate concentrations S_i with $i = 1, \ldots, N$.

Figures 3a-3b show examples of fitting seven different SCS models to a simulated Monod rate law. The fit is poor for the SCS model with a few knots, but it improves with the increasing number of knots. Eventually, the improvement levels off as the fit is nearly perfect.

3.4.2. Part 2: fitting candidate models to noisy concentration measurements

Since it is highly unrealistic that growth rates can be measured directly, a different approach is taken to handle concentration measurements in the second part of our study. The growth-rate model becomes part of the system of ordinary differential equations (1)-(2) describing the effects of both growth and decay of biomass in a biological reactor. The rate models are fitted by comparing the concentrations obtained via integration of 1 and 2 with the corresponding measured concentrations. This means that the parameter estimation now involves the specific decay rate and the yield coefficients in addition to the parameters of the growth-rate expressions. The quality of the model fit is described by the weighted root mean square residual (WRMSR):

$$\mathrm{WRMSR}_{j,k} = \sqrt{\frac{1}{3\,N}\left(\sum_{i=1}^{N}\left(\frac{\tilde{S}_j(t_i) - \hat{S}_k(t_i)}{\sigma_S}\right)^2 + \sum_{i=1}^{N}\left(\frac{\tilde{P}_j(t_i) - \hat{P}_k(t_i)}{\sigma_P}\right)^2 + \sum_{i=1}^{N}\left(\frac{\tilde{X}_j(t_i) - \hat{X}_k(t_i)}{\sigma_X}\right)^2\right)} \tag{18}$$

where $j \in \{R, M, MT, T, HT, H, PE, S\}$ indicates the simulated rate law, $k \in \{M, T, SCS2, SCS3\}$ indicates the candidate growth-rate model, $\tilde{S}_{j,i}$ ($\tilde{P}_{j,i}$, $\tilde{X}_{j,i}$) is the noisy measured substrate (product, biomass) concentration computed with the rate j, and $\hat{S}_{k,i}$ ($\hat{P}_{j,i}$, $\hat{X}_{j,i}$) is the modeled substrate (product, biomass) concentration obtained with the candidate rate law k, at time point t_i with $i = 1, \ldots, N$.

Parameter estimation is executed in three major steps, namely, *(i)* computation of denoised and interpolated substrate and biomass profiles, *(ii)* computation of initial guesses of all parameters, and *(iii)* nonlinear optimization of the WRMSR value. The first step amounts to fitting SCS functions as proposed in Villez et al. (2013). The second and third steps make use of the trust-region-reflective algorithm (in Matlab: `lsqnonlin`). This procedure is explained in detail in *Supplementary Information.*

3.4.3. Part 3: model validation

To validate the identified models, a new batch experiment is simulated with different initial conditions (S^0 = 15 mg N/L, P^0 = 0 mg N/L, X^0 = 3 mg X/L) and all remaining parameters being the same as

before (Table 2). This is executed for each of the increasing-concave growth-rate expressions. The length of the experiment and the measurement sampling are the same as for the calibration experiments, except that no noise is added to the measurements. This way, the capacity of each model to predict the ground truth (accuracy) is evaluated. Two fitted models are considered for each simulation: *(i)* the best-fitting conventional model (either Monod or Tessier) and *(ii)* the SCS2 model. For each simulated experiment and fitted model, a new WRMSR value is computed with (18) to summarize the performance of each model. This kind of validation experiment was not provided in Mašić et al. (2016a).

3.5. Software availability

All computations are performed with the Matlab R2015a (The Mathworks, 2015) environment, including the Optimization Toolbox, as well as the following additional software: the Functional Data Analysis toolbox by Ramsay and Silverman (2002), the Mosek optimization software by MOSEK ApS (2012), and the SCS toolbox by Villez and Habermacher (2016). All software necessary to reproduce the results presented in this work is available as part of the Efficient Model Identification (EMI) software package for Matlab. This package is *(i)* self-sufficient, apart from the Mosek optimization software, *(ii)* is published under the GPL v3 open-source license, and *(iii)* is added in *Supplementary Information*.

4. Results

4.1. Part 1: fitting rate models to noise-free rate measurements

The growth rate is simulated first using Monod kinetics (see Figure 4). The simulated noise-free rate values are used to fit the candidate models. The Monod and Tessier models are fitted first using the approach described in Section 3.4. Then, two SCS models are fitted to the same data. Figure 4 shows the simulated rate law, the Monod and Tessier models, and the two SCS models. It is easy to see that the Monod model fits the data perfectly, while the Tessier model cannot be adjusted to represent the Monod rate law. Figure 4 also shows that SCS2, with more knots, fits the data slightly better than SCS1. An analogous figure is obtained when the Tessier rate law is used to simulate the rate measurements, for which the Tessier model is found to be the best rate in the library (data not shown).

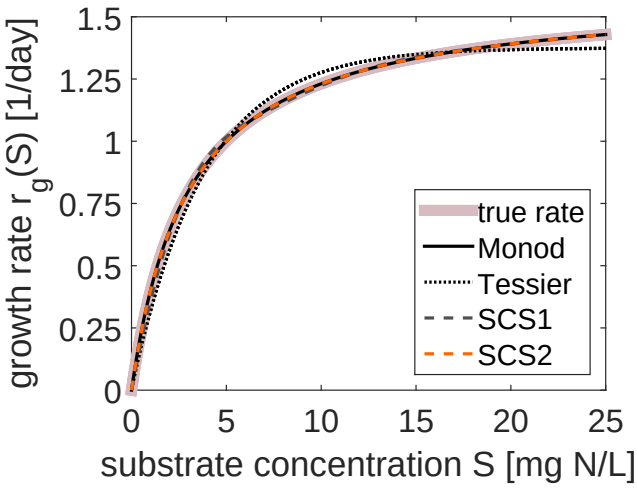


Figure 4: The Monod growth rate [1/day] as a function of the substrate concentration [mg N/L]: noise-free simulated growth rate, Monod model, Tessier model, SCS1 model (5 knots), and SCS2 model (17 knots).

We now assume that the growth rate, which is not known a priori, is any of the monotonically increasing rate laws, namely, the root law r_R, the M+T law r_{MT} and the hyperbolic tangent law r_{HT}, in addition to the Monod and Tessier laws. For each simulated rate, four candidate models are tested, namely, the Monod, Tessier, SCS1, and SCS2 models.

