Adv Biochem Eng Biotechnol (2013) 132: 137–166
DOI: 10.1007/10_2012_166

Published Online: 20 December 2012

Applying Mechanistic Models in Bioprocess Development

Rita Lencastre Fernandes, Vijaya Krishna Bodla, Magnus Carlquist, Anna-Lena Heins, Anna Eliasson Lantz, Gürkan Sin and Krist V. Gernaey

Abstract The available knowledge on the mechanisms of a bioprocess system is central to process analytical technology. In this respect, mechanistic modeling has gained renewed attention, since a mechanistic model can provide an excellent summary of available process knowledge. Such a model therefore incorporates process-relevant input (critical process variables)–output (product concentration and product quality attributes) relations. The model therefore has great value in planning experiments, or in determining which critical process variables need to be monitored and controlled tightly. Mechanistic models should be combined with proper model analysis tools, such as uncertainty and sensitivity analysis. When assuming distributed inputs, the resulting uncertainty in the model outputs can be decomposed using sensitivity analysis to determine which input parameters are responsible for the major part of the output uncertainty. Such information can be used as guidance for experimental work; i.e., only parameters with a significant influence on model outputs need to be determined experimentally. The use of mechanistic models and model analysis tools is demonstrated in this chapter. As a practical case study, experimental data from *Saccharomyces cerevisiae* fermentations are used. The data are described with the well-known model of Sonnleitner and Käppeli (Biotechnol Bioeng 28:927–937, 1986) and the model is analyzed further. The methods used are generic, and can be transferred easily to other, more complex case studies as well.

Keywords Fermentation · Identifiability · Modeling · Monte Carlo · PAT · *Saccharomyces cerevisiae* · Sensitivity · Uncertainty

R. Lencastre Fernandes · V. K. Bodla · G. Sin · K. V. Gernaey (✉)
Department of Chemical and Biochemical Engineering, Technical University of Denmark,
Building 229, 2800 Lyngby, Denmark
e-mail: kvg@kt.dtu.dk

M. Carlquist
Division of Applied Microbiology, Department of Chemistry, Lund University,
22100 Lund, Sweden

A.-L. Heins · A. E. Lantz
Center for Microbial Biotechnology, Department of Systems Biology,
Technical University of Denmark, Building 223, 2800 Lyngby, Denmark

Abbreviations

API	Active pharmaceutical ingredient
EMEA	European Medicines Agency
FDA	Food and Drug Administration
MW	Molecular weight
NBE	New biological entity
NCE	New chemical entity
PAT	Process analytical technology
PSE	Process systems engineering
QbD	Quality by design
RTR	Real-time release
OD	Optical density
DW (Biomass)	Dry weight

Contents

1 Introduction

The pharmaceutical industry is changing rapidly nowadays. One important change, compared with the situation 10 or 20 years ago, is undoubtedly the increased focus on development of more efficient production processes. The introduction of process analytical technology (PAT) by the Food and Drug Administration [2] forms an important milestone here, since its publication ended a long period of regulatory uncertainty. The PAT guidance indeed makes it clear that regulatory bodies are in favor of more efficient production methods, as long as a safe product can be guaranteed. This opens up new and exciting possibilities for innovation in pharmaceutical production processes.

One of the central concepts in PAT is the *design space*, which is defined as "the multi-dimensional combination of critical input variables and critical process parameters that lead to the right critical quality attributes" [2]. The term "critical" should be interpreted as "having a significant influence on final product quality." Changing the process within the design space is therefore not considered as a change. As a consequence, no regulatory postapproval of the process is required for a change within the design space. Almost naturally, this opens up the possibility of increased use of optimization methods for pharmaceutical processes in the future, methods that have been used for a long time in, for example, the chemical industry [3].

Small-molecule (MW $<$ 1,000) drug substances (APIs, NCEs) are typically produced via organic synthesis. In such a production system, the available process knowledge is often relatively large. Process systems engineering (PSE) methods and tools—especially those relying on mechanistic models to represent available process knowledge—are therefore increasingly applied in the frame of pharmaceutical process development and innovation of small-molecule drugs [4], with the aim of shortening time to market while yielding an efficient production process. In essence, mechanistic models rely on deterministic principles to represent available process knowledge on the basis of mass, energy, and momentum balances; given initial conditions, future system behavior can be predicted.

It is, however, not the intention here to provide a detailed review on mechanistic models for biobased production processes of pharmaceuticals. There are excellent textbooks and review articles on the general principles of mechanistic modeling of fermentation processes [5–8], biocatalysis [9, 10], and mammalian cell culture [11].

Biotechnology research has resulted in a new class of biomolecular drugs—typically larger molecules, also called biologics or NBEs—which includes monoclonal antibodies, cytokines, tissue growth factors, and therapeutic proteins. The production of biomolecular drugs is usually complicated and extremely expensive. The level of process understanding is therefore in many cases lower, compared with small-molecule drug substances, and as a consequence, PSE methods and tools relying on mechanistic models are usually not applied to the same extent in production of biomolecular drugs, despite the fact that quite a number of articles have been published throughout the years on the development of mechanistic models for such processes.

This chapter focuses on the potential use of mechanistic models within biobased production of drug products, as well as the use of good modeling practice (GMoP) when using such mechanistic models [12]. A case study with the yeast model by Sonnleitner and Käppeli [1] is used to illustrate how a mechanistic model can be formulated in a well-organized and easy-to-interpret matrix notation. This model is then analyzed using uncertainty and sensitivity analysis, an analysis that serves as a starting point for a discussion on the potential application of such methods. Strategies for mechanistic model-building are highlighted in the final discussion.

2 Case Study: Aerobic Cultivation of Budding Yeast

Saccharomyces cerevisiae is one of the most relevant and intensively studied microorganisms in biotechnology and bioprocess engineering; For example, out of 151 recombinant biopharmaceuticals that had been approved by the FDA and EMEA in January 2009, 28 (or 18.5 %) were produced in *S. cerevisiae* [13]. Sonnleitner and Käppeli [1] proposed a widely accepted mechanistic model describing the aerobic growth of budding yeast, and this model is used here to exemplify how a mechanistic model of a bioprocess can be applied to create more in-depth process knowledge. Optimally, the process knowledge should be translated into a mechanistic model, and the model should be updated whenever additional details of the process are unraveled. This model should capture the key phenomena taking place in the process, and be further employed in the development of process control strategies.

However, when developing and using mechanistic models, reliability of the model (hence the credibility of model-based applications) is an important issue, which needs to be assessed using appropriate methods and tools including identifiability, sensitivity, and uncertainty analysis techniques. Unfortunately, literature reporting on mechanistic model developments often lacks the results of such analysis—confidence intervals on estimated parameters, for example, are only sporadically reported—and as a consequence it is not possible to conclude about the quality of the model and its predictions. Seen from a PAT perspective, it is of utmost importance to document that one has constructed a reliable mechanistic model; For example, in case this model would be used later for simulations to help in determining where to put the borders of the design space, it would be difficult to defend the resulting design space—for example, towards the FDA—in case the reliability of the model cannot be documented sufficiently.

