Identification of Biokinetic Models using the Concept of Extents

Alma Masic, Sriniketh Srinivasan, Julien Billeter, Dominique Bonvin, and Kris Villez

Environ. Sci. Technol., Just Accepted Manuscript • DOI: 10.1021/acs.est.7b00250 • Publication Date (Web): 01 Apr 2017
Downloaded from http://pubs.acs.org on April 12, 2017

Just Accepted

“Just Accepted” manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides “Just Accepted” as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. “Just Accepted” manuscripts appear in full in PDF format accompanied by an HTML abstract. “Just Accepted” manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). “Just Accepted” is an optional service offered to authors. Therefore, the “Just Accepted” Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these “Just Accepted” manuscripts.

This document is the accepted manuscript version of the following article:
Graphical abstract for TOC
Graphical abstract for TOC
47x26mm (600 x 600 DPI)
Identification of Biokinetic Models
using the Concept of Extents

Alma Mašić,† Sriniketh Srinivasan,‡ Julien Billeter,‡ Dominique Bonvin,‡ and
Kris Villez*,†

†Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133,
CH-8600 Dübendorf, Switzerland
‡Laboratoire d’Automatique, École Polytechnique Fédérale de Lausanne, CH-1015
Lausanne, Switzerland

E-mail: Kris.Villez@eawag.ch

Abstract

The development of a wide array of process technologies to enable the shift from
conventional biological wastewater treatment processes to resource recovery systems is
matched by an increasing demand in predictive capabilities. Mathematical models are
excellent tools to meet this demand. However, obtaining reliable and fit-for-purpose
models remains a cumbersome task due to the inherent complexity of biological wastew-
ater treatment processes. In this work, we present a first study in the context of envi-
ronmental biotechnology that adopts and explores the use of extents as a way to sim-
plify and streamline the dynamic process modeling task. In addition, the extent-based
modeling strategy is enhanced by optimal accounting of nonlinear algebraic equilibria
and nonlinear measurement equations. Finally, a thorough discussion of our results
aims at demonstrating the benefits of extent-based modeling and its potential to turn
environmental process modeling into a highly automated task.
Introduction

Dynamic models are increasingly used to better understand, design, and operate environmental processes.\textsuperscript{1,2} For biological wastewater treatment processes, the available activated sludge model family\textsuperscript{3,4} has been used widely despite reported challenges in model identification. These challenges relate to (i) the information content and the quality of calibration data that limit practical identifiability,\textsuperscript{5-10} (ii) the lack of mechanistic understanding,\textsuperscript{11,12} and (iii) nonlinear and non-convex properties.\textsuperscript{13-15} These issues are even more severe in the case of decentralized treatment processes that are proposed to address fast societal dynamics by providing straightforward upscaling of wastewater treatment operations.\textsuperscript{16} In addition, both economical and political motives are driving a paradigm shift in objectives from environmental protection to a need to generate added-value products from wastewaters. To ensure both product quality and economically optimal operations, resource recovery from wastewater requires tight management and control of the involved processes. The urine nitrification process for fertilizer production developed at Eawag is an example of this.\textsuperscript{17} Advanced control of such high-rate processes is not possible without detailed process understanding and predictive power. In addition, the diversity of available technologies is rapidly increasing. For this reason, fast development of reliable models is paramount to attain sustainable urban water cycles.

In the past, model complexity has been been tackled by means of model identification protocols. Examples include (i) protocols that split model identification into steps corresponding to major fractions of the medium\textsuperscript{18} and (ii) protocols based on iterative model building.\textsuperscript{19} Despite these efforts, the aforementioned model identification challenges have only been partly addressed. In this work, we focus on the development of a method that deals with the nonlinear and non-convex nature of kinetic identification in biological process modeling. In previous work,\textsuperscript{20} a deterministic optimization method was found well suited to estimate parameters in a simple model for biological nitrite oxidation. This optimization method led to globally optimal parameter estimates. The same study demonstrated that a
standard approach based on gradient-based optimization fails to find good parameter estimates. Unfortunately, deterministic global optimization is cumbersome when the number of parameters is large.

To deal with the model structure selection and parameter estimation challenge, we adopt an extent-based framework\textsuperscript{21-24} to enable the application of deterministic optimization methods to biological process models involving multiple reactions. The concept of extents allows transforming multivariate time series into a set of individual time series, each one reflecting the progress of a single reaction. This, in turn, enables the individual identification of the rate law and the corresponding parameters for each of the biological reactions. In addition, the use of extents facilitates model diagnosis. The proposed extent-based modeling methodology is demonstrated and benchmarked against a conventional approach by means of a simulated experiment with a urine nitrification process model.\textsuperscript{25} All symbols used in this text are listed in Table 1.

Other factors complicating model identification include (i) the stochastic nature of environmental processes and (ii) significant lack of identifiability of model structures and parameters, further leading to significant uncertainty and correlated parameter estimates. These issues are certainly important but not studied in this work. Instead, we focus on solving model identification problems to global optimality given experimental data. This also means that we assume that a proper experimental design has been executed.

**Methods**

**Definitions**

**Species and components**

Consider a batch reactor with volume $V$ containing $S$ chemical species involved in $R$ reactions. The numbers of moles are given as the $S \times 1$ vector $\mathbf{n}$. Among the $R$ reactions,
Reactions are kinetically controlled, and reactions are considered to be at equilibrium, with \( R = R_k + R_e \). The species are split into kinetic species that are only involved in kinetically controlled reactions (i.e., not in equilibrium reactions) and equilibrium species that are involved in equilibrium (and possibly also in kinetically controlled) reactions \( (S = S_k + S_e) \). The corresponding numbers of moles are \( n_k \) and \( n_e \). Equilibrium components are defined as the molecular constituents that are involved in equilibrium reactions and whose concentrations are conserved. The \( S = S_k + S_e \) numbers of moles of the kinetic species \( n_k \) and the equilibrium components \( n_e \) can be written as:

\[
\mathbf{n} = \begin{bmatrix} n_k \\ n_e \end{bmatrix} = \mathbf{E} \mathbf{n}
\]

where \( \mathbf{E} \) of dimension \( S \times S \) relates the numbers of moles of all species \( \mathbf{n} \) to those of the kinetic species and equilibrium components \( \mathbf{n} \).