The RMSR values shown in Figure 5 indicate some differences in performance between the four candidate models. While the Monod model fits the Monod rate law best, it is not suited well to fit the other rate laws. Similarly, the Tessier model fits the Tessier rate law best. The SCS1 model, with only 5 knots, delivers a fit

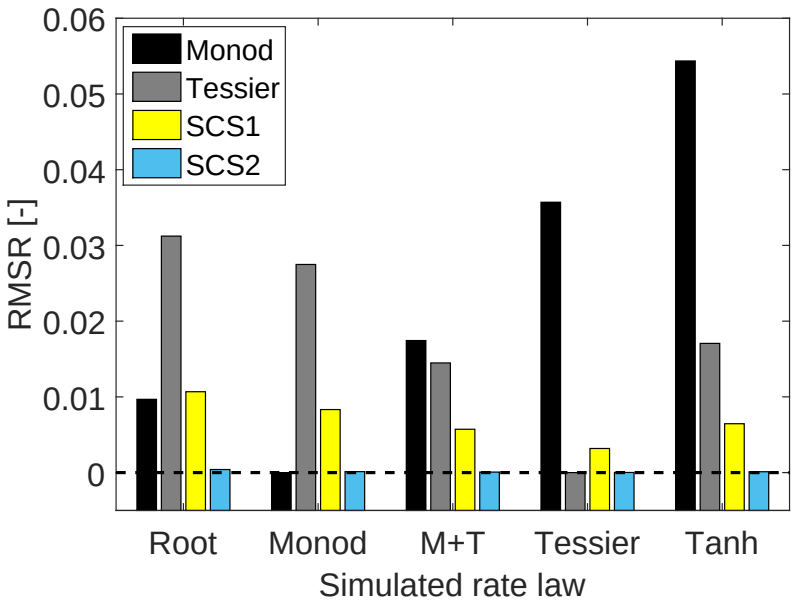
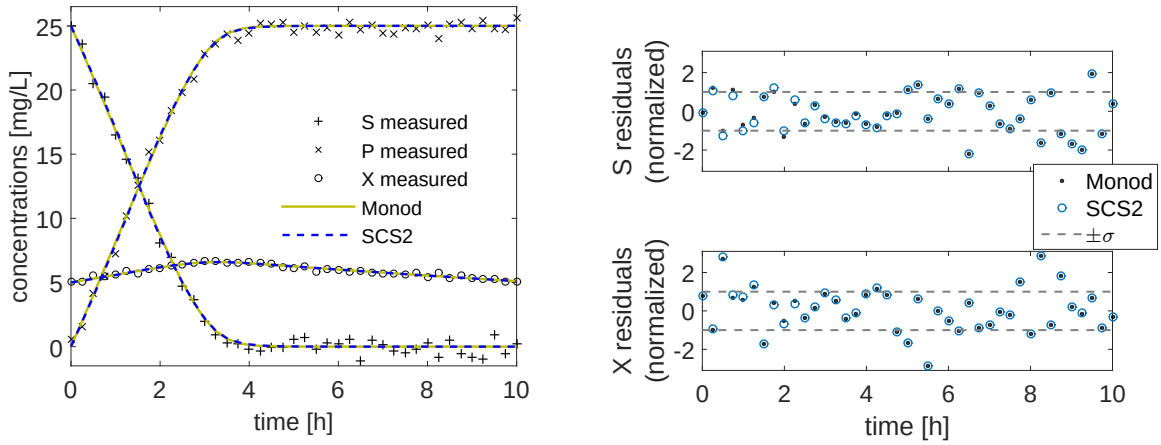


Figure 5: Quality of fit upon fitting the Monod, Tessier, and two SCS models to noise-free growth rates corresponding to five simulated rate laws, namely, root, Monod, M+T, Tessier, and hyperbolic tangent. The SCS models exhibit 5 (SCS1) and 17 (SCS2) knots.

that is better than the incorrect Monod or Tessier model, but not as good as the correct Monod or Tessier model. In contrast, the SCS2 model fits all rate laws very well. Based on these results, and in line with the observations in Figures 3a-3b, SCS2 is selected as the candidate model for fitting the rate parameters to measured concentrations in Section 4.2. SCS2 shows a good trade-off between an excellent fit and a small number of knots.



(a) Substrate, product, and biomass concentrations simulated with the M+T growth rate. Fitting of the Monod and SCS2 models to noisy concentrations.

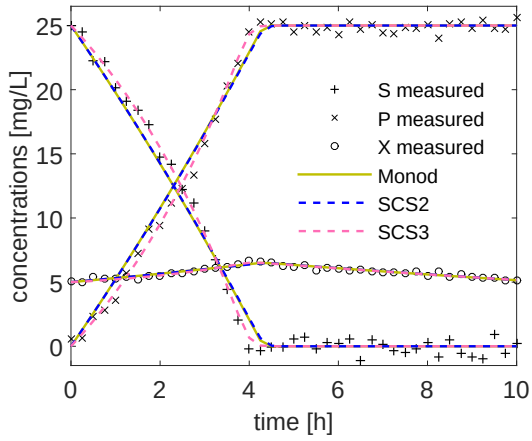
(b) Substrate and biomass concentration residuals with the $1\,\sigma$ measurement error band.

Figure 6: Fitting of Monod and SCS2 models to noisy concentrations obtained with the M+T rate law.

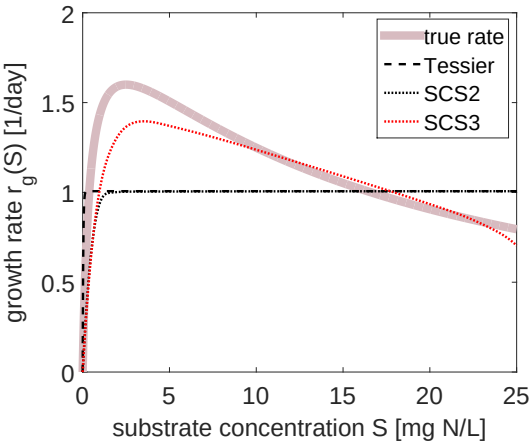
4.2. Part 2: fitting candidate models to noisy concentration measurements

As a first example, the M+T growth-rate law is used to simulate the concentrations. The simulated data points are then corrupted with additive noise ($\sigma_S, \sigma_P, \sigma_X$, see Table 2) to mimic realistic experimental conditions. Figure 6a shows that the substrate concentration decreases with time until it is completely depleted. The product concentration increases with time until all the substrate has been converted to product. The biomass concentration initially increases until all substrate has been depleted, after which it starts decreasing. The figure also shows the fit of the best library model – in this case the Monod model – and the fit of the SCS2 model, for all three concentrations. Figure 6b shows the fits of the Monod and SCS2 models by displaying the residuals for the substrate and biomass concentrations. These residuals are compared to the measurement errors via the indication of the normalized $1\,\sigma$ measurement error bands. Both models approximate the measured data well as most residuals are within the $1\,\sigma$ band. There is no indication of strongly autocorrelated residuals.