One of the challenges in modeling is the identifiability problem, defined as “given a set of data, how well can the unknown model parameters be estimated, hence identified.” Typically, the number of parameters in a mechanistic model is relatively high, and therefore it is often not possible to uniquely estimate all the parameters by fitting the model predictions to experimental measurements. An indication of the parameters that can be estimated based on available data can be obtained by performing an identifiability analysis prior to the parameter estimation.

Furthermore, the model predictions will depend on the values of all parameters. Some of the parameters will, however, have a stronger influence than others. An uncertainty and sensitivity analysis can be performed to determine which are the parameters whose variability contributes most to the variance of the different model outputs.

In this case study, a systematic model analysis is performed following the workflow presented in Fig. 1. This workflow is rather generic, and could easily be transferred to another case study with a similar model.

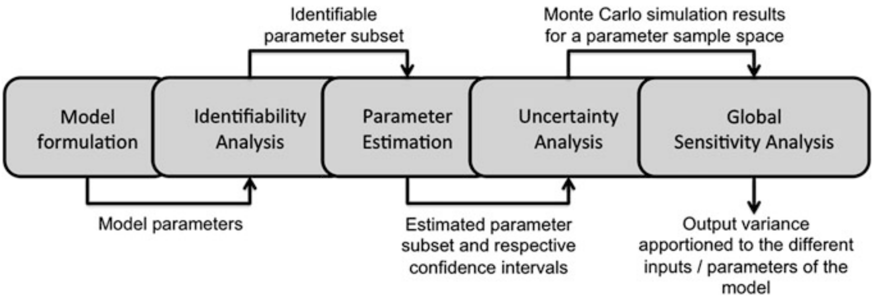

Identifiable parameter subset
Monte Carlo simulation results for a parameter sample space
Model formulation
Identifiability Analysis
Parameter Estimation
Uncertainty Analysis
Global Sensitivity Analysis
Model parameters
Estimated parameter subset and respective confidence intervals
Output variance apportioned to the different inputs / parameters of the model

Fig. 1 Schematic workflow for the model analysis

2.1 Model Formulation

Under aerobic conditions, budding yeast may exclusively oxidize glucose (respiratory metabolism), or simultaneously oxidize and reduce glucose (fermentative metabolism) if the respiratory capacity of the cells is exceeded. The described overflow metabolism is commonly referred to as the Crabtree effect. Cells preferably oxidize glucose, as the energetic yield is more favorable for respiration than fermentation. In case the respiratory capacity is reached, the excess of glucose (i.e., overflow of glucose) is reduced using fermentative pathways that result in the production of ethanol. Moreover, in a second growth phase, yeast will then consume the produced ethanol, but only after depletion of glucose, as the latter inhibits the consumption of any other carbon source. Also acetate and glycerol are formed and consumed, although the corresponding concentrations are typically much lower than for ethanol.

The Sonnleitner and Käppeli [1] model describes the glucose-limited growth of *Saccharomyces cerevisiae*. This model is able to account for the overflow metabolism, and to predict the concentrations of biomass, glucose, ethanol, and oxygen throughout an aerobic cultivation in a stirred tank reactor. Acetate and glycerol are not included for simplification purposes. The model relies on three stoichiometric reactions describing the growth of biomass on glucose by respiration (Eq. 1) and by fermentation (Eq. 2), as well as the growth of biomass on ethanol by respiration (Eq. 3). The stoichiometry of the three different pathways can be summarized in a matrix form (Table 1) describing how the consumption of glucose, ethanol, and oxygen are correlated with the production of biomass and ethanol, i.e., the yields of the reactions. The mol-based stoichiometric coefficients can be converted into the corresponding mass-based yields, e.g., $Y_{\mathrm{XG}}^{\mathrm{Oxid}} = b \times$ MW(biomass)/MW(glucose).

$$C_6H_{12}O_6 + aO_2 + b\,0.15\,[NH_3] \rightarrow bC_1H_{1.79}O_{0.57}N_{0.15} + cCO_2 + dH_2O \quad (1)$$

$$C_6H_{12}O_6 + g\,0.15\,[NH_3] \rightarrow gC_1H_{1.79}O_{0.57}N_{0.15} + hCO_2 + iH_2O + jC_2H_6O \quad (2)$$

Table 1 Stoichiometric matrix describing aerobic growth of budding yeast

Component i →	C_1 Glucose		C_2 Ethanol		C_3 Oxygen		C_4 Biomass	
Symbols	G		E		O		X	
Units	mol l^{-1}	g l^{-1}	mol l^{-1}	g l^{-1}	mol l^{-1}	g l^{-1}	C-mol l^{-1}	g l^{-1}
Process j ↓								
Biomass growth by glucose oxidation (Eq. 1)	-1	-1	0	0	a	Y_{OG}	b	Y_{XG}^{Oxid}
Biomass growth by glucose reduction (Eq. 2)	-1	-1	j	Y_{EG}	0	0	g	Y_{XG}^{Red}
Biomass growth by ethanol oxidation (Eq. 3)	0	0	-1	-1	k	Y_{OE}	l	Y_{XE}

$$C_6H_6O + kO_2 + l\,0.15[NH_3] \rightarrow lC_1H_{1.79}O_{0.57}N_{0.15} + mCO_2 + nH_2O \tag{3}$$

For each pathway, a mass balance can be established for each atomic element (e.g. C or N). To solve such elemental balances for carbon, hydrogen, and oxygen, one stoichiometric coefficient for each pathway has to be assumed. Since the biomass yield coefficients are often easily estimated from experimental data, they are typically the ones that are assumed. Therefore, only the coefficients b, g and l, or the corresponding mass yields Y_{XG}^{Oxid}, Y_{XG}^{Red}, and Y_{XE} will be considered as model parameters; i.e., the other stoichiometric coefficients are fixed based on Eqs. 1–3.

Furthermore, a process matrix can be used to describe the rates of consumption and production of each of the model variables (glucose, ethanol, oxygen, and biomass), as well as the fluxes in each pathway. Details on the use of this matrix notation are provided by Sin and colleagues [14]. The interested reader can find additional details on elemental mass and energy balances applied to fermentation processes elsewhere [15, 16].

In the case of the model used as an example here, the total glucose consumption and ethanol consumption rates (when considered individually) are mathematically described using Monod-type kinetics (Eqs. 4–6). The maximum uptake rates for glucose, ethanol, and oxygen ($r_{i,\max}$) are model parameters, and they are characteristic of the *S. cerevisiae* strain being used. The same goes for the substrate saturation constants: K_G, K_E, and K_O. The maximum oxygen uptake rate ($r_{O,\max}$) corresponds to the respiratory capacity, as it reflects the maximum rate for oxidation of glucose or ethanol when any of these carbon sources is in excess. The ethanol uptake rate includes a term accounting for glucose repression; i.e., ethanol consumption is only observed for low concentrations of glucose. The strength of inhibition (i.e., how low the glucose concentration should be before ethanol consumption is allowed) is defined by the inhibition constant K_i. The specific growth rate of biomass is defined as the sum of the growth resulting from each pathway, and is estimated based on the yield of biomass on the substrate and the corresponding uptake rate (Eq. 7).