**Example.** Let us illustrate the notations through a simplified urine nitrification process model that is used in this work to simulate experimental data. This process model is selected because it is an excellent example of a biological process model based on the activated sludge model family and involving rate-controlling acid-base equilibria. There are \( R = 6 \) reactions involving \( S = 10 \) species dissolved in water. The kinetically controlled reactions are the biological nitritation and nitratation by ammonia oxidizing bacteria (AOB) and nitrite oxidizing bacteria (NOB), respectively, that is, \( R_k = 2 \):

\[
\begin{align*}
\text{NH}_3 + \frac{3}{2} \text{O}_2 & \xrightarrow{\text{AOB}} \text{HNO}_2 + \text{H}_2\text{O} \\
\text{HNO}_2 + \frac{1}{2} \text{O}_2 & \xrightarrow{\text{NOB}} \text{HNO}_3
\end{align*}
\]
The remaining reactions consist of $R_e = 4$ instantaneous acid-base equilibrium reactions:

\[
\begin{align*}

\text{NH}_4^+ & \rightleftharpoons H^+ + \text{NH}_3 \\
\text{HNO}_2 & \rightleftharpoons H^+ + \text{NO}_2^- \\
\text{HNO}_3 & \rightleftharpoons H^+ + \text{NO}_3^- \\
\text{H}_2\text{O} & \rightleftharpoons H^+ + \text{OH}^- \\
\end{align*}
\]

The net growth of bacteria is assumed negligible. According to this reaction scheme, the $S = 10$ species are oxygen, ammonium, ammonia, nitrous acid, nitrite, nitric acid, nitrate, proton ions, hydroxyl ions, and water. Oxygen is only involved in the kinetically controlled reactions ($S_k = 1$). The remaining species are equilibrium species ($S_e = 9$). The numbers of moles are computed from the concentrations as follows:

\[
\mathbf{n} = V \begin{bmatrix} [O_2] & [\text{NH}_4^+] & [\text{NH}_3] & [\text{HNO}_2] & [\text{NO}_2^-] & [\text{HNO}_3] & [\text{NO}_3^-] & [H^+] & [\text{OH}^-] & [\text{H}_2\text{O}] \end{bmatrix}^T
\]

\[
= \begin{bmatrix} \mathbf{n}_k \\ \mathbf{n}_e \end{bmatrix} = V \begin{bmatrix} [O_2] \\ [\text{NH}_3] \\ [\text{HNO}_2] \\ [\text{NO}_2^-] \\ [\text{NO}_3^-] \\ [H^+] \\ [\text{OH}^-] \\ [\text{H}_2\text{O}] \end{bmatrix}.
\]

(2)

with $V$ the volume.

The $S_e = 5$ molecular constituents that are conserved in the equilibrium reactions are total ammonia (total ammonia nitrogen, TAN), total nitrite (TNO2), total nitrate (TNO3), total proton (TH), and total hydroxyl (TOH). With $S_k = 1$ (oxygen), the $6 \times 10$ matrix $\mathbf{E}$ reads:
Balance equations

For a batch reactor, the differential mole balance equations are written as

$$\dot{n}(t) = V N^T r(n(t)/V), \quad n(0) = n_0$$  \hspace{1cm} (4)

with $N$ the $R \times S$ stoichiometric matrix, $V$ the volume (assumed to be constant), $r$ the $R \times 1$ reaction rates, and $n_0$ the $S \times 1$ initial numbers of moles. Upon pre-multiplying (4) by $\bar{E}$, one obtains:

$$\dot{n}(t) = \bar{E} \dot{n}(t) = V \bar{E} N^T r(n(t)/V) = V N^T r(n(t)/V), \quad \dot{n}(0) = \bar{E} n_0$$  \hspace{1cm} (5)

with $\bar{N}$ the corresponding stoichiometric matrix of dimension $R \times \bar{S}$ and $\bar{n}_0 = \bar{E} n_0$. Given $\bar{n}$, the vector $n$ is obtained by solving the following system of $S = S + R_e$ algebraic equations:\(^26\)
\[
\bar{E} \mathbf{n}(t) = \bar{\mathbf{n}}(t) \quad (6)
\]
\[
g(\mathbf{n}(t)/V) = 0_{Re} \quad (7)
\]

where \( g(\cdot) \) expresses the \( Re \) instantaneous equilibria. The dynamics of the component concentrations are functions of the kinetically controlled reactions only, that is, the rows of \( \mathbf{N} \) corresponding to the equilibrium reactions contain only zeros.\(^{26}\) Hence, a reduced stoichiometric matrix \( \mathbf{N}_k \) can be defined as the matrix consisting of the rows of \( \mathbf{N} \) with at least one non-zero element. Following this, (5) reduces to:

\[
\dot{\mathbf{n}}(t) = V \mathbf{N}_k^T \mathbf{r}_k(\mathbf{n}(t)/V), \quad \mathbf{n}(0) = \mathbf{n}_0 \quad (8)
\]

with \( \mathbf{r}_k \) the kinetically controlled reaction rates.

**Example.** Following aforementioned definitions, the stoichiometric matrix for all reactions is

\[
\mathbf{N} = \begin{bmatrix}
-3/2 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
-1/2 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\
\end{bmatrix} \quad (9)
\]

with the reduced stoichiometric matrix
\[ N_k = \begin{bmatrix} -3/2 & -1 & 1 & 0 & 2 & 1 \\ -1/2 & 0 & -1 & 1 & 0 & 0 \end{bmatrix} \] (10)

describing the \( R_k = 2 \) kinetically controlled reactions in terms of the \( S_k = 1 \) kinetic species and the \( S_c = 5 \) equilibrium components.

The rate laws for the biological oxidation reactions are:

\[
\mathbf{r}_k = \begin{bmatrix} r_{\text{AOB}} \\ r_{\text{NOB}} \end{bmatrix} = \begin{bmatrix} [\text{NH}_3] / \left( \theta_{\text{AOB},1} + \theta_{\text{AOB},2} [\text{NH}_3] + \theta_{\text{AOB},3} [\text{NH}_3]^2 \right) \\ [\text{HNO}_2] / \left( \theta_{\text{NOB},1} + \theta_{\text{NOB},2} [\text{HNO}_2] \right) \end{bmatrix} \] (11)

with the kinetic parameters \( \theta_{\text{AOB},1}, \theta_{\text{AOB},2}, \theta_{\text{AOB},3}, \theta_{\text{NOB},1}, \) and \( \theta_{\text{NOB},2} \). The time dependence of rates and concentrations is omitted for the sake of conciseness. The two kinetic expressions correspond to Haldane and Monod kinetics, respectively. Since we assume that oxygen is sufficient for both oxidation processes, rate-limiting effects of oxygen can be safely ignored. The balance equations describing the equilibria cover four acid-base reactions so that (7) is

\[
g(\mathbf{n}/V) = \begin{bmatrix} \left( [\text{H}^+] [\text{NH}_3] \right)/[\text{NH}_4^+] - 10^{-pK_{a, \text{NH}_4^+}} \\ \left( [\text{H}^+] [\text{NO}_2^-] \right)/[\text{HNO}_2] - 10^{-pK_{a, \text{HNO}_2}} \\ \left( [\text{H}^+] [\text{NO}_3^-] \right)/[\text{HNO}_3] - 10^{-pK_{a, \text{HNO}_3}} \\ [\text{H}^+] [\text{OH}^-] - 10^{-pK_w} \end{bmatrix} = \mathbf{0}_4. \] (12)

The initial numbers of moles are \( \mathbf{n}_{k,0} = V \begin{bmatrix} c_{\text{O}_2,0} & c_{\text{TAN},0} & 0 & 0 & c_{\text{TH},0} & c_{\text{TOH},0} \end{bmatrix}^T \). As there is no liquid entering or leaving the reactor during the reaction, the simulated data correspond to a typical batch test \( 27-30 \) with a single pulse of ammonia dosed at the start.
of the experiment. The initial concentrations of proton and hydroxyl component \((c_{\text{TH},0} \quad \text{and} \quad c_{\text{TOH},0})\) are set to values that satisfy the equilibrium equations and deliver a zero ion balance. The initial oxygen and water concentrations can be set arbitrarily and do not affect the reaction rates nor the equilibria.