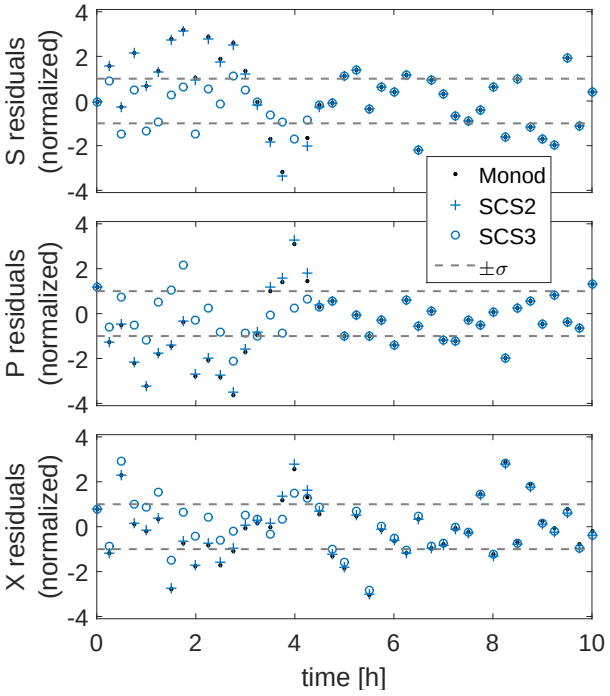
In a second example, noisy measurements are generated using the Steele rate law and following the same procedure as before. The substrate, product, and biomass concentrations are shown in Figure 7a. We can see that the concentration profiles are slightly different from those in Figure 6a, in particular the consumption of substrate is slower. Figure 7b shows that the Tessier and SCS2 models are not able to fit the simulated Steele growth rate well. This is also illustrated by the residuals in Figure 7c, where we can see that they are large and autocorrelated. The residuals go far outside the $1\,\sigma$ measurement error band, particularly in the period 0-4 h.



(a) Substrate, product, and biomass concentrations simulated with the Steele growth rate. Fitting of the Monod, SCS2, and SCS3 models to noisy concentrations.



(b) Fitted Tessier, SCS2, and SCS3 growth-rate models compared with the simulated Steele rate law.



(c) Substrate, product, and biomass concentration residuals with the $1\,\sigma$ measurement error band.

Figure 7: Fitting of Monod, SCS2, and SCS3 models to noisy concentrations obtained with the Steele rate law.

The fitting exercise for the Steele rate law is repeated with the SCS3 model (17 knots). Figure 7a shows that we obtain a much better fit with the concave SCS3 model than with any of the previously used models. Furthermore, Figure 7b illustrates that the SCS3 model is the only one that follows the curve of the Steele rate law, which has a typical inhibition shape. Due to their increasing shape, the other growth rate models are unable to adjust to the correct shape and instead level off once they reach their maximum. The excellent fit of the SCS3 model is also shown by the residuals in Figure 7c, which largely remain within the $1\,\sigma$ measurement error band.

The same exercise is repeated for each of the eight growth-rate laws given in Table 1 and the results of this are summarized by the WRMSR values (18) shown in Figure 8. The behavior is similar to that

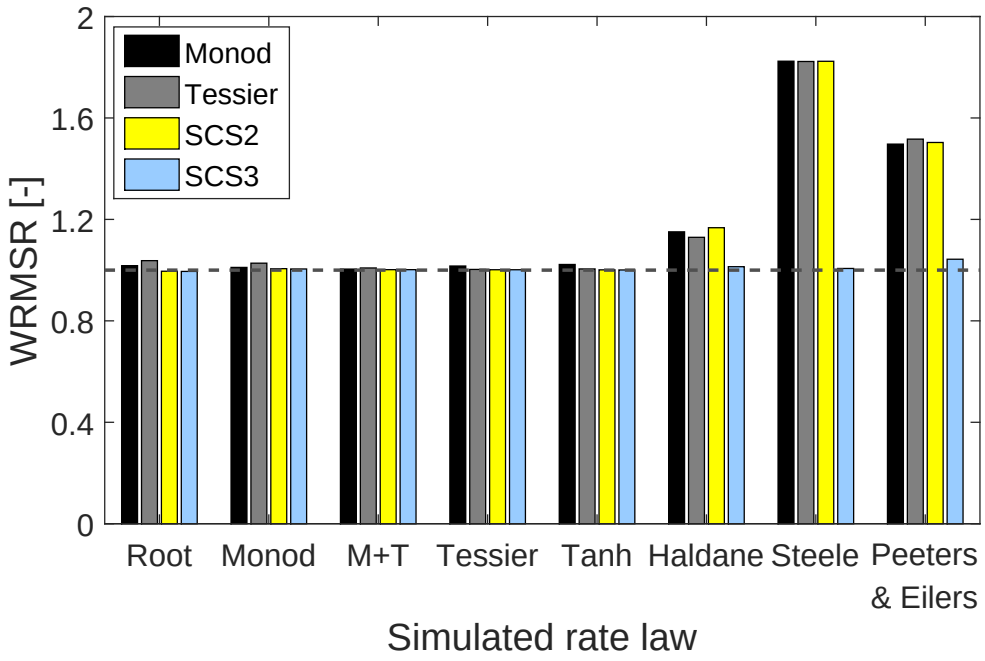


Figure 8: WRMSR values for the Monod, Tessier, SCS2, and SCS3 models fitted to noisy concentrations generated by the eight simulated growth-rate laws in Table 1.

observed in Figure 5, with the Monod model delivering the best fit for the Monod rate law, and the Tessier model delivering the best fit for the Tessier rate law, although the difference is not that large. When the other growth-rate laws are simulated, the performance of the Monod and Tessier models can deteriorate. In contrast, the SCS2 model has an excellent fit for all cases where the shape is increasing and concave. When the shape is not monotonically increasing – for the Haldane, Steele, and Peeters & Eilers rate laws – the quality of fit for the Monod, Tessier, and SCS2 models is much worse. However, the SCS3 model leads to a good fit for all cases, including the rate laws that exhibit inhibition.

4.3. Part 3: model validation

In this section, the performance of the fitted growth-rate models is investigated by means of validation tests. To this end, we compare how well the best-fitting conventional model and the SCS2 model predict the noise-free ground truth simulation for newly simulated experiments with different initial conditions.

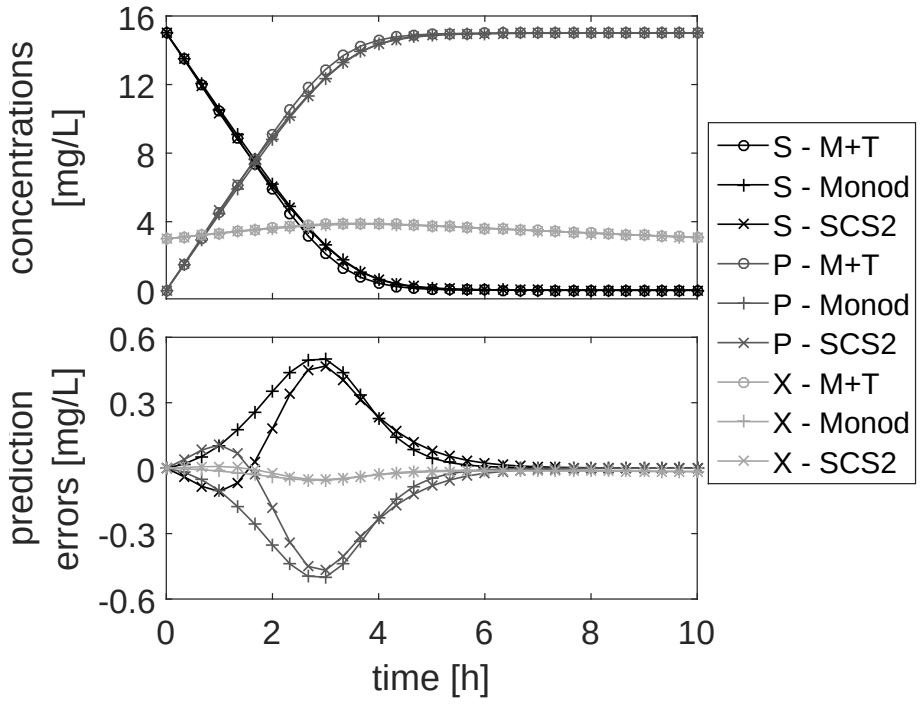


Figure 9: (Top) Simulated concentrations with the true M+T rate law (o), the fitted Monod model (+), and the fitted SCS2 model (x) starting from an initial concentration vector not experienced during the calibration experiment. (Bottom) Prediction errors comparing the fitted Monod and SCS2 models with the true M+T model.