$$r_G^{Total} = r_{G,\max}\frac{G}{G + K_G} = r_G^{Oxid} + r_G^{Red} \tag{4}$$

$$\mathrm{r_E} = \mathrm{r_{E,max}} \frac{\mathrm{E}}{\mathrm{E} + K_\mathrm{E}} \frac{K_i}{\mathrm{G} + K_i} \tag{5}$$

$$\mathrm{r_O} = \mathrm{r_{O,max}} \frac{\mathrm{O}}{\mathrm{O} + K_\mathrm{O}} \tag{6}$$

$$\mu_{\mathrm{Total}} = +Y_{\mathrm{XG}}^{\mathrm{Oxid}} \times \mathrm{r}_{\mathrm{G}}^{\mathrm{Oxid}} + Y_{\mathrm{XG}}^{\mathrm{Red}} \times \mathrm{r}_{\mathrm{G}}^{\mathrm{Red}} + Y_{\mathrm{XE}} \times \mathrm{r}_{E}^{Oxid} \tag{7}$$

The rate of oxidation and the rate of reduction of glucose are defined based on the maximum oxygen uptake rate: if the oxygen demand that is stoichiometrically required for oxidation of the total glucose flux ($Y_{\mathrm{OG}} \times r_{\mathrm{G}}{}^{\mathrm{Total}}$) exceeds the maximum oxygen uptake rate ($r_{\mathrm{O,max}}$), the difference between the two fluxes corresponds to the overflow reductive flux. With regard to the oxidation of ethanol, the observed rate of ethanol oxidation depends on the ethanol availability (Eq. 5) and it is further limited by the respiratory capacity: not only the maximum capacity of the cell, but also the capacity remaining after considering metabolism of glucose (Table 2).

In addition to the reactions taking place in the cells, oxygen is continuously supplied to the bioreactor. This supply is described based on the mass transfer coefficient ($k_{\mathrm{L}}a$) and the difference between the dissolved oxygen concentration (O) and the saturation concentration of oxygen in water (O^*) as a driving force. $k_{\mathrm{L}}a$ is dependent on the aeration intensity and the mixing conditions in a given fermentor. It is also dependent on the biomass concentration, although this dependence is often disregarded. The rates for each component can be obtained from the process model matrix (Table 2) by multiplying the transpose of the stoichiometric matrix (**Z**’) by the process rate vector ($\boldsymbol{\rho}$): $r_{m,1} = \mathbf{Z}\prime_{nxm} \times \boldsymbol{\rho}_{nx1}$, where m corresponds to the number of components (or model variables) and n is the number of processes. In Table 3, a nomenclature list of vectors and matrices is presented.

The model matrix in Table 2 provides a compact overview of the model equations. In the example here, it contains information about the biological reactions and the transfer of oxygen from the gas to the liquid phase. Of course, depending on the purpose of the model, the model matrix could be extended with additional equations, for instance, aiming at a more detailed description of the biological reactions, e.g., by including additional state variables, or aiming at the description of the mass transfer of additional components, e.g., $\mathrm{CO_2}$ stripping from the fermentation broth. Sin and colleagues [14] provided an example of the extension of the model matrix with chemical processes for the kinetic description of mixed weak acid–base systems. The latter is important in case pH prediction is part of the purpose of the model. In the work of Sin and colleagues [14], the yield coefficients are all part of the stoichiometric matrix. In our case here, an alternative rate vector is presented, where all rates are normalized with regard to glucose.

Table 2 Process matrix describing the conversion rates and stoichiometry for each model variable: glucose, ethanol, oxygen, and biomass

Component i →	C_1 Glucose	C_2 Ethanol	C_3 Oxygen	C_4 Biomass	Rate vector (ρ)
Symbols	G	E	O	X	
Units	g l^{-1}	g l^{-1}	g l^{-1}	g l^{-1}	g l^{-1} h^{-1}
Process j ↓					
Biomass growth by glucose oxidation	-1	0	$-Y_{\mathrm{OG}}$	$-Y_{\mathrm{XG}}^{\mathrm{Oxid}}$	$\frac{1}{Y_{\mathrm{OG}}}\left(\min\left(r_{\mathrm{O,max}}\frac{\mathrm{O}}{\mathrm{O}+K_{\mathrm{o}}}, Y_{\mathrm{OG}} \times r_{\mathrm{G,max}}\frac{\mathrm{G}}{\mathrm{G}+K_{\mathrm{G}}}\right)\right)$
Biomass growth by glucose reduction	-1	Y_{EG}	0	$Y_{\mathrm{XG}}^{\mathrm{Red}}$	$r_{\mathrm{G,max}}\frac{\mathrm{G}}{\mathrm{G}+K_{\mathrm{G}}} - \frac{1}{Y_{\mathrm{OG}}}\left(\min\left(r_{\mathrm{O,max}}\frac{\mathrm{O}}{\mathrm{O}+K_{\mathrm{o}}}, Y_{\mathrm{OG}} \times r_{\mathrm{G,max}}\frac{\mathrm{G}}{\mathrm{G}+K_{\mathrm{G}}}\right)\right)$
Biomass growth by ethanol oxidation	0	-1	$-Y_{\mathrm{OE}}$	Y_{XE}	$\frac{1}{Y_{\mathrm{OE}}}\left(\min\left(r_{\mathrm{O,max}}\frac{\mathrm{O}}{\mathrm{O}+K_{\mathrm{o}}} - \min\left(r_{\mathrm{O,max}}\frac{\mathrm{O}}{\mathrm{O}+K_{\mathrm{o}}}, Y_{\mathrm{OG}} \times r_{\mathrm{G,max}}\frac{\mathrm{G}}{\mathrm{G}+K_{\mathrm{G}}}\right), Y_{\mathrm{OE}} \times r_{\mathrm{E,max}}\frac{\mathrm{E}}{\mathrm{E}+K_{\mathrm{E}}}\frac{K_i}{\mathrm{G}+K_i}\right)\right)$
Oxygen supply	–	–	1	–	$k_L a$ (O*–O)

Table 3 Nomenclature list of matrices and vectors used in the model formulation and model analysis

Unit	Description
$\boldsymbol{Z}$	Stoichiometric matrix
$\boldsymbol{\rho}$	Process rate vector
$\boldsymbol{\theta}$	Vector of model parameters
$\mathbf{S}^{sc}$	Scaled sensitivity matrix
$\mathbf{s}_j$	Column vector of the sensitivity matrix: corresponding to sensitivity of the various model outputs to the parameter j
s_{ij}	Scaled sensitivity of the output i to the parameter j
δ_j	Importance index of parameter j
sc	Scaling factors

2.2 Parameter Identifiability Analysis

The model described in the previous sections has four variables—glucose (G), ethanol (E), oxygen (O), and biomass (X)—and 11 parameters. In addition, the oxygen saturation concentration in water (at growth temperature) is necessary for solving the model. A list of the parameters and their descriptions is provided in Table 4.

The maximum specific growth rate on ethanol $(\mu_{E,max})$ is defined as the product of the yield of biomass on ethanol (Y_{XE}) and the maximum specific ethanol uptake rate $(r^{Oxid}_{E,max})$. For consistency between parameters, the ethanol specific uptake rate is used as a parameter in this example.