**Measurement equations**

During the simulated batch experiment, \(M\) measurements are obtained at \(H\) distinct time instants \(t_h\), with \(h = 1, \ldots, H\) and \(t_1 = 0\), as:

\[
\tilde{y}(t_h) = y(t_h) + \epsilon(t_h) = f(n(t_h)/V) + \epsilon(t_h), \quad \epsilon(t_h) \sim \mathcal{N}(0, \Sigma_h) \tag{13}
\]

with \(\tilde{y}(t_h)\) the \(M\)-dimensional vector of measurements, and \(y(t_h)\) the noise-free measured variables at time \(t_h\). In words, the measurements are nonlinear functions of the species concentrations and are subject to additive Gaussian noise. The functions \(f(\cdot)\) are assumed continuous and differentiable.

**Example.** Measurements of the total ammonia, total nitrite and total nitrate concentrations and of pH are obtained. The noise-free measurements (13) are given as:

\[
y = \begin{bmatrix}
y_{\text{TAN}} \\
y_{\text{TNO}_2} \\
y_{\text{TNO}_3} \\
y_{pH}
\end{bmatrix} = \begin{bmatrix}
[\text{NH}_4^+] + [\text{NH}_3] \\
[\text{HNO}_2] + [\text{NO}_2^-] \\
[\text{HNO}_3] + [\text{NO}_3^-] \\
-\log_{10} ([\text{H}^+])
\end{bmatrix} = \begin{bmatrix}
G n/V \\
-\log_{10} ([\text{H}^+])
\end{bmatrix} \tag{14}
\]

where \(G\) is the measurement matrix.
Clearly, the first three measurements are linear in the species concentrations. In contrast, the pH measurement depends nonlinearly on the proton concentration, which in turn depends nonlinearly on the component concentrations via the algebraic equilibrium relationships in (7). In our simulated experiment, the measurement error variance-covariance matrix is considered time-invariant and defined as follows:

\[
\Sigma_h = \Sigma = \text{diag}\left(\begin{bmatrix} \sigma_{T_{AN}}^2 & \sigma_{T_{NO2}}^2 & \sigma_{T_{NO3}}^2 & \sigma_{pH}^2 \end{bmatrix}^T\right), \quad h = 1, \ldots, H
\]  

where the \text{diag}(\cdot) operator creates a diagonal matrix from a column vector argument.

Data generation and problem formulation

Data generation

The main objective of this paper is to compare a new method for model identification to a more conventional approach. To properly compare the two methods, simulated measurements are used. These measurements are obtained by solving the DAE system (6)-(8) with (3) and (10)-(12) from \(t_1 = 0\) to \(t_H = 10\ h\). Measurements are obtained by means of (13)-(16) at regular intervals of 10 minutes so that \(H = 61\). All parameter values used for simulation are given in Table 1.
Problem formulation

The model identification problem consists of finding an appropriate model based on the measurements from a pulse experiment. For each kinetically controlled reaction, a set of five candidate rate laws are proposed. These are the zeroth-order, first-order, Monod, Tessier, and Haldane rate laws given in Table 2. The initial conditions $\mathbf{n}_0$, the stoichiometric matrix $\mathbf{N}$, the equilibrium equations $\mathbf{g}(\cdot)$, the measurement equations $\mathbf{f}(\cdot)$, and the measurement error variance-covariance matrices $\mathbf{H}_h$ are assumed to be known. The aim is therefore to identify which of the considered rate laws are appropriate for the two reactions, while also estimating the corresponding kinetic parameters. In this work, feasible values for the parameters are considered to be in the interval $[10^{-6}, 10^2]$.

Notation. The $j^{th}$ candidate rate law for the $i^{th}$ kinetically controlled reaction is referred to as $r_{k,i}^{(j)}$. The corresponding parameter vectors are $\mathbf{\theta}_i^{(j)}$. The number of candidate rate laws for the $i^{th}$ reaction is $J_i$, so that $j = 1, 2, \ldots, J_i$. For a given choice of rate laws for each kinetically controlled reaction, the parameter vector composed of the joint set of parameter vectors for all reactions is denoted as $\Theta$.

Method 1: Simultaneous model identification

The simultaneous model identification procedure is an exhaustive method that consists of building a model for every possible combination of the considered rate laws ($r_k$) followed by the estimation of all kinetic parameters ($\Theta$) for each model. As indicated above, we assume that the stoichiometry and equilibrium relations are known and the rate laws and their parameters are to be identified. For a given selection of candidate rate laws, parameter estimation is formulated mathematically as the following weighted least squares (WLS) estimation problem:
\[ \hat{\Theta} = \arg \min_{\Theta} \sum_{h=1}^{H} (\tilde{y}(t_h) - y(t_h))^T \Sigma_h^{-1} (\tilde{y}(t_h) - y(t_h)) \]  

(17)

subject to \[ y(t_h) = f(n(t_h)/V) \]  

(18)

\[ g(n(t)/V) = 0_{R_w} \]  

(19)

\[ \tilde{n}(t) = \bar{n}(t) \]  

(20)

\[ \bar{n}(t) = V \int_0^t N_k^T r_k(n(\tau)/V, \Theta) d\tau, \quad \bar{n}(0) = \bar{n}_0 \]  

(21)

\[ \Theta = \begin{bmatrix} \theta_1^{(j)^T}, \ldots, \theta_i^{(j)^T}, \ldots, \theta_R^{(j)^T} \end{bmatrix}^T \]  

(22)

During this estimation, the simulated system (18)-(22) is the same as the data-generating process, except for the rate laws in \( r_k \) and the parameters therein. Because the measurement errors are assumed to be normally distributed according to (13), minimizing the WLS objective corresponds to a maximum-likelihood estimation (MLE).

The optimization problem (17)-(22) is solved by means of the Nelder-Mead simplex algorithm.\(^{31}\) This algorithm is initiated with parameter values at the center of the feasible intervals considered above. The total number of models whose parameters need to be estimated equal the product of the numbers of candidate rate laws, \( \prod_i J_i \). Following the parameter estimation for each of these models, a well-fitting model is selected from the complete set of models by trading off the WLS objective (17) against parsimony. To this end, the WLS objective is equivalently expressed as the weighted root mean squared residual (WRMSR):

\[ WRMSR = \sqrt{\frac{1}{H \cdot M} \sum_{h=1}^{H} (\tilde{y}(t_h) - y(t_h))^T \Sigma_h^{-1} (\tilde{y}(t_h) - y(t_h))} \]  

(23)

Example. To model the simulated process, five different candidate rate laws are consid-
Method 2: Incremental model identification via extents

This subsection introduces the concept of extents of reaction and shows how to compute them from the measured numbers of moles. The computed extents are then used to identify the kinetics of each reaction individually, thereby making the procedure incremental. Finally, the same measurements are used to fine-tune the kinetic parameters for the global model.