Figure 9 shows the substrate, product, and biomass concentrations generated with the M+T rate law as well as the fitted Monod and SCS2 models for the simulated validation experiment. In the top part of Figure 9 we see small differences between the true and the fitted concentration values. These differences are better illustrated in the bottom part of the figure, which shows the prediction errors. We see that the largest errors are smallest (closest to zero) for the fitted SCS2 model. The WRMSR values are computed as before, however using the ground truth as measurements. The WRMSR values are 0.32 and 0.28 for the Monod and SCS2 models, respectively. Thus, the SCS2 model performs better in a validation test compared to the best-fitting conventional model, in this case the Monod model. A plot of WRMSR values for all validation tests is given Figure 10. This graph shows that the WRSMR values are highest when the data are generated using the M+T model. The prediction errors are well below the measurement uncertainty in all cases, i.e.

WRMSR<1. The WRMSR values for data generated using the Tessier and the hyperbolic tangent models are again smallest with the SCS2 model. However, when the data are generated using the Monod and Root models, the opposite is obtained, namely, the best-fitting conventional model predicts the noise-free ground truth simulation better than the SCS2 model does. Viable hypotheses explaining these observations include *(i)* that there is a lack of convergence during parameter estimation, and *(ii)* there is a trade-off between bias and variance which is different for every ground truth model simulation.

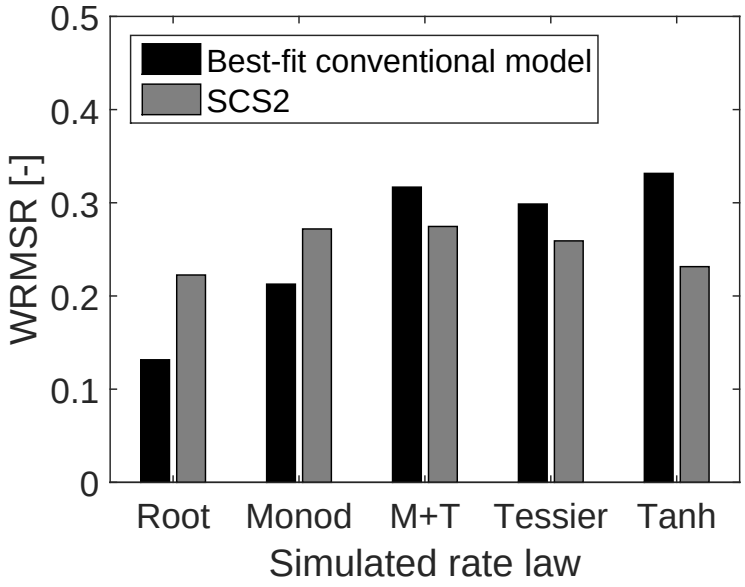


Figure 10: WRMSR values for the best-fitting conventional model (Monod or Tessier) and the SCS2 obtained with *noise-free measurements* in an independent simulation generated by the five increasing-concave growth-rate laws in Table 1.

5. Discussion

The numerical studies performed in this paper indicate that the SCS functions can be a very useful tool for biokinetic process modeling. Specific case studies focus on the most important benefits of SCS-based rate laws, namely, *(i)* their ability to universally approximate functions of a predetermined shape, as demonstrated with rate laws expressing substrate affinity effects without inhibition, and *(ii)* their usefulness as a diagnostic tool during model development, as demonstrated by a clear improvement of the lack-of-fit for rate laws expressing substrate inhibition upon removing the "increasing" shape constraints. The first benefit was demonstrated in Mašić et al. (2016a). In addition to the related conclusions in Mašić et al. (2016a), this work also demonstrates how one can choose an appropriate number of knots in the SCS functions. The use of the SCS functions as a diagnostic tool is newly proposed with this work. The next paragraphs provide an in-depth analysis of all results.

5.1. Near-universal approximation

As expected, the fitting of the Monod and Tessier models to noise-free rate measurements generated with the Monod and Tessier models, respectively, is excellent. When the assumed growth-rate model is incorrect, the quality of fit can decrease significantly. If, for model identification purposes, one would only rely on conventional rate models such as Monod or Tessier, slightly deviating bacterial kinetics might not be captured to a satisfactory degree. In other words, the parametric flexibility of classic biokinetic models proves insufficient to capture a wide range of qualitatively identical rate laws. The limited model coverage shrinks even further if only the (default) Monod model is used for all modeling purposes. In contrast, the SCS models are shown to provide improved generalization properties. In fact, the SCS models are *near-universal* approximators in the sense that any rate law (*universal*) can be approximated to arbitrary precision by adding knots to the fitted SCS function as long as it satisfies the considered shape (*near-*). In our study, 17 knots were shown to be sufficient to approximate the considered increasing-concave rate laws well.

The difference in quality between the Monod and Tessier models on the one hand, and the SCS models on the other, is not as pronounced when the concentration data exhibit the increasing-concave shape (Section 4.2). However, the SCS models offer the advantage that only a single model needs to be fitted. In addition, selecting the best of a few candidate models does not guarantee that the overall best model is found. In contrast, the SCS models deliver a fit that is guaranteed to be better than any other model with the same increasing-concave shape, provided that a sufficient number of knots are used. By means of a number of

validation tests, it was established that the SCS models can provide predictive accuracy on par with the best-fitting conventional model selected from a library. However, a detailed analysis of the bias-variance trade-off is still missing.

5.2. Diagnostic capability

Section 4.2 also evaluated the fit of the proposed SCS models to concentration data generated using rate laws that are not increasing-concave but only concave in shape. The fit of the increasing-concave SCS2 model is clearly worse when the imposed shape is incorrect. In contrast, the only-concave SCS3 model delivers comparable fits whether or not an increasing growth rate can safely be assumed. As a result, the fitting of SCS models with different shapes can be helpful to automatically assess what kind of effects are present in the modeled process. In particular, similar fit of the SCS2 and SCS3 models would suggest the absence of a decreasing trend in the rate law, that is, the absence of substrate inhibition, whereas a dissimilar fit would suggest the presence of substrate inhibition. This diagnostic capability is a major benefit of the proposed SCS modeling framework, given that parameter estimation for only two models (with and without the increasing constraint) is sufficient to arrive at this conclusion.