The number of parameters is considerably larger than the number of model variables (or outputs), which is typical for this type of model. It is therefore questionable whether all parameters can be estimated based on experimental data, even if the four model variables were to be measured simultaneously. This is the subject of identifiability analysis, which seeks to identify which of the parameters

Table 4 Model parameters, corresponding units, and numerical values [12]

	Parameter	Value	Units
$r_{\mathrm{G,max}}$	Maximal specific glucose uptake rate	3.5	$\mathrm{g\ G\ g^{-1}\ X\ h^{-1}}$
$r_{\mathrm{O,max}}$	Maximal specific oxygen uptake rate	8×10^{-3}	$\mathrm{mol\ O\ g^{-1}\ X\ h^{-1}}$
$Y^{\mathrm{Oxid}}_{\mathrm{XG}}$	Yield of biomass on glucose (oxidation)	0.49	$\mathrm{g\ X\ g^{-1}\ G}$
$Y^{\mathrm{red}}_{\mathrm{XG}}$	Yield of biomass on glucose (reduction)	0.05	$\mathrm{g\ X\ g^{-1}\ G}$
Y_{XE}	Yield of biomass on ethanol	0.72	$\mathrm{g\ X\ g^{-1}\ E}$
$\mu_{\mathrm{E,max}}$	Maximal specific growth rate on ethanol	0.17	$\mathrm{h^{-1}}$
K_{G}	Saturation parameter for glucose uptake	0.5	$\mathrm{g\ l^{-1}}$
K_{O}	Saturation parameter for oxygen uptake	1×10^{-4}	$\mathrm{g\ l^{-1}}$
K_{E}	Saturation parameter for ethanol uptake	0.1	$\mathrm{g\ l^{-1}}$
K_i	Inhibition parameter: free glucose inhibits ethanol uptake	0.1	$\mathrm{g\ l^{-1}}$
k_La	Mass transfer coefficient	1,000	$\mathrm{h^{-1}}$

can be estimated with high degree of confidence based on the available experimental measurements.

The main purpose of such an identifiability analysis is in fact to increase the reliability of parameter estimation efforts from a given set of data [17]. One method available to perform such an analysis is the two-step procedure based on sensitivity and collinearity index analysis proposed by Brun and colleagues [18]. Accordingly, the method calculates two identifiability measures: (1) the parameter importance index (δ) that reflects the sensitivity of the model outputs to single parameters, and (2) the collinearity index (γ) which reflects the degree of near-linear dependence of the sensitivity functions of parameter subsets. A parameter subset (a combination of model parameters) is said to be identifiable if (1) the data are sufficiently sensitive to the parameter subset (above a cutoff value), and (2) the collinearity index is sufficiently low (below a cutoff value).

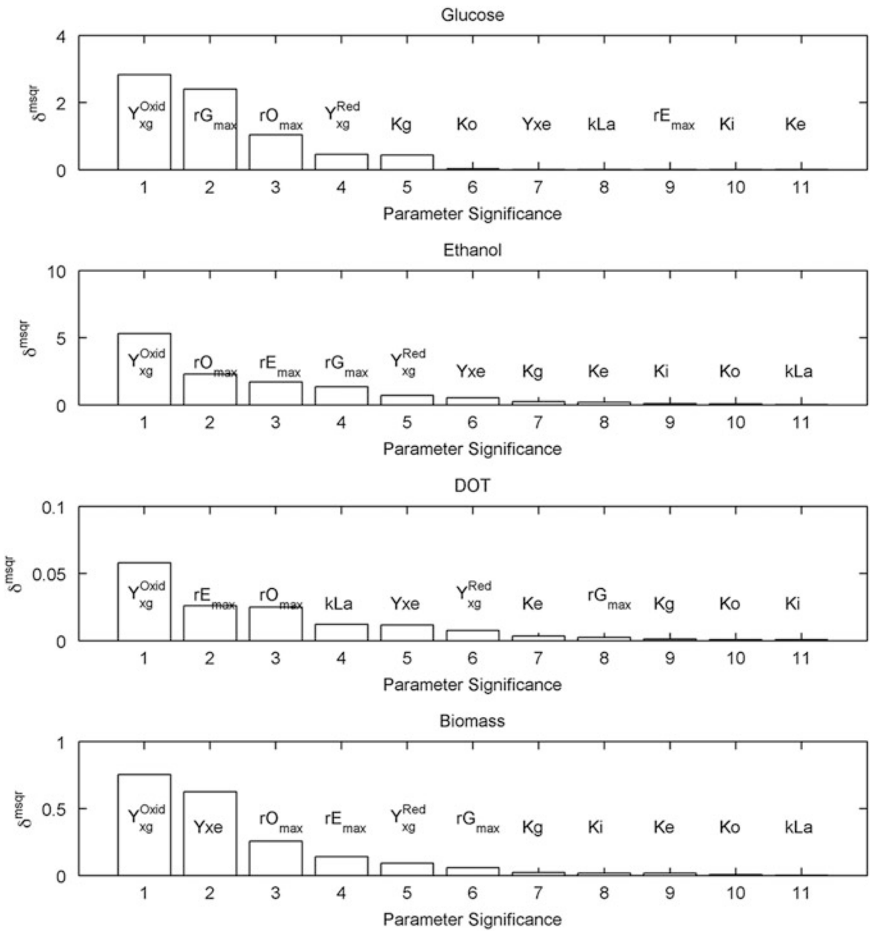
2.2.1 Local Sensitivity Analysis: Parameter Importance Indices δ

The local importance of an individual parameter to a model output for small changes ($\Delta\boldsymbol{\theta}$) in the parameter values ($\boldsymbol{\theta}$) at a specific location ($\boldsymbol{\theta}_0$) can be measured by the estimation of a dimension-free scaled sensitivity matrix $\mathbf{S}^{\mathrm{sc}} = \{s_{ij}\}$, where the index i refers to a specific model variable (output) and j denotes the model parameter. For further details, the reader is referred to the original paper of Brun and colleagues [18]. The mean squared norm of column $\boldsymbol{s}_j$, denoted by δ_j, is a measure of the importance of parameter θ_j (see Eqs 8–10). A large norm indicates that the parameter is identifiable with the available data if all other parameters are fixed. A parameter importance ranking can be obtained by ranking the parameters according to their δ indices. The lower the value of δ, the lower the importance of that parameter.

For this first analysis, the parameter values (Table 4) provided in the original paper [1] are used as nominal values at which sensitivity functions are calculated. The scaled sensitivity matrix **S** and the resulting rank of δ importance indices were calculated using Eqs. 8–10, and are graphically compared in Fig. 2. It is noteworthy that the δ indices are very sensitive to: (1) the choice of variation range defined for each parameter ($\Delta\boldsymbol{\theta}$), (2) scaling factors ($sc$) used to calculate the sensitivity matrix, and (3) the original set of parameters ($\boldsymbol{\theta}_0$), naturally as this is a local analysis. In this example the sc were defined as the mean of the experimental observations for each variable.

$$v_{ij} = \left.\frac{\partial \eta_i(\theta_j)}{\partial \theta_j}\right|_{\theta - \theta_0} \tag{8}$$

$$S_{ij} = v_{ij}\frac{\Delta\theta_j}{SC_j} \tag{9}$$



Glucose

Ethanol

DOT

Biomass

Fig. 2 Parameter importance indices (δ) for the four model variables: glucose, ethanol, dissolved oxygen and biomass

$$\delta_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} s_{ij}^2} \tag{10}$$

The results of the parameter significance ranking indicate that the yield coefficient $Y_{\mathrm{XG}}^{\mathrm{Oxid}}$ is the parameter that most affects all four model outputs. Variations in the maximum uptake rates will also have a significant effect on the model outputs. As may be expected, the glucose maximum uptake rate is most significant with regard to the model prediction for glucose, whereas the maximum uptake rate of ethanol is most important for the prediction of ethanol and dissolved oxygen.