Definition of extents

In batch reactors, the extents of reaction $x(t)$ can be defined by means of the following integral:

$$\mathbf{n}(t) = \mathbf{n}_0 + V \int_0^t \mathbf{N}^T \mathbf{r} \left( \mathbf{n}(\tau)/V \right) d\tau = \mathbf{n}_0 + \mathbf{N}^T \mathbf{x}(t). \quad (24)$$

In words, an extent of reaction expresses the progress of the corresponding reaction in terms of the numbers of moles of the product it has produced since $t = 0$. This definition can be applied to multiphase systems as well. In what follows, unless mentioned otherwise, the term extent refers specifically to the extent of a kinetically controlled reaction. Equation (8) can be integrated to give:

$$\mathbf{n}(t) = \mathbf{n}_0 + V \int_0^t \mathbf{N}_k^T \mathbf{r}_k \left( \mathbf{n}(\tau)/V \right) d\tau = \mathbf{n}_0 + \mathbf{N}_k^T \mathbf{x}_k(t). \quad (25)$$

Reformulating the balance equations (21)-(20) in terms of extents allows selecting rate laws and estimating parameters for each reaction individually. To do so, the available measurements are first transformed into experimental extents. After this transformation, and

\[\text{ered for each of the two biological reactions } (J_1 = J_2 = 5). \text{ The number of distinct models}\]

whose parameters are estimated is therefore $\prod_i J_i = J_1 \cdot J_2 = 5 \cdot 5 = 25$.\]
for each reaction individually, selected rate laws can be fitted to the experimental extents. These steps are explained next.

Step 1 – Computation of experimental extents

The extents of reaction for the kinetically controlled reactions can be computed by solving the following WLS problem for each sampling instant $t_h$:

$$
\hat{x}_k(t_h) = \arg \min_{x_k(t_h)} (\hat{y}(t_h) - y(t_h))^T \Sigma_h^{-1} (\hat{y}(t_h) - y(t_h))
$$

s.t. $y(t_h) = f(n(t_h)/V)$

$$
(26)
$$

$$
g(n(t_h)/V) = 0_{Re}
$$

$$
E n(t_h) = n_0 + N_T^T x_k(t_h)
$$

where (26) is the objective function, (27) expresses the expected measurements as functions of the numbers of moles of the species, (28) expresses the algebraic equilibria, and (29) relates the extents of the kinetically controlled reactions to the number of moles of the species. As above, minimizing the WLS objective to its global minimum corresponds to MLE. In general, the above problem is nonlinear, non-convex, and therefore solved numerically. In special cases, an analytic solution can be provided.26

The initial numbers of moles $n_0$, and volume $V$ are assumed to be known. Hence, one can compute the point-wise approximation $\Lambda_h$ to the variance-covariance matrix of the experimental extents as the inverse of the Fisher information matrix $J_h^T \Sigma_h^{-1} J_h$, where $J_h$ is the Jacobian matrix, with $J_h(m, i) = \frac{\partial y_m}{\partial x_{k,i}}|_{\hat{x}_k(t_h)}$. The elements of $J_h$ are computed by numerical differentiation unless analytical derivatives are available. This procedure allows writing the following approximate distribution for the extent estimation errors, that is, the difference between the experimental extents $\hat{x}_k$ and the true extents $x_k$:
Example. In the simulated experiment, it follows from (1), (14), and (25) that the Jacobian consists of three rows that are computed analytically and a fourth row that is evaluated numerically:

\[ \mathbf{J}_h = \begin{bmatrix} \frac{1}{\mathcal{V}} \mathbf{G E}^+ \mathbf{N}_k^T \\
\frac{\partial pH}{\partial \bar{x}_k} \bigg|_{\bar{x}_k(t_h)} \end{bmatrix} \]  

(31)

with the superscript \((\cdot)^+\) indicating the Moore-Penrose pseudo-inverse. The \(i\)th element of the last row is computed as \(\frac{pH_1 - pH_0}{\delta}\) with \(pH_0\) and \(pH_1\) the pH values obtained by solving (27)-(29) at \(\bar{x}_k(t_h)\) and \(\bar{x}_k(t_h) + \Delta_i\), with \(\Delta_i\) a vector with the small number \(\delta\) in its \(i\)th position and zeros elsewhere.

Step 2 – Extent modeling

The original identification problem (17)-(22) is now simplified by fitting the rate laws to the experimental extents instead of to the original measurements and by estimating the parameters for each reaction individually. The idea is to model each reaction by optimizing the fit to the corresponding experimental extent, \(\bar{x}_{k,i}\), with \(i = 1, \ldots, R_k\). However, since the reaction rate \(r_{k,i}\) is a function of concentrations that might depend on the progress of several reactions, one estimates the contribution of the other reactions from measurements.\(^{21}\)

This results in the following optimization problem for the \(j\)th candidate rate law for the \(i\)th kinetically controlled reaction:
\[ \hat{\theta}^{(j)}_i = \arg \min_{\theta^{(j)}_i} \text{ssq}_i := \sum_{h=1}^{H} \frac{(\bar{x}_{k,i}(t_h) - x_{k,i}(t_h))^2}{\lambda_{i,h}} \]  

s.t. \[ g(n(t)/V) = 0 \]  

\[ \bar{E}n(t) = \bar{n}_0 + \bar{N}_k^r x_k(t) \]  

\( \forall r = 1, \ldots, R_k : \)  

\[ x_{k,r}(t) = \begin{cases} 
V \int_{0}^{t} r_{k,i}^{(j)} \left( \frac{n(t)}{V}, \theta^{(j)}_i \right) \, dt, & x_{k,i}(0) = 0 \quad \text{if } r = i \\
I(t, \bar{x}_{k,r}, t), & \text{if } r \neq i 
\end{cases} \]  

where \( \lambda_{i,h} := \Lambda_h(i, i) \), \( t = [t_1, t_2, \ldots, t_h, \ldots, t_H] \) and with the operator \( I(\cdot) \) defined as

\[ \forall t \in \{ t : t_i \leq t \leq t_{i+1} \} : \quad I(t, \bar{x}_{k,r}, t) := \bar{x}_{k,r}(t_i) + \left( \bar{x}_{k,r}(t_{i+1}) - \bar{x}_{k,r}(t_i) \right) \frac{t - t_i}{t_{i+1} - t_i}. \]  

In the above problem, (32) is the objective function expressing that the \( i \)th predicted extent should be as close as possible to the corresponding experimental extents in the WLS sense. As before, (33) and (34) express the algebraic equilibria and the relationships between the extents of the kinetically controlled reactions and the number of moles of all species. Equation (35) indicates that the predicted extents stem from (i) the simulated \( i \)th reaction, and (ii) piecewise-linear interpolation of the experimental extents for the other kinetically controlled reactions. The most important consequence of this method is that only the kinetic parameters of the \( i \)th candidate rate law appear in the optimization problem. Indeed, the interpolation of the experimental extents (36) implies that the kinetic parameters of the corresponding reactions are not needed. The original optimization problem is thereby replaced by multiple optimization problems involving a univariate system including only one reaction. Furthermore, the modification also means that one does not need to know the structure of the rate laws corresponding to the interpolated experimental extents, that
is, the best candidate rate law for each reaction can be found independently of the rate laws for the other reactions.