5.3. Limitations of the study

The results of this study are limited in certain aspects, including *(i)* that all rate and concentration measurements are obtained by numerical simulation and *(ii)* that high-quality and frequent measurements are assumed available for all process states, including the biomass. In this study, highly informative data was necessary to effectively demonstrate all of the reported benefits of SCS models. While lack of informative data does not prevent the use of SCS rate models, it may lead to lack of structural or practical parameter identifiability (Dochain et al., 1995; Vanrolleghem et al., 1995; Bonvin et al., 2016). Lack of structural identifiability may be addressed by model transformation and lumping of parameters as applied in Mašić et al. (2016b). Lack of practical identifiability is typically addressed by means of experimental design or by selecting a smaller number of identifiable parameters, while using a reasonable guess for the remaining parameters. Note also that the models studied in this work are similar to any conventional activated sludge model (ASM) in the sense that the complex metabolism of bacteria is approximated by means of a single differential equation for growth and decay of functionally similar bacterial clades.

In order to establish shape-constrained spline models as alternative biokinetic models, it is necessary to *(i)* study more complex problems involving multiple biomass concentrations, *(ii)* use real measurements from

a batch experiment, and *(iii)* demonstrate the suitability of SCS models for rate law shape determination, as a preparatory step for detailed biokinetic modeling.

5.4. Perspectives and outlook

Future work will consider additional shapes for the SCS functions, always trying to mimic the rate laws in the library, but with more flexibility. Although the SCS functions are described by more parameters than conventional growth-rate laws, the computation time is comparable to that obtained with library models. This new methodology is expected to be beneficial in various applications, including the biological nitrification of collected source-separated urine for resource recovery, the success of which relies on a good model of the nitrification process.

An important application of the diagnostic capability of the SCS functions lies in the determination of the model shape as a preparatory modeling step. By firstly identifying the shape of the underlying rate law, thereby narrowing down the possible model choices, future work could aim at finding a conventional model that fits the data best. This represents a significant advantage of the SCS modeling approach.

Last but not least, the proposed SCS rate laws satisfy requirements that enable global optimization (Mašić et al., 2016b). Furthermore, it is considered likely that the flexibility of SCS rate laws can reduce model bias to the point that reliable predictions can be ensured without the need to account explicitly for model bias as in Reichert and Schuwirth (2012) and Villez et al. (2015). However, these features remain to be demonstrated.

6. Conclusions

In this work, shape-constrained spline models have been integrated into a set of differential equations to simulate and model biological wastewater treatment processes. The proposed dynamic SCS models can fit qualitatively similar growth-rate laws in a universal manner and require less computational efforts than searching through a library of rate laws. The SCS model is a black-box model in essence, but the ease of interpretation gives it a white-box flavor. When faced with an unconventional growth rate that is not part of a library, it is still possible to estimate a good predictive model with the SCS approach, as long as the assumed overall shape remains valid. In addition, when faced with an unknown shape, the SCS approach is useful for shape-based diagnosis of the model by determining whether the observed growth rate possesses certain features. Such a shape-based exclusion is practically impossible with the library approach.

Future work aims at using the SCS approach with different and more complex shapes, with particular emphasis on the determination of the model shape as a preparatory step in conventional white-box modeling.

Moreover, one can use the SCS model and perform a sensitivity analysis that provides insights on the model suitability for process design and optimization. Finally, experimental evaluation is necessary to investigate the performance of SCS on real measured data.

Acknowledgments

This study was made possible by Eawag Discretionary Funds (grant no.: 5221.00492.009.03, project: DF2015/EMISSUN). We thank the authors of the Functional Data Analysis toolbox by Ramsay and Silverman (2002), the Mosek optimization software MOSEK ApS (2012), and the SCS toolbox by Villez et al. (2013) and Villez and Habermacher (2016).

References

Bhatt, N., Kerimoglu, N., Amrhein, M., Marquardt, W., Bonvin, D., 2012. Incremental identification of reaction systems – A comparison between rate-based and extent-based approaches. Chemical Engineering Sci. 83, 24–38.

Bonvin, D., Georgakis, C., Pantelides, C., Barolo, M., Grover, M., Rodrigues, D., Schneider, R., Dochain, D., 2016. Linking models and experiments. Ind. Eng. Chem. Res. 55 (25), 6891–6903.

Bredeweg, B., Linnebank, F., Bouwer, A., Liem, J., 2009. Garp3 – Workbench for qualitative modelling and simulation. Ecological Informatics 4(5), 263–281.

Contois, D. E., 1959. Kinetics of bacterial growth: relationship between population density and specific growth rate of continuous cultures. J. Gen. Microbiol. 21, 40–50.

de Boor, C., 1978. A Practical Guide to Splines. Springer.

Dochain, D., Vanrolleghem, P. A., Van Daele, M., 1995. Structural identifiability of biokinetic models of activated sludge respiration. Water Research 29 (11), 2571–2578.

Donckels, B. M., De Pauw, D. J., De Baets, B., Maertens, J., Vanrolleghem, P. A., 2009. An anticipatory approach to optimal experimental design for model discrimination. Chemometrics and Intelligent Laboratory Systems 95 (1), 53–63.

Guay, M., Dochain, D., Perrier, M., 2004. Adaptive extremum seeking control of continuous stirred tank bioreactors with unknown growth kinetics. Automatica 40 (5), 881–888.

Haldane, J. B. S., 1930. Enzymes. Reprint by MIT, Cambridge.

Henze, M., van Loosdrecht, M., Ekama, G., Brdjanovic, D. (Eds.), 2008. Biological Wastewater Treatment: Principles, Modelling and Design. IWA Publishing.

Kuipers, B. J., 1994. Qualitative Reasoning: Modeling and simulation with incomplete knowledge. MIT Press, Cambridge, MA, USA.

Mašić, A., Eberl, H. J., 2014. On optimization of substrate removal in a bioreactor with wall attached and suspended bacteria. Math. Biosci. Eng. 11 (5), 1139–1166.

Mašić, A., Srinivasan, S., Billeter, J., Bonvin, D., Villez, K., 2016a. On the use of shape-constrained splines for biokinetic process modeling. 11th IFAC Symposium on Dynamics and Control of Process Systems Including Biosystems DYCOPS-CAB 2016, Trondheim, Norway, 6-8 June 2016 – IFAC-PapersOnLine 49 (7), 1145–1150.

Mašić, A., Udert, K., Villez, K., 2016b. Global parameter optimization for biokinetic modeling. Environmental Modelling & Software 85, 356–373.

Meyer, M. C., 2008. Inference using shape-restricted regression splines. Annals of Applied Statistics 2, 1013–1033.

Monod, J., 1949. The growth of bacterial cultures. Ann. Rev. Microbiol. 3, 371–394.

MOSEK ApS, 2012. MOSEK Optimization Software for MATLAB, Version 6.0. Available from http://www.mosek.com/.

Moser, A., 1985. Kinetics of batch fermentations. In: Rehm, H.-J., Reed, G. (Eds.), Biotechnology. Vol. 2. Fundamentals of Biochemical Engineering. VCH, Weinheim.