The prediction of biomass is also greatly affected by the yield of biomass on ethanol, in addition to the yield on glucose (oxidative metabolism). The impact of the saturation constants is rather limited for any of the model variables.

2.2.2 Identifiability of Parameter Subsets: Collinearity Index γ_K

In addition to understanding the importance of individual parameters to the model output, it is necessary to take the joint influence of all parameters into account as well ($[\theta_1, \ldots, \theta_{j=J}]$). If columns $\boldsymbol{s}_j$ are nearly linearly dependent, the change of a parameter θ_j can be compensated by a change in the other parameter values. This means that the parameters $[\theta_1, \ldots, \theta_J]$ are not uniquely identifiable.

The collinearity index γ_K assesses the degree of near-linear dependence between a subset of K ($2 \le K \le J$) parameters, i.e., columns of the scaled sensitivity matrix.[1]

A high value of a collinearity index indicates that the parameter set is poorly identifiable. In practice, γ_K is calculated for all subsets of K parameters out of the 11 parameters and is plotted in Fig. 3. Also the subset size for each case is shown. In this case, a subset was considered identifiable if the corresponding collinearity index was smaller than 5. This threshold has to be defined a priori. Brun and colleagues [18] suggested as a rule of thumb that this threshold should lie in the range 5–20, where the lowest collinearity index corresponds to the strictest criterion. In practice, this decision on the threshold value is dependent on prior experience of the model user, and thus an iterative process.

All the model variables were considered in this analysis, implying as well that all could be measured experimentally. As illustrated in Fig. 3, a maximum of eight parameters can be identified, and the collinearity index increases with the number of parameters. The maximum collinearity index observed for combinations of eight parameters was 22.34, while the best identifiable sets of eight parameters correspond to a γ_K value of approximately 2.65. These parameter subsets are listed in Table 5.

It is indeed known that a change in the maximum uptake rate of glucose can be compensated with a change of biomass yield coefficients. Also, based on the model structure, it is clear that changes in yields for the oxidative and reductive consumptions of glucose can compensate each other. It is therefore not surprising that the parameter subsets that have higher collinearity index include these parameters. When comparing the subset of six parameters with the lowest collinearity index (last row in Table 5) with the "best" subset of eight parameters (shaded row in Table 5), the two parameters that have been removed in the subset of six parameters are the maximum uptake rates of ethanol and oxygen.

[1] Further discussion and equations are provided in the paper by Brun et al. (2002).

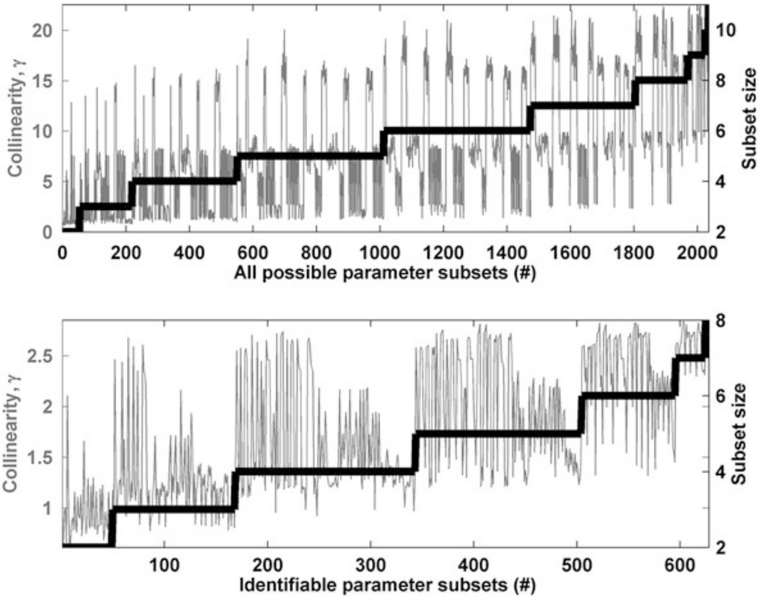


Fig. 3 Collinearity index and size corresponding to parameter subsets of increasing size. The top plot refers to all the parameter subsets evaluated in the analysis, whereas the bottom figure refers exclusively to the subsets that complied with the a priori defined collinearity threshold

The collinearity between the uptake rates and the yield coefficients explains why, even though they are the parameters with greatest importance for the model outputs (Fig. 2), they are not all included in the identifiable parameter subsets.

2.3 Parameter Estimation

Two datasets corresponding to two replicate batch fermentations of *S. cerevisiae* were available. For further details on the experimental data collection methods the reader is referred to the work of Carlquist et al. [19]. The dynamic profiles of glucose, ethanol, and biomass (as optical density, OD) were available for the two datasets, while oxygen data were only available for one of them. The OD measurements were converted into biomass dry weight (DW) values using a previously determined linear correlation (DW = 0.1815 × OD).

The parameters in the "best" identifiable subset were estimated by minimization of the weighted least-square errors. The weights for each variable *i* were defined by $w_i = 1\Big/(sc_i)^2$, and the scaled factors (also used in Eq. 9) were defined as the mean of the experimental observations for each given variable. The estimation was done simultaneously for the two datasets. The new estimates of the identifiable parameters

Table 5 Identifiable parameter subsets with maximum number of parameters and corresponding collinearity index

Parameter subset								Collinearity index	Identifiable parameter Set
$r_{\mathrm{G,max}}$	$r_{\mathrm{E,max}}$	$Y_{\mathrm{XG}}^{\mathrm{Oxid}}$	$Y_{\mathrm{XG}}^{\mathrm{Red}}$	Y_{XE}	K_{G}	K_{E}	$k_{L}a$	22.34	No
$r_{\mathrm{G,max}}$	$r_{\mathrm{O,max}}$	$r_{\mathrm{E,max}}$	$Y_{\mathrm{XG}}^{\mathrm{Oxid}}$	$Y_{\mathrm{XG}}^{\mathrm{Red}}$	Y_{XE}	K_{G}	$k_{L}a$	22.10	No
$r_{\mathrm{G,max}}$	$r_{\mathrm{E,max}}$	$Y_{\mathrm{XG}}^{\mathrm{Oxid}}$	$Y_{\mathrm{XG}}^{\mathrm{Red}}$	Y_{XE}	K_{G}	K_{i}	$k_{L}a$	22.10	No
$r_{G,\mathrm{max}}$	$r_{\mathrm{O,max}}$	$r_{\mathrm{E,max}}$	Y_{XE}	K_{G}	K_{E}	K_{i}	$k_{L}a$	2.65	Yes
$r_{\mathrm{G,max}}$	$r_{\mathrm{E,max}}$	$Y_{\mathrm{XG}}^{\mathrm{Red}}$	Y_{XE}	K_{G}	K_{E}	K_{i}	$k_{L}a$	2.75	Yes
$r_{G,\mathrm{max}}$	$r_{\mathrm{E,max}}$	Y_{XE}	K_{G}	K_{O}	K_{E}	K_{i}	$k_{L}a$	2.75	Yes
$r_{G,\mathrm{max}}$	$r_{\mathrm{E,max}}$	$Y_{\mathrm{XG}}^{\mathrm{Oxid}}$	Y_{XE}	K_{G}	K_{E}	K_{i}	$k_{L}a$	2.85	Yes
$r_{\mathrm{G,max}}$	Y_{XE}	K_{G}	K_{E}	K_{i}	$k_{L}a$	–	–	1.75	Yes