The second method allows solving each individual parameter estimation problem to global optimality by means of the branch-and-bound algorithm proposed earlier. This way, the best parameter values are guaranteed to be found within the considered feasible intervals. The bounding procedures required for this algorithm are given in the Supporting Information.

With each candidate rate law and the associated optimal parameters $\hat{\theta}_i^{(j)}$, one obtains the modeled extent $\hat{x}_{k,i}^{(j)}$ and the following extent-specific WRMSR:

$$WRMSR_i^{(j)} = \sqrt{\frac{1}{H} \sum_{h=1}^{H} \left( \frac{\hat{x}_{k,i}(t_h) - \hat{x}_{k,i}^{(j)}(t_h)}{\lambda_{i,h}} \right)^2}.$$  \hspace{1cm} (37)

The rate law $\hat{r}_{k,i}$ is selected by trading off the WRMSR against parsimony. This is repeated for every reaction, which means that the number of parameter estimation problems to be solved now equals the sum of the numbers of candidate rate laws, $\sum_i J_i$. In addition, the number of parameters that are estimated in each problem is generally lower than the number of parameters estimated with the first method (17)-(22).

**Example.** With 5 candidate rate laws considered for each reaction, $\sum_i J_i = J_1 + J_2 = 10$ instances of the parameter estimation problem need to be solved. The number of parameters that are estimated in each problem ranges from 1 (e.g. zeroth-order rate law) to 3 (Haldane). In comparison, the first method requires the estimation of 2 (zeroth-order rate law for both reactions) up to 6 (Haldane rate law for both reactions) parameters at once.

**Step 3 – Model fine-tuning**

Following the rate-law selection, the model parameters are fine-tuned by simultaneously estimating all kinetic parameters via (17)-(22). As in the first method, this is done using
the Nelder-Mead simplex algorithm. In contrast to the first method, this algorithm is now executed for only one model containing the rate laws selected in Step 2 and is initiated with the corresponding parameter estimates obtained in Step 2.

**Results**

**Process simulation**

The nitrification model (6)-(15) is used to generate concentration and pH measurement series. The results are shown in Fig. 1. One can see a fairly distinct separation in time of the two reactions with the TNO2 concentration rising to 50% of the original TAN concentration at about 4 h. Before (after) this time, a net production (consumption) of TNO2 is observed. The figure also shows the free ammonia concentration [NH\textsubscript{3}]. The ammonia oxidation stops when this concentration reaches zero. The nitrite and nitrate ion concentrations are nearly indistinguishable from the total nitrite and total nitrate concentrations (not shown). At the end of the experiment, about half of the available TAN is converted via nitrite to nitrate. The limited buffering capacity in the simulated system causes fairly large changes in pH. Additive Gaussian noise is simulated to obtain the measurements in regularly-spaced time intervals according to the description above.

**Method 1: Simultaneous model identification**

The kinetic parameters of 25 different models, each with a unique pair of rate laws for the first and second reactions, are estimated by solving (17)-(22). The resulting WRMSR values shown in Fig. 2 range from 6.57 to 37.86. These values indicate the model prediction error standard deviation relative to the measurement error standard deviation. Assuming the correct model, it exhibits a $\chi^2$-distribution with a mean value of 1 and a right-sided 99% confidence limit of 1.11. The graph also shows the WRMSR value of 1.01 obtained with the noisy measurements and the true model including its parameters. This WRMSR is very
close to the expected value of 1. This value is however 6.57 times lower than the best model obtained with the first model identification method. Clearly, this method is unfit to find a good model. In all cases, including the case involving the true rate laws used for simulation, only a locally optimal parameter set could be found. In addition, the best model (Model 7) includes the first-order rate law for both reactions, which does not correspond to the true rate laws. Additional results, including simulations by each of the 25 models after parameter estimation, are included in the Supporting Information.

Method 2: Incremental model identification via extents

**Step 1 – Extent computation**

The extents computed by solving (26)-(29) using the TAN, TNO2, TNO3, and pH measurements are shown in Fig. 3(a). The confidence bands for the experimental extents vary with time, in particular for the first extent. High precision is obtained at the beginning and during most of the second half of the experiment. However, during the first half, the uncertainty first increases and then decreases. At the end of the experiment, the uncertainty increases again. These effects are due to the nonlinear propagation of the pH measurement error through the measurement and algebraic equations. The ellipsoidal confidence regions at 0.5, 1.5, 2.5, 3.5, and 4.5 h are shown in Fig. 3(b). The orientation of the confidence region becomes more oblique with increased uncertainty in the first extent.

**Step 2 – Extent modeling**

The global solutions to (32)-(35), obtained for every reaction and every candidate rate law, are discussed next.

**Modeling the first extent.** The best fits of the first extent obtained with the various candidate rate laws are shown in Fig. 4(a). It is clear that the zeroth- and first-order models do not fit the experimental extents well. The Monod and Tessier models fit better, yet they
over-estimate the experimental extent. This is clearly visible in Fig. 4(b-c), where the model errors are shown. In contrast, the Haldane rate law fits the extent profile well. As such, the Haldane model is easily selected as the best among the model candidates. In Fig. 5, the WRMSR values (37) are given with a 95% upper control limit based on the corresponding $\chi^2$-statistics. Based on this statistic, all models except for the Haldane model are rejected for the first extent.

**Modeling the second extent.** The best fits of the second extent are visualized in Fig. 6(a). Here, all rate models fit the experimental extents reasonably well, except for the zeroth-order model. The first-order model leads to visibly auto-correlated residuals Fig. 6(b-c). This is also evident from the WRMSR values (Fig. 5), on the basis of which the zeroth-order and first-order models are rejected. In this case, a parsimonious model is chosen among the three remaining candidates. The Monod model delivers the best fit among the simplest candidates (Monod and Tessier). An alternative approach may consist in designing an experiment that enables better discrimination of the remaining rate laws. This is not explored in this work.

**Step 3 – Model fine-tuning**

The model structure consisting of the two selected rate laws, namely, Haldane and Monod, is used next to fine-tune the model parameter via the simultaneous approach (17)-(22). Fig. 7 compares the simulated concentration and pH values with the predictions of the identified models prior and after fine-tuning. These three simulations are hard to distinguish from each other. The resulting overall WRMSR (23) equals 1.0013 and is shown in Fig. 2. Most importantly, the extent-based model identification procedure has delivered a well-fitting set of rate laws and kinetic parameter estimates. Furthermore, the selected rate laws are exactly those used to generate the simulated experimental measurements. The parameter estimates deviate at most 10% from their true values, except $\theta_{AOB,2}$ which deviates with 30%. Such
deviations are typical for biokinetic wastewater treatment models and are in part explained by correlation between parameter estimates.

Discussion

The results presented above are now interpreted in a broader biokinetic modeling context.