Moser, H., 1958. The dynamics of bacterial populations maintained in the chemostat. Washington: Carnegie Institution.

Nesterov, Y., 2000. High Performance Optimization, Applied Optimization. Vol. 33. Kluwer Academic Publishers, Ch. Squared functional systems and optimization problems, pp. 405–440.

Neumann, M., Gujer, W., 2008. Underestimation of uncertainty in statistical regression of environmental models: Influence of model structure uncertainty. Environ. Sci. Technol. 42 (11), 4037–4043.

Papp, D., Alizadeh, F., 2014. Shape-constrained estimation using nonnegative splines. Journal of Computational and Graphical Statistics 23, 211–231.

Peeters, J. C. H., Eilers, P., 1978. The relationship between light intensity and photosynthesis – a simple mathematical model. Hydrobiological Bulletin 12 (2), 134–136.

Ramsay, J., Silverman, B., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag, New York.

Refsgaard, J., der Sluijs, J. V., Brown, J., der Keur, P. V., 2006. A framework for dealing with uncertainty due to model structure error. Adv. Water Resour. 29 (11), 1586–1597.

Reichert, P., Schuwirth, N., 2012. Linking statistical bias description to multiobjective model calibration. Water Resources Research 48, 1–20.

Sin, G., Van Hulle, S. W., De Pauw, D. J., Van Griensven, A., Vanrolleghem, P. A., 2005. A critical comparison of systematic calibration protocols for activated sludge models: A swot analysis. Water Research 39 (12), 2459–2474.

Sin, G., Villez, K., Vanrolleghem, P. A., 2006. Application of a model-based optimisation methodology for nutrient removing SBRs leads to falsification of the model 53(4-5), 95–103.

Steele, J. H., 1965. Notes on some theoretical problems in production ecology. In: Goldman, C. R. (Ed.), Primary Productivity in Aquatic Environments (Mem. Ist. Ital. Idrobiol., 18 Suppl.). University of California Press, Berkeley, pp. 383–398.

Tessier, G., 1942. Croissance des populations bactériennes et quantité d'aliment disponible. In: Review Science extrait du No. 3208. pp. 209–214.

The Mathworks, 2015. MATLAB online documentation. Accessed on October 28, 2015.
URL `http://www.mathworks.com/help/matlab/`

Vanrolleghem, P. A., Van Daele, M., Dochain, D., 1995. Practical identifiability of a biokinetic model of activated sludge respiration. Water Research 29 (11), 2561–2570.

Villez, K., Giudice, D. D., Neumann, M. B., Rieckermann, J., 2015. The statistical description of model bias as a method to account for model structure errors. Proceedings of the 9th IWA Symposium on Systems Analysis and Integrated Assessment (Watermatex2015), Gold Coast, Queensland, Australia, June 14-17, 2015, appeared on-line.

Villez, K., Habermacher, J., 2016. Shape anomaly detection for process monitoring of a sequencing batch reactor. Computers & Chemical Engineering 91, 365–379.

Villez, K., Venkatasubramanian, V., Rengaswamy, R., 2013. Generalized shape constrained spline fitting for qualitative analysis

of trends. Comp. Chem. Eng. 58, 116–134.

Supplementary information

In this section, we describe the construction of spline functions and the application of shape constraints to such functions. Furthermore, we describe in detail the parameter estimation procedure used to obtain reasonable parameter initial guesses and the final nonlinear optimization.

S.1. Splines: construction

Spline functions are constructed as smooth curves that approximate a set of points. In essence, each spline function is a sum of several piecewise polynomial curves, each defined on a certain segment of the domain. The number of segments is determined by the number of knots that divide the domain into intervals. In this paper, we use cubic basis spline (B-spline) functions.

Cubic B-splines. Let the domain be divided into $n_k - 1$ intervals with n_k knots bounding the intervals. The location of these knots in the domain are referred to as $S_1, S_2, \ldots, S_{n_k}$. We define a B-spline of order 4 as a piecewise polynomial function of degree $n_d = 3$. For the sake of simple demonstration, we will use the SCS1 function which is defined by its $n_k = 5$ knots. The SCS1 function can thus be expressed as a linear combination of B-splines

$$SCS1(S) = \sum_{i=1}^{n_k+n_d-1} B_{i,3}(S)\ \theta_i \tag{S.1}$$

where θ_i are the spline coefficients (parameters) and $B_{i,3}$ are the B-splines of degree 3. These B-splines can be constructed by the recurrence relation

$$B_{i,n_d}(S) = \frac{S - S_i}{S_{i+n_d} - S_i} B_{i,n_d-1}(S) + \frac{S_{i+n_d+1} - S}{S_{i+n_d+1} - S_{i+1}} B_{i+1,n_d-1}(S) \tag{S.2}$$

where the functions $B_{i,0}$ are given by

$$B_{i,0}(S) = \begin{cases} 1, & S_i \leq S \leq S_{i+1} \\ 0, & \text{otherwise.} \end{cases} \tag{S.3}$$

For the purpose of such a construction, the first knot is repeated $n_d + 1$ times in the series of knots. Thus, to construct a cubic spline function recursively the values for S_i used in the above recursion are $S_1, S_1, S_1, S_1, S_2, S_3, \ldots, S_{n_k}$ (de Boor, 1978). Conventionally, the spline coefficients are fitted by minimizing a convex objective function. Most typically, a least-squares fit is obtained (as in Section 3.4.1 of our work).

Regularized fitting is also popular (e.g., Ramsay and Silverman, 2002). Figure S.1 shows the basis functions of SCS1 and SCS1 itself.

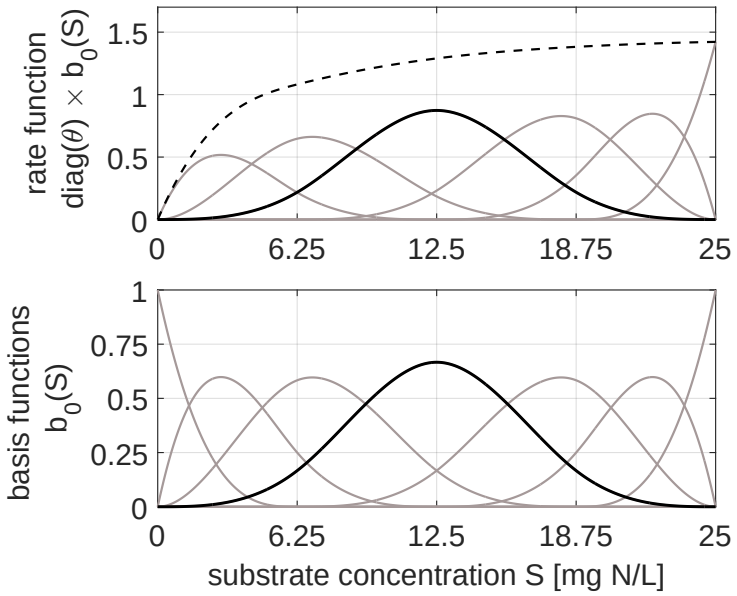


Figure S.1: Visualization of the SCS1 function with 7 cubic B-spline functions and 5 knots. (Bottom) All the 7 basis functions. (Top) The basis functions multiplied by their parameters. The dashed line shows the sum of the basis functions.