Table 6 Estimated values for the identifiable subset of parameters

Parameter	Initial guess	Estimated value	Units
$r_{\mathrm{G,max}}$	3.5	2.9	$\mathrm{g\ G\ g^{-1}\ X\ h^{-1}}$
$r_{\mathrm{O,max}}$	8×10^{-3}	5.5×10^{-3}	$\mathrm{mol\ O\ g^{-1}\ X\ h^{-1}}$
$r_{\mathrm{E,max}}$	0.24	0.32	$\mathrm{g\ E\ g^{-1}\ X\ h^{-1}}$
Y_{XE}	0.72	0.47	$\mathrm{g\ X\ g^{-1}\ E}$
K_{G}	0.5	0.17	$\mathrm{g\ G\ l^{-1}}$
K_{E}	0.1	0.56	$\mathrm{g\ E\ l^{-1}}$
K_{i}	0.1	0.31	$\mathrm{g\ G\ l^{-1}}$
$k_{L}a$	1,000	930	$\mathrm{h^{-1}}$

are presented in Table 6. In Fig. 4, the model predictions obtained with the estimated set of parameters are compared with the experimental data.

Generally, the model predictions are in good agreement with the experimental data. An overprediction of the biomass concentration and a slight underestimation of the ethanol concentration are however observed. The oxygen profile describes the drop of the dissolved oxygen concentration during the growth, and a steep increase upon the depletion of ethanol and the resulting growth arrest. The dynamics of oxygen described by the model assumes a constant mass transfer coefficient ($k_{\mathrm{L}}a$) and equilibrium between the gas and liquid phases. It is worth mentioning that the formation of other metabolites (i.e., glycerol and acetate) that are not considered in the model may explain the discrepancies to some degree. In fact, the overestimation of biomass which can be observed in Fig. 4 may be caused by the fact that other carbon-containing metabolites have not been taken into account.

When assessing the goodness of fit of the mechanistic model, it is important to consider that the experimental measurements have an associated error as well. Model predictions may not give a "perfect" fit at first sight, but they may well be within the experimental error. While such error might be relatively low for the measurement of glucose and ethanol by high-performance liquid chromatography

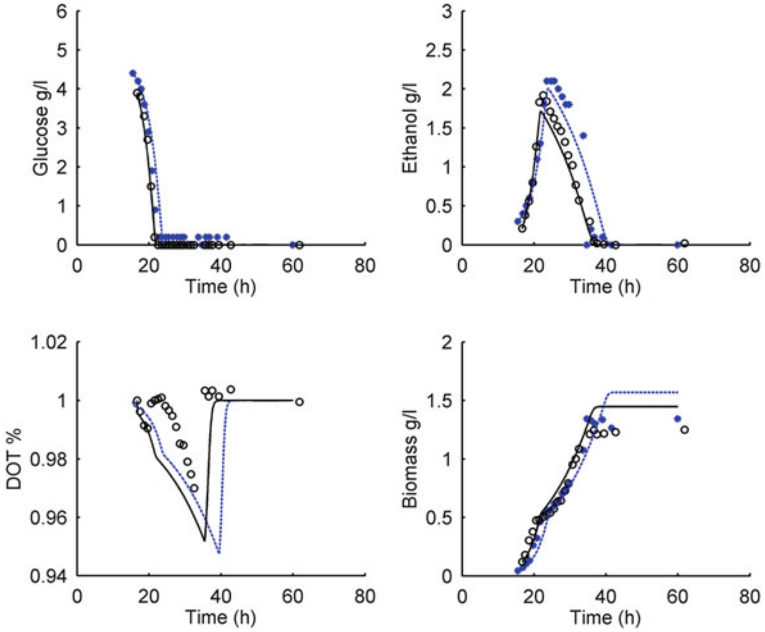


Fig. 4 Comparison of model predictions versus experimental data collected for cultivation 1 (*black line* model prediction, *black circles* experimental data) and cultivation 2 (*blue dashed line* model prediction, *blue stars*, experimental data)

(HPLC), it is significantly higher for dry weight measurements, which are less reliable, especially for low biomass concentrations (too large sample volumes would be required for increasing accuracy). Additionally, at the end of the fermentation, the biomass dry weight may include a fraction of nonviable and/or dormant cells.

2.3.1 Confidence Intervals for Estimated Parameters

The estimated parameter values as such only have limited value if they are not presented in combination with a measure of the degree of confidence that one can have in them. Therefore, the confidence intervals for each of the parameters are defined based on the covariance matrix and Student *t*-probability distribution. The covariance matrix is calculated using the residuals between model predictions and the standard deviations of the experimental measurements (further details are provided by Sin et al. [14]). An experimental error of 5 % was assumed for glucose and ethanol measurement by high-performance liquid chromatography (HPLC), as well as for the oxygen measurements using a gas analyzer for determining the composition of the exhaust gas, and a 20 % error for the determination of the cell dry weight. The confidence intervals at $(1 - \alpha)$ confidence level were

Table 7 Confidence intervals for the identifiable subset of parameters for 95 % confidence level

Parameter	Estimated value	Confidence interval	Units
$r_{\mathrm{G,max}}$	2.9	±9.8×10^{-2} (3.4 %)	g G g^{-1} X h^{-1}
$r_{\mathrm{O,max}}$	5.5×10^{-3}	±6.3×10^{-4} (11.6 %)	mol O g^{-1} X h^{-1}
$r_{\mathrm{E,max}}$	0.32	±0.24 (75.7 %)	g E g^{-1} X h^{-1}
Y_{XE}	0.47	±3.1×10^{-2} (6.6 %)	g X g^{-1} E
K_{G}	0.17	±8.4×10^{-2} (50.2 %)	g G l^{-1}
K_{E}	0.56	±0.44 (78.9 %)	g E l^{-1}
K_i	0.31	±0.30 (97.5 %)	g G l^{-1}
k_La	930	±49 (5.2 %)	h^{-1}

calculated using Eq. 11, where COV is the covariance matrix of the parameter estimators, $t(N - M, \alpha/2)$ is the t-distribution value corresponding to the $\alpha/2$ percentile, N is the total number of experimental observations (45 samples for the two cultivations), and M is the total number of parameters. The confidence intervals for the estimated parameters are presented in Table 7.

$$\theta_{1-\alpha} = \theta \pm \sqrt{\mathrm{diag}(\mathrm{COV}(\theta))} \cdot t\left(N - M, \frac{\alpha}{2}\right). \tag{11}$$

None of the confidence intervals include zero, giving a first indication that all parameters are significant to a certain degree and the model does not seem to be overparameterized. In the case of the inhibition constant K_{i}, the confidence interval is rather large. This is most likely a consequence of the low sensitivity of model outputs to this variable (Fig. 2). Furthermore, the confidence intervals of the Monod half-saturation constants K_{G} and K_{E} are quite large as well, which might be related to the fact that their estimated values are rather low. The latter means that the collected data do not contain that many data points which can be used during the parameter estimation for extracting information on the exact values of K_{G} and K_{E} Indeed, only the data corresponding to relatively low glucose and ethanol concentrations can be used, since the specific rates will be relatively constant and close to maximum for higher substrate concentrations.