Interpretation of the results. In this study, the concept of extents is introduced for the first time for the purpose of dynamic modeling of an environmental biochemical process. By means of a simplified biokinetic model of the urine nitrification process and simulated batch experiments, several benefits of the extent-based modeling approach have been demonstrated. Concretely, the identification of biokinetic models via extents:

- Allows using deterministic optimization methods to obtain excellent parameter estimates. Despite the fact that the individual extent modeling steps only approximate the original model identification problem, one can obtain a well-fitting model. Most importantly, the convergence to local optima as observed with a conventional parameter estimation method can be avoided.

- Provides an intuitive diagnostic tool for modeling. Indeed, extent-modeling indicates whether a reaction can be modeled appropriately with a given candidate rate law, thus allowing modelers to pay more attention to reactions that are more difficult to model. Similarly, this approach indicates whether sufficient information is available within a given experimental data set to discriminate between candidate rate laws.

- Reduces a model selection problem that is polynomial in the number of candidate rate laws into a model selection problem that is linear in this number. In this study, the extent-based modeling method required solving 10 parameter estimation problems involving 1 to 3 parameters, whereas the conventional simultaneous approach required solving 25 parameter estimation problems involving 2 to 6 parameters.
It is of special importance that the extent-based model identification method is the only method delivering an acceptable model. Indeed, the conventional model identification method did not result in an acceptable model, despite the apparent simplicity of the studied process and simulated experiment.

**Links with prior work.** While the concept of extents is new in the context of dynamic modeling of environmental processes, it is important to note that a number of important concepts in use today are somewhat similar. For instance, the integral defined by the area under the oxygen uptake rate curve, a.k.a. respirogram, is matched to the total accumulated oxygen uptake in typical respirometric experiments.\(^{32}\) Similar concepts include accumulated methane production\(^{33}\) and number of base pulses.\(^{34}\) It is also interesting to note that the accumulated cellulose solubilisation has been described as extent of solubilisation,\(^{35}\) however without links to the general concept of extents. The most important difference between extents and the concepts already in use is that extents reflect individual processes rather than several simultaneous processes. So far, model reduction on the basis of the concept of reaction invariants\(^{36}\) is the only related application known in the environmental engineering sciences. We expect tangible benefits from a broader and systematic use of extents, including those mentioned above.

**Analysis of the extent-based modeling method.** In the general case, the extent-based modeling method does not solve the exact same problem as the conventional simultaneous modeling method. Extent-based modeling solves the same problem (17)-(22) only if (i) the measurement equations are linear and there are nonlinear algebraic equations involved in the extent computations (26)-(29), (ii) the off-diagonal elements of the matrices \( \Lambda_h \) are equal to zero, that is, in absence of correlation between experimental extents, and (iii) the reaction rates can be expressed as functions of the modeled extents. These requirements are rarely satisfied so that the resulting parameter estimates likely deviate from those obtained by solving (17)-(22). However, the extent-based modeling framework is particularly useful
when solving (17)-(22) to global optimality is difficult or computationally prohibitive.

In computing the solution to (26)-(29), one can distinguish different situations:

1. The first situation occurs when the available measurements are linear in the extents of the kinetically controlled reactions and do not depend on the equilibrium species concentrations. In this case, one can discard all nonlinear (equilibrium) equations and an analytic solution for the extents can be found.

2. The second situation occurs when the number of measured variables matches the number of computed extents exactly, thereby resulting in a fully determined system (hence no need for optimization). In this situation, one can find extents that make the objective function (26) equal to zero, while satisfying (27)-(29). The solution can therefore be obtained by solving the equation system (27)-(29) numerically. In the process considered in this work, this situation would occur if the pH and one of the remaining variables (TAN, TNO2, TNO3) were measured (not demonstrated).

3. The third situation occurs when the number of measured variables exceeds the number of computed extents (overdetermined system). This corresponds to the case studied in this work (TAN, TNO2, TNO3, and pH measured). One approach consists of discarding (26) and solving (27)-(29) in a least-squares sense. When doing so, the experimental extents are not the solution to (26)-(29). We recommend solving (26)-(29) exactly, as in this work, to obtain experimental extents that are WLS-optimal.

In its current form, the provided extent-based modeling method assumes a closed batch process whose stoichiometric matrix and the algebraic equilibrium equations are known or estimated precisely. However, this is not true in general. The method presented here can easily be expanded to account for mass transfer as well as gas-liquid transfer as demonstrated already. The main reason this has not been included here is to maintain a clear presentation of the developed method. Not knowing the stoichiometric matrix or the algebraic equilibrium equations means that, prior to modeling via extents, one may use target factor
analysis\textsuperscript{38} to identify the stoichiometric matrix and/or detailed physico-chemical analysis to obtain a model for acid-base and salt speciation. However, certain situations allow using the extent-based modeling framework to estimate equilibrium parameters\textsuperscript{26} as well as stoichiometric parameters.\textsuperscript{39} Even more critical is the fact that extent-based model identification requires at least as many measured variables as there are kinetically controlled reactions. When this requirement is not met, one can opt to partition the model identification problem into smaller problems which include more than one reaction.\textsuperscript{39}

**Methodological improvements.** Methodologically speaking, this work adds four elements to the extent-based modeling framework, namely:

- Extent computation with measurements that are nonlinear in the species concentrations.
- Optimal estimation of the experimental extents when (27)-(29) involves more measured variables than extents, that is, in the overdetermined case.
- Accounting for nonlinear effects during experimental extent computation by means of a Laplacian approximation of their distribution.
- Extent-based modeling and deterministic global optimization are combined for the first time into a single model identification framework.

**Future work.** The developments in this study are considered critical steps towards a first real-world application of the extent-based modeling of environmental processes. However, the following aspects call for further development and testing of the method prior to experimental validation in full-scale wastewater treatment systems:

- Realistic sensor data. So far, measurement devices are considered to exhibit an instantaneous response within the extent-based modeling framework. However, typical devices respond dynamically to the measured variable.\textsuperscript{40,41} An explicit account of sensor dynamics is not provided yet in the extent-based modeling framework.
Laboratory validation. Several elements of realistic biological process have been ignored to facilitate the introduction of extent-based modeling. The ignored elements include (i) bacterial growth and decay processes, (ii) complex composition of actual wastewater, and (iii) complex physico-chemical reaction systems in high-strength wastewater such as source-separated urine. The first element only affects the extent-based methodology due to a lack of extent observability. This can be accounted for in special cases but may prove difficult to address in general. The second and third element affect both modeling methods used in this study and are being addressed currently by adopting a more realistic physico-chemical urine composition and associated reaction system in view of a lab-scale validation.

Prior knowledge. In this work, the reactor volume $V$, the initial conditions $n_0$, and the stoichiometric matrix $N$ are considered known. Methods permitting the estimation of these variables and parameters remain to be investigated.

Completeness of the candidate rate laws. In this work, we have assumed that the set of candidate rate laws includes the true rate laws in the data-generating process. This is not true in general. An alternative model structure based on shape constrained splines can address this problem. So far, this type of models has only been applied to monoculture processes. Its use in connection with the extent-based model identification remains to be evaluated.