S.2. Splines: applying shape constraints

In order to enforce a shape on the spline function, one can constrain the parameters in such a way that the desired shape is maintained. In this study, we use a concave shape and a monotonically increasing concave shape. For the concave shape, we require the first derivative in all knots to be less than or equal to zero (see (15) in Section 3.3). To obtain the monotonically increasing shape, we require the second derivative in the final knot to be nonnegative (see (16)). In general, such shape constraints amount to semi-definite cone constraints which can be reduced to second order cone inequality constraints, quadratic inequality constraints, or linear inequality constraints. This is discussed at length in Nesterov (2000); Villez et al. (2013); Papp and Alizadeh (2014). In the specific cases studied in this work, these shape constraints are described completely as linear inequality constraints in the parameters.

S.3. Parameter estimation procedure

The following parameter estimation procedure provides reasonable estimates for the parameters for every considered model. It is noted that this procedure does not guarantee that globally optimal parameter estimates are found. An extension of the global parameter estimation method in Mašić et al. (2016b) for

the models considered in this work is currently being developed. The following steps are repeated for every simulated experiment and every considered model.

S.3.1. Step 1 – Denoising and interpolation

Denoising and interpolation of the substrate concentration and the biomass concentration is executed by fitting two SCS functions with the method provided in Villez et al. (2013). The first function (F1) is fitted to the series obtained by subtracting the product concentration measurements minus the initial substrate concentration from the substrate concentration measurements followed by division by two. This series is a least-squares estimate of the substrate concentration given the stoichiometric balance. This function is constrained to consist of two episodes (as defined in Villez et al. (2013)). The first episode has a decreasing-concave shape; the second one a decreasing-convex shape. This shape corresponds to the true shape of any substrate profile that can be obtained with any of the considered models and irrespective of the presence of inhibition. This can be proven via qualitative simulation (not shown, Kuipers, 1994; Bredeweg et al., 2009). The second function (F2) is fitted to the biomass concentration series and is constrained to consist of four episodes which are increasing-convex, increasing-concave, decreasing-concave, and decreasing-convex. This is again the only feasible shape of the biomass profile for any of the considered models. After fitting the two functions, the spline functions are resampled at equidistant times with an interval of 0.01 h. The two fitted functions are shown in Figure S.2 for the calibration experiment simulated with the M+T rate expression.

S.3.2. Step 2 – Initial parameter guesses

Initial parameter guesses are obtained via a sequence of three substeps (Step 2(a), Step 2(b), and Step 3(c)). This estimation problems solved in each of these steps are executed with the trust-region-reflective algorithm (in Matlab: `lsqnonlin.m`).

Step 2(a) – Guesses for the specific substrate utilization rate parameters. An initial guess for the growth rate parameters are obtained with the F1 and F2 obtained in Step 1. An interpolated substrate consumption rate is computed analytically as the 1st derivative of F1 at the interpolating points. A corresponding biomass concentration is obtained by evaluation of F2 at the interpolating points. We divide the substrate consumption rate by the biomass concentration to obtain an estimate of the substrate utilization rate. These substrate utilization rates are estimates of $r_g(S)/Y$ and are set out against the corresponding substrate concentrations for nonlinear regression. To this end, Y is arbitrarily assumed to be 1 in this step. For the conventional models, this means that the obtained parameter estimate for μ_{max} should be divided by Y

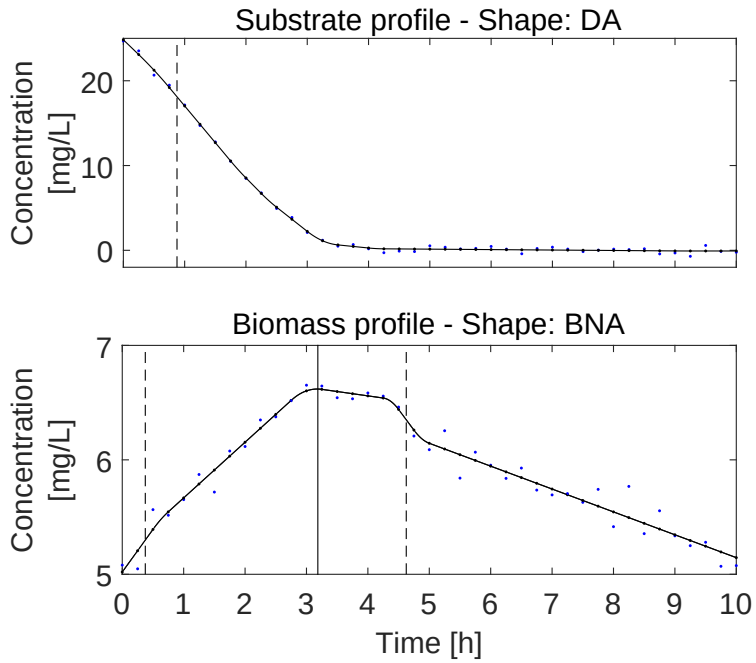


Figure S.2: Fitting of shape constrained spline functions to denoise and interpolate the substrate measurement profile (top) and the biomass measurement profile (bottom). Vertical dashed lines indicate the identified inflection points. The full line indicates the maximum in the biomass profile.

when an estimate for Y becomes available. Similarly, all spline coefficients for the SCS functions should be divided by Y as well. An example is shown in Figure S.3.

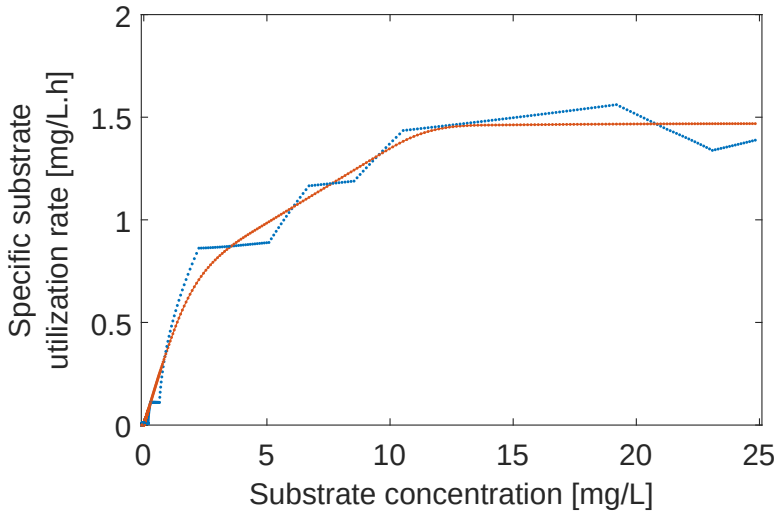


Figure S.3: Computation of an initial guesses for the substrate utilization parameters. The estimated substrate utilization rate against the substrate concentration (blue dots) is approximated by manipulating the parameters of $r_g(S)$ in the expression $r_g(S)/Y$ for the specific utilization rate (red line).