It is furthermore also a good idea to analyze the values of the parameter confidence intervals simultaneously with the correlation matrix (Table 8); For example, the correlation matrix shows that $r_{\mathrm{E,max}}$ is correlated with K_{E} and that $r_{\mathrm{O,max}}$ is correlated with K_{O}. Both correlations are inherent to the model structure; i.e., correlation between the parameters related to the maximum specific growth rate and the substrate affinity constant in Monod-like kinetics expressions are quite common, and point towards a structural identifiability issue.

Note also that the significant correlations found between some of the model parameters (Table 8) seem to conflict with the results of the collinearity index analysis which was reported earlier (Fig. 3; Table 5). That is one of the reasons also for the identifiability analysis to be an iterative process.

Table 8 Correlation matrix for all model parameters

	$r_{\mathrm{G,max}}$	$r_{\mathrm{O,max}}$	$r_{\mathrm{E,max}}$	Y_{xg}^{Oxid}	Y_{xg}^{Red}	Y_{xe}	K_{G}	K_{O}	K_{E}	K_i	k_La
$r_{\mathrm{G,max}}$	1	0.166	0.081	0.538	−0.546	−0.168	−0.356	−0.145	−0.047	−0.134	−0.076
$r_{\mathrm{O,max}}$	0.166	1	0.546	−0.163	0.202	0.314	0.276	−0.904	−0.419	−0.548	0.114
$r_{\mathrm{E,max}}$	0.081	0.546	1	−0.038	0.105	0.338	0.134	−0.648	−0.950	−0.652	0.434
Y_{xg}^{Oxid}	0.538	−0.163	−0.038	1	−0.987	−0.421	0.165	0.017	0.095	−0.387	−0.041
Y_{xg}^{Red}	−0.546	0.202	0.105	−0.987	1	0.493	−0.206	−0.108	−0.145	0.342	0.131
Y_{Xe}	−0.168	0.314	0.338	−0.421	0.493	1	0.042	−0.515	−0.352	0.003	0.490
K_{G}	−0.356	0.276	0.134	0.165	−0.206	0.042	1	−0.255	−0.091	−0.425	0.009
K_{O}	−0.145	−0.904	−0.648	0.017	−0.108	−0.515	−0.255	1	0.499	0.583	−0.398
K_{E}	−0.047	−0.419	−0.950	0.095	−0.145	−0.352	−0.091	0.499	1	0.552	−0.373
K_i	−0.134	−0.548	−0.652	−0.387	0.342	0.003	−0.425	0.583	0.552	1	−0.047
k_La	−0.076	0.114	0.434	−0.041	0.131	0.490	0.009	−0.398	−0.373	−0.047	1

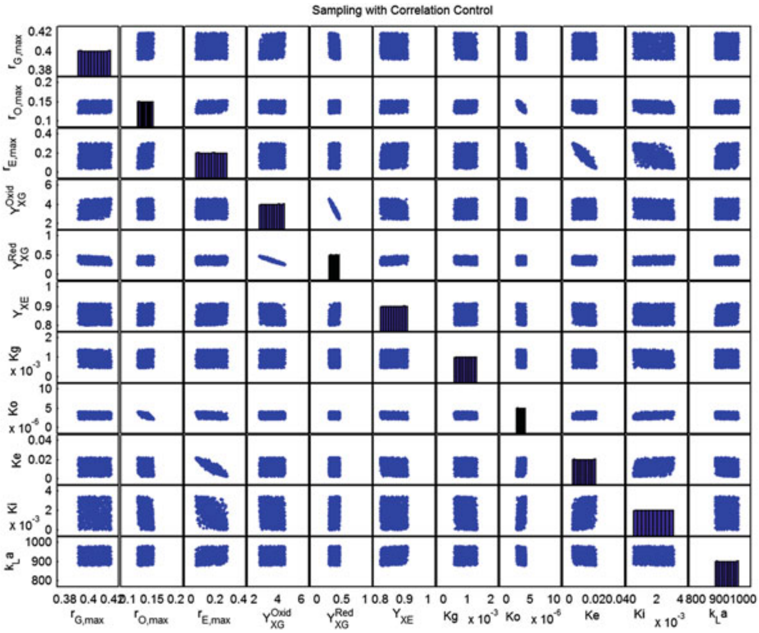


Fig. 5 Latin hypercube sampling for the model parameters, taking into account the correlation between them

2.4 Uncertainty Analysis

Uncertainty analysis allows for understanding the variance of the model outputs as a consequence of the variability in the input parameters. Such an analysis can be performed using the Monte Carlo procedure, which consists of three steps: (1) definition of the parameter space, (2) generation of samples of the parameter space, i.e., combinations of parameters, and (3) simulation of the model using the set of samples generated in the previous step. In this case study, a sample set of 1,000 combinations of parameter values was generated using the Latin hypercube sampling procedure [20]. This sampling technique can be set up such that it takes the correlations between parameters, i.e., information resulting from the parameter estimation, into account (as explained by Sin et al. [12]). The correlation matrix for all the parameters was estimated and is presented in Table 8. For each parameter, minimum and maximum values have to be defined: for the estimated parameters the limits of the 95 % confidence intervals were used, while a variability of 30 % around the default values was assumed for the remaining parameters.

The correlation between two parameters can take values between −1 and 1. A positive correlation indicates that an increase in the parameter value will result in an increase in the value of the other parameter as well. On the contrary, a negative value indicates an inverse proportionality. In Fig. 5, the sampling space is illustrated by scatter plots of combinations of two parameters. A high correlation

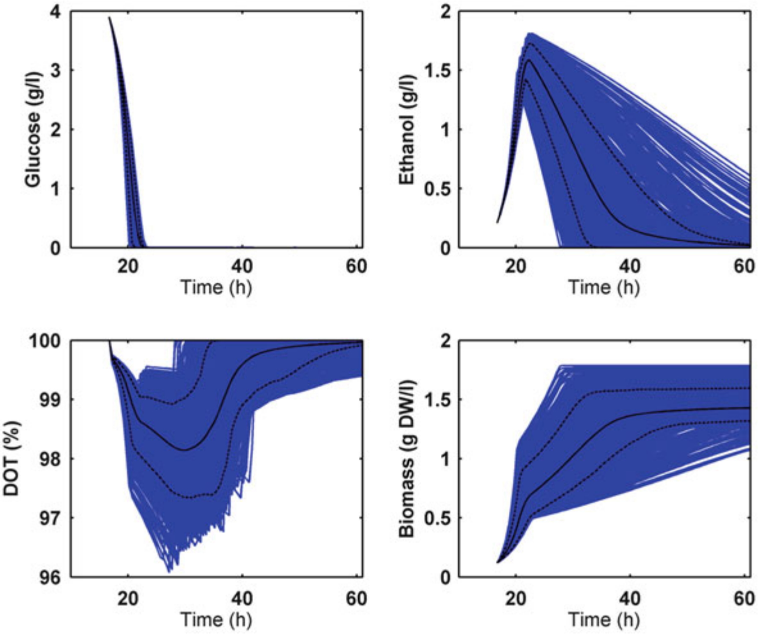


Fig. 6 Representation of uncertainty in the model predictions for glucose, ethanol, dissolved oxygen, and biomass: Monte Carlo simulations (*blue*), mean, and the 10th and 90th percentile of the predictions (*black*)

(in absolute value) will lead to an elliptical or linear cloud of sampling points, as, for example, for $Y_{\mathrm{XG}}^{\mathrm{Oxi}d}$ and $Y_{\mathrm{XG}}^{\mathrm{Red}}$ [corr($Y_{\mathrm{XG}}^{\mathrm{Oxid}}$, $Y_{\mathrm{XG}}^{\mathrm{Red}}$) $= -0.98$ in Table 8], as well as $r_{\mathrm{E,max}}$ and K_{E}, and $r_{\mathrm{O,max}}$ and K_{O}.