Acknowledgement

This study was made possible by Eawag Discretionary Funds (grant no.: 5221.00492.009.03, project: DF2015/EMISSUN). All results were obtained by use of Matlab and the Spike_O toolbox for optimization.
Table 1: List of symbols and parameter values used for simulation. Values in parentheses refer to the best-available estimates.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_s$</td>
<td>Substrate concentration</td>
<td>–</td>
<td>mol·L$^{-1}$</td>
</tr>
<tr>
<td>$c_{\text{TAN},0}$</td>
<td>Initial TAN concentration</td>
<td>0.35</td>
<td>mol·L$^{-1}$</td>
</tr>
<tr>
<td>$c_{\text{TH},0}$, $c_{\text{TOH},0}$</td>
<td>Initial concentrations</td>
<td>–</td>
<td>mol·L$^{-1}$</td>
</tr>
<tr>
<td>$c_{\text{O}_2,0}$</td>
<td></td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\bar{E}$</td>
<td>Matrix defining the set of kinetic species and conserved molecular constituents</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$f(\cdot)$</td>
<td>Measurement expressions</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$G$</td>
<td>Measurement gain matrix</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$g(\cdot)$</td>
<td>Algebraic equilibrium expressions</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$H$</td>
<td>Number of samples</td>
<td>61</td>
<td>–</td>
</tr>
<tr>
<td>$h$</td>
<td>Measurement sample index</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$i$</td>
<td>Reaction index</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$J_h$</td>
<td>Jacobian matrix at the $h$th sample time</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$J_i$</td>
<td>Number of rate law candidates for the $i$th reaction</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$j$</td>
<td>Rate law candidate index</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$M$</td>
<td>Number of measurements</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$m$</td>
<td>Measured variable index</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$N$, $\bar{N}$, $N_k$</td>
<td>Stoichiometric matrices</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$n$</td>
<td>Numbers of moles of all species</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$\bar{n}$</td>
<td>Numbers of moles of kinetic species and conserved molecular constituents</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$n_0$, $n_{k,0}$, $n_0$</td>
<td>Initial numbers of moles</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$n_c$</td>
<td>Numbers of moles of conserved molecular constituents</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$n_e$</td>
<td>Numbers of moles of equilibrium species</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$n_k$</td>
<td>Numbers of moles of kinetic species</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$pK_{a, \text{NH}_4^+}$</td>
<td>Logarithmic acid dissociation constant of NH$_4^+$</td>
<td>+9.24</td>
<td>–</td>
</tr>
<tr>
<td>$pK_{a, \text{HNO}_2}$</td>
<td>Logarithmic acid dissociation constant of HNO$_2$</td>
<td>+3.25</td>
<td>–</td>
</tr>
<tr>
<td>$pK_{a, \text{HNO}_3}$</td>
<td>Logarithmic acid dissociation constant of HNO$_3$</td>
<td>−1.40</td>
<td>–</td>
</tr>
</tbody>
</table>
Table 1: List of symbols and parameter values used for simulation. Values in parentheses refer to the best-available estimates.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$pK_w$</td>
<td>Logarithmic water dissociation constant</td>
<td>14</td>
<td>–</td>
</tr>
<tr>
<td>$R$</td>
<td>Number of reactions</td>
<td>6</td>
<td>–</td>
</tr>
<tr>
<td>$R_e$</td>
<td>Number of equilibrium reactions</td>
<td>4</td>
<td>–</td>
</tr>
<tr>
<td>$R_k$</td>
<td>Number of kinetic reactions</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>$r$</td>
<td>Reaction index</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$r_k$</td>
<td>Reaction rates</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$r_k (r_{k,i})$</td>
<td>Reaction rates of the kinetically controlled reactions (of the $i$th kinetically controlled reaction)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\hat{r}_{k,i}$</td>
<td>Selected rate law for the $i$th kinetically controlled reaction</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$S$</td>
<td>Number of chemical species</td>
<td>10</td>
<td>–</td>
</tr>
<tr>
<td>$\tilde{S}$</td>
<td>Number of kinetic species and conserved molecular constituents</td>
<td>6</td>
<td>–</td>
</tr>
<tr>
<td>$S_c$</td>
<td>Number of conserved molecular constituents</td>
<td>5</td>
<td>–</td>
</tr>
<tr>
<td>$S_e$</td>
<td>Number of equilibrium species</td>
<td>9</td>
<td>–</td>
</tr>
<tr>
<td>$S_k$</td>
<td>Number of kinetic species</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>$t, t_h$</td>
<td>Time (continuous, discrete)</td>
<td>–</td>
<td>h</td>
</tr>
<tr>
<td>$V$</td>
<td>Volume</td>
<td>1</td>
<td>L</td>
</tr>
<tr>
<td>$WRMSR$</td>
<td>Weighted root mean squared residual</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$WRMSR_i^{(j)}$</td>
<td>WRMSR of the $i$th reaction with the $j$th candidate rate law</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$x_k (x_{k,i}, x_{k,r})$</td>
<td>Extents of the kinetically controlled reactions (of the $i/r$th kinetically controlled reaction)</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$\tilde{x}<em>k (\tilde{x}</em>{k,i}, \tilde{x}_{k,r})$</td>
<td>Experimental extents of kinetically controlled reactions (of the $i/r$th kinetically controlled reaction)</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$\tilde{x}_{k,r}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tilde{x}_{k,i}^{(j)}$</td>
<td>Extent estimate for the $i$th kinetically controlled reaction with the $j$th rate law candidate</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$y$</td>
<td>Measured variables</td>
<td>–</td>
<td>mol · L$^{-1}$ / –</td>
</tr>
<tr>
<td>$\tilde{y}$</td>
<td>Measurements</td>
<td>–</td>
<td>mol · L$^{-1}$ / –</td>
</tr>
</tbody>
</table>
Table 1: List of symbols and parameter values used for simulation. Values in parentheses refer to the best-available estimates.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_{TAN}$, $y_{TNO_2}$</td>
<td>Measured variables</td>
<td>–</td>
<td>mol · L$^{-1}$ / –</td>
</tr>
<tr>
<td>$y_{TNO_3}$, $y_{pH}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$ ($\Delta_i$)</td>
<td>Perturbation vector (for the $i$th reaction)</td>
<td>–</td>
<td>mol</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Perturbation parameter</td>
<td>1 · 10$^{-12}$</td>
<td>mol</td>
</tr>
<tr>
<td>$\Theta$</td>
<td>Kinetic parameters for all rate laws</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\hat{\Theta}$</td>
<td>Kinetic parameter estimates for all rate laws</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\theta$ ($\theta^{(j)}_i$)</td>
<td>Kinetic parameters (for the $j$th rate law candidate of $i$th reaction)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\hat{\theta}^{(j)}_i$</td>
<td>Parameter estimates for $j$th kinetic rate law candidate for the $i$th kinetically controlled reaction</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\theta_{AOB,1}$</td>
<td>Kinetic parameter for AOB activity</td>
<td>0.025 (0.024)</td>
<td>d · mol · L$^{-1}$</td>
</tr>
<tr>
<td>$\theta_{AOB,2}$</td>
<td>Kinetic parameter for AOB activity</td>
<td>0.1 (0.13)</td>
<td>d</td>
</tr>
<tr>
<td>$\theta_{AOB,3}$</td>
<td>Kinetic parameter for AOB activity</td>
<td>2.5 (2.4178)</td>
<td>d · mol$^{-1}$ · L</td>
</tr>
<tr>
<td>$\theta_{NOB,1}$</td>
<td>Kinetic parameter for NOB activity</td>
<td>0.11 · 10$^{-3}$ (0.1 · 10$^{-3}$)</td>
<td>d · mol · L$^{-1}$</td>
</tr>
<tr>
<td>$\theta_{NOB,2}$</td>
<td>Kinetic parameter for NOB activity</td>
<td>1.1 (1.13)</td>
<td>d</td>
</tr>
<tr>
<td>$\Lambda$ ($\Lambda_h$)</td>
<td>Extent variance-covariance matrix (for the $h$th sample)</td>
<td>–</td>
<td>mol$^2$ / –</td>
</tr>
<tr>
<td>$\lambda$ ($\lambda_{i,h}$)</td>
<td>Extent variance (for the $i$th reaction and the $h$th sample)</td>
<td>–</td>
<td>mol$^2$</td>
</tr>
<tr>
<td>$\Sigma$ ($\Sigma_h$)</td>
<td>Measurement error variance-covariance matrix (for the $h$th sample)</td>
<td>–</td>
<td>(mol · L$^{-1}$)$^2$ / –</td>
</tr>
<tr>
<td>$\sigma_{TAN}$</td>
<td>Measurement standard deviation for TAN</td>
<td>0.01</td>
<td>mol · L$^{-1}$</td>
</tr>
<tr>
<td>$\sigma_{TNO_2}$</td>
<td>Measurement standard deviation for TNO2</td>
<td>0.01</td>
<td>mol · L$^{-1}$</td>
</tr>
<tr>
<td>$\sigma_{TNO_3}$</td>
<td>Measurement standard deviation for TNO3</td>
<td>0.01</td>
<td>mol · L$^{-1}$</td>
</tr>
<tr>
<td>$\sigma_{pH}$</td>
<td>Measurement standard deviation for pH</td>
<td>0.05</td>
<td>–</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Integrand (time)</td>
<td>–</td>
<td>h</td>
</tr>
<tr>
<td>$[\cdot]$</td>
<td>Concentration symbol equivalent to $c$</td>
<td>–</td>
<td>mol · L$^{-1}$</td>
</tr>
</tbody>
</table>
Table 2: List of candidate rate laws used for both nitritation and nitratation reactions. The substrate concentration $c_S$ is the free ammonia concentration ([NH$_3$]) for the nitritation and the free nitrous acid concentration ([HNO$_2$]) for the nitratation.