Step 2(b) – Guess for the decay rate expression parameters. An initial guess for the decay rate parameter is obtained by evaluating F2 in the second half of the experiment (t=5-10h). During this time, the biomass concentration dynamics are governed by biomass decay only as all of the provided substrate has been depleted. The first derivative of F2 is set out against F2 and a linear line going through the origin is fitted in the least-squares sense to these data pairs. The slope of this line corresponds to a guess for the specific decay rate (see Figure S.4).

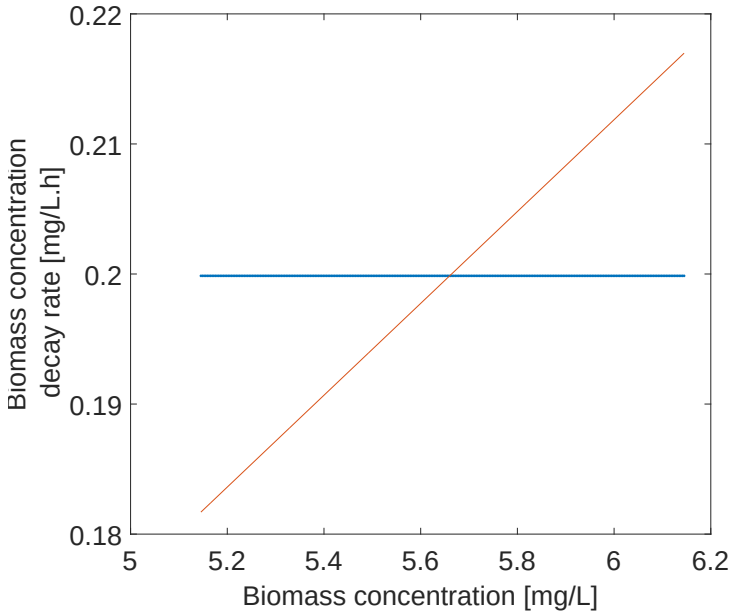


Figure S.4: Computation of an initial guess of the decay parameter. The estimated decay rate against the biomass concentration (blue dots) is approximated by a linear line through the origin (red line). Note that the fitted line does not fit the estimates well in this case.

Step 2(c) – Guess for the yield parameter. The cumulative volume-specific substrate consumption is computed by evaluating F1 and subtracting it from the initial substrate concentration. The accumulated biomass lost through decay is obtained by integrating the decay rate with its parameter guess obtained in Step 2(b) and using the function F2 to obtain biomass concentrations at every time. This lost biomass is added to the interpolated values of the biomass profile (F2) and the initial biomass concentration is subtracted. This delivers the accumulated volume-specific biomass production that was generated through the growth process. The accumulated volume-specific biomass production and the cumulative volume-specific substrate consumption are plotted against each other and a line is fitted in the least-squares sense to these data pairs. The slope of this line corresponds to a guess for the yield coefficient. This is demonstrated in Figure S.5. The initial guesses for μ_{max} or the spline function coefficients describing $r_g(S)/Y$ identified in Step 2(a) are

now modified by multiplying them with the guess for the yield coefficient. These scaled parameters now describe the initial guess for $r_g(S)$. This completes the computation of initial parameter guesses.

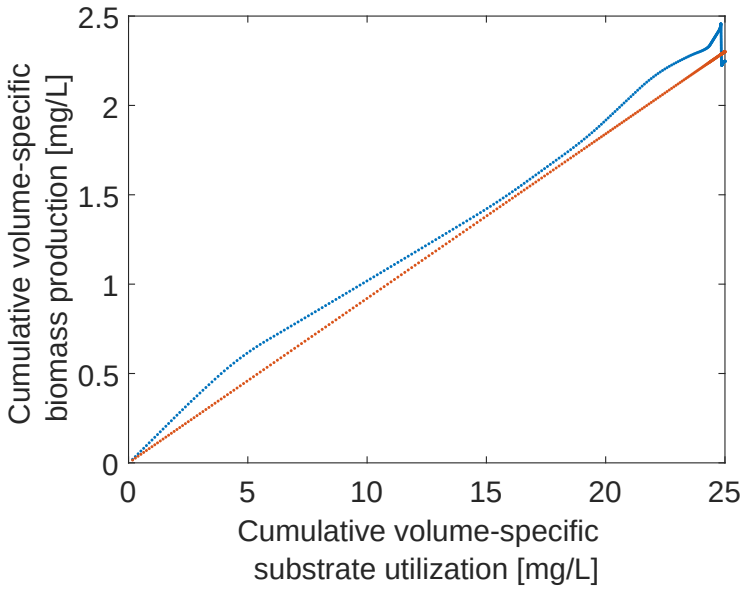


Figure S.5: Computation of an initial guess of the yield parameter. The estimated cumulative biomass production against the cumulative substrate utilization (blue dots) is approximated by a linear line through the origin (red line).

S.3.3. Nonlinear optimization

The final parameter estimates are obtained by manipulating all parameters simultaneously in order to minimize the WRMSR value. This is executed by means of the trust-region-reflective algorithm (in Matlab: `lsqnonlin.m`) which uses the parameter guesses obtained in Step 2 as the initial parameter vector. In Figure S.6, a simulation with both initial guesses and final parameters of the SCS2 model is shown.

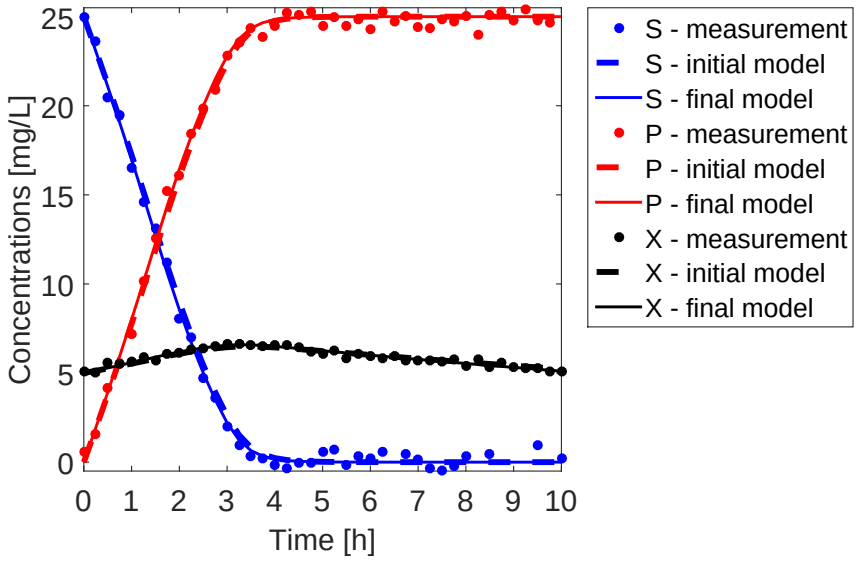


Figure S.6: Measurements (dots), initial model simulation (dashed lines), and final model simulation (full lines).