<table>
<thead>
<tr>
<th>Name</th>
<th>Index</th>
<th>Candidate rate laws $r_{k,i}^{(j)}(c_S, \theta_1^{(j)}), i \in {1, 2}$</th>
<th>Parameter vector $\theta_1^{(j)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zeroth order</td>
<td>1</td>
<td>$\begin{cases} 1/\theta_{i,1}^{(1)} &amp; \text{if: } c_S \geq 0 \ 0 &amp; \text{otherwise} \end{cases}$</td>
<td>$[\theta_{i,1}^{(1)}]$</td>
</tr>
<tr>
<td>First order</td>
<td>2</td>
<td>$c_S / \theta_{i,1}^{(2)}$</td>
<td>$[\theta_{i,1}^{(2)}]$</td>
</tr>
<tr>
<td>Monod</td>
<td>3</td>
<td>$c_S / (\theta_{i,1}^{(3)} + \theta_{i,2}^{(3)} c_S)$</td>
<td>$[\theta_{i,1}^{(3)} \theta_{i,2}^{(3)}]^T$</td>
</tr>
<tr>
<td>Tessier</td>
<td>4</td>
<td>$1 - \exp \left( -c_S \theta_{i,1}^{(4)} \theta_{i,2}^{(4)} / \theta_{i,1}^{(4)} \right)$</td>
<td>$[\theta_{i,1}^{(4)} \theta_{i,2}^{(4)}]^T$</td>
</tr>
<tr>
<td>Haldane</td>
<td>5</td>
<td>$c_S / (\theta_{i,1}^{(5)} + \theta_{i,2}^{(5)} c_S / \theta_{i,3}^{(5)} c_S)$</td>
<td>$[\theta_{i,1}^{(5)} \theta_{i,2}^{(5)} \theta_{i,3}^{(5)}]^T$</td>
</tr>
</tbody>
</table>
Figure 1: **Data generation:** Simulated concentration (in gN/L) and pH (continuous and dashed lines) with corresponding measurements (dots, squares, triangles).
Figure 2: **Method 1: Simultaneous model identification.** WRMSR values for 25 models. The markers indicate the selected rate law for the first rate law. Shading of the bars indicates the selected rate law for the second reaction. The WRMSR values for the true model and for the best model obtained with Method 2 are indicated by a full and a dashed line, respectively.
Figure 3: Method 2: Extent-based modeling – Step 1: Computation of experimental extents. (a) True (lines) and experimental (dots) extents with $3\sigma$ confidence intervals for the nitritation (gray) and nitratation (black) reactions. (b) Variance-covariance matrix as $3\sigma$ confidence region (ellipsoid) for the experimental extent errors around $(0,0)$; colored lines corresponding to $t_h = 0.5, 1.5, 2.5, 3.5$ and $4.5$ h indicated with matching colors and styles in (a).
Figure 4: **Method 2: Extent-based modeling – Step 2: Modeling of 1st extent.** (a) Experimental (circles, with error bars) and modeled (continuous lines) extents as functions of time; (b) Residuals between modeled and experimental extents as functions of time; (c) Normalized residuals between modeled and experimental extents as functions of time.
Figure 5: Method 2: Extent-based modeling – Step 2: Modeling of extents – Lack-of-Fit. WRMSR for all extents and all candidate rate laws (bars) and 95% upper control limits of the associated $\chi^2$-distribution (lines).
Figure 6: **Method 2: Extent-based modeling – Step 2: Modeling of 2nd extent.** (a) Experimental (circles, with error bars) and modeled (continuous lines) extents as functions of time; (b) Residuals between modeled and experimental extents as functions of time; (c) Normalized residuals between modeled and experimental extents as functions of time.
Figure 7: **Method 2: Extent-based modeling – Step 3: Model fine-tuning.** Simulation of the TAN, TNO2, and TNO3 concentrations and pH for (i) the true data-generating model, (ii) the model obtained before fine-tuning, and (iii) the model obtained after fine-tuning. Differences between these simulations are barely noticable.
Supporting Information Available

Supporting Information includes bounding procedures, additional figures, and all code to produce our results.

This material is available free of charge via the Internet at http://pubs.acs.org/.

References


(32) Chandran, K.; Smets, B. F. Estimating biomass yield coefficients for autotrophic am-