The number of samples and the assumed range of variability of each parameter (i.e., the parameter space) is defined by the expert performing the analysis. The higher the number of samples, the more effectively the parameter space will be covered, at the expense of increased computational time. The range of the parameter space should rely on previous knowledge of the process: (1) the initial guess of the parameter numerical values can be obtained from the literature or estimated in a first rough estimation where all parameters are included; (2) the variability (range) for each parameter can be determined by the confidence intervals, in case a parameter estimation has been done, or be defined based on expert knowledge as discussed by Sin et al. [12].

The estimations for the four model variables (outputs) and the corresponding mean and a prediction band defined by 10 and 90 % percentiles are presented in Fig. 6. The narrow prediction bands (including 80 % of the model predictions) for glucose reflect the robustness of the predictions for this model variable, while the wide bands observed, for example, for oxygen show the need for a more accurate estimate of the parameters in order to obtain a good model prediction.

2.5 Sensitivity Analysis: Linear Regression of Monte Carlo Simulations

Based on the Monte Carlo simulations, a global sensitivity analysis can be conducted. The aim of the sensitivity analysis is to break down the output uncertainty with respect to input (parameter) uncertainty. The linear regression method is a rather simple yet powerful analysis that assumes a linear relation between the parameter values and the model outputs. The sensitivity of the model outputs to the individual parameters, for a given time point, is summarized by a ranking of parameters according to the absolute value for the standardized regression coefficient (SRC). In a dimensionless form, the linear regression is described by Eq. 12, where sy_{ik} is the scalar value for the kth output, β_{jk} is the SRC of the jth input parameter, θ_j, for the kth model output, y_k, and its magnitude relates to how strongly the input parameter contributes to the output.

$$\frac{sy_{ik} - \mu_{sy_k}}{\sigma_{sy_k}} = \sum_{j=1}^{M} \beta_{jk} \times \frac{\theta_{ij} - \mu_{\theta j}}{\sigma_{\theta_j}} + \varepsilon_{ik} \qquad (12)$$

In the case of nonlinear dependence of the model variable on a parameter, this method can still be used, although with caution. As a rule of thumb, if the model coefficient of determination (R^2) is lower than 0.7, this analysis is not conclusive. The SRC for each parameter has, by definition, a value between -1 and 1, where a negative sign indicates that the output value will decrease when there is an increase in the value of the parameter. Oppositely, a positive SRC indicates direct proportionality between the parameter value and the model output. Sin et al. [12] describe further details on how to perform the analysis.

In the model example, different growth phases are described, and therefore the importance of the parameters is expected to change with time. Therefore, the analysis was performed for a selection of time points up to 62 h.

The suitability of applying the linear regression method was in this case also assessed for each time point and each output. The R^2 values are presented in Fig. 7 as a function of time.

While the regression method seems to be suitable for all time points in the case of biomass, the same is not observed for glucose, ethanol, and oxygen.. With regard to glucose, the model uncertainty is very small (narrow spread of the model predictions plotted in Fig. 6). The depletion of glucose is estimated to occur at time of approximately 22 h for all cases. The sensitivity analysis when the glucose concentrations are virtually zero is not expected to be significant, and it is thus not surprising that the R^2 value decreases abruptly at approximately the same time point that glucose is depleted. Simultaneously, the uncertainty in ethanol concentration predictions increases substantially. This may explain the temporary drop in the R^2 value for ethanol at this time point. A similar drop in R^2 is observed for oxygen around the time that ethanol is depleted, and a sudden rise in the dissolved oxygen concentration is observed. Upon ethanol depletion, the R^2 value

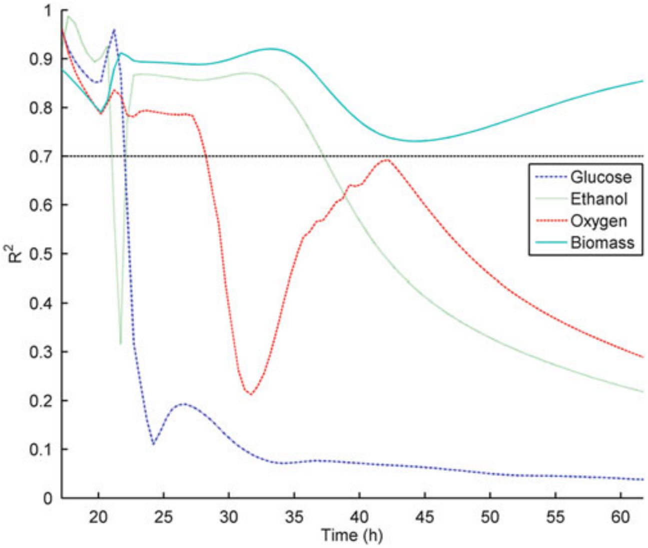


Fig. 7 Regression correlation coefficient (R^2) for each model output, indicating the goodness of the linear regression used for estimating the sensitivity of each model output to various parameters. For R^2 values lower than 0.7, the corresponding standardized regression coefficient (SRC) may yield erroneous information

for ethanol falls under the threshold, similarly to what was observed for glucose at its depletion.

In Fig. 8, an overview of the SRCs for each parameter and model output is presented. Interpretation of parameter ranking and SRC should be made cautiously. All model outputs seem to be sensitive to the yield coefficient of biomass on oxidized glucose, even during the growth phase on ethanol (after glucose depletion).

The ranking of each parameter according to the SRC for each model output is illustrated in Fig. 9. When analyzing this ranking, it is possible to see the decrease in sensitivity of the glucose prediction towards the maximum glucose uptake rate, as well as the simultaneous increase in sensitivity towards the maximum oxygen uptake rate, during the growth phase on glucose. This is in agreement with the fact that the consumption of glucose is initially only limited by the maximum uptake rate (excess of glucose in the media), and afterwards as the biomass concentration increases and glucose concentration decreases, the observed uptake rate is no longer maximal. Similar figures for the parameter ranking regarding ethanol, oxygen and biomass can be drawn.

With regard to the model predictions for ethanol, this model output is most sensitive to the maximum glucose uptake rate and biomass yield on glucose (reduction pathway) during the first growth phase, and later on the maximum ethanol uptake rate. This is in good agreement with the fact that the production of ethanol is a result of the reduction of glucose, and its consumption only takes place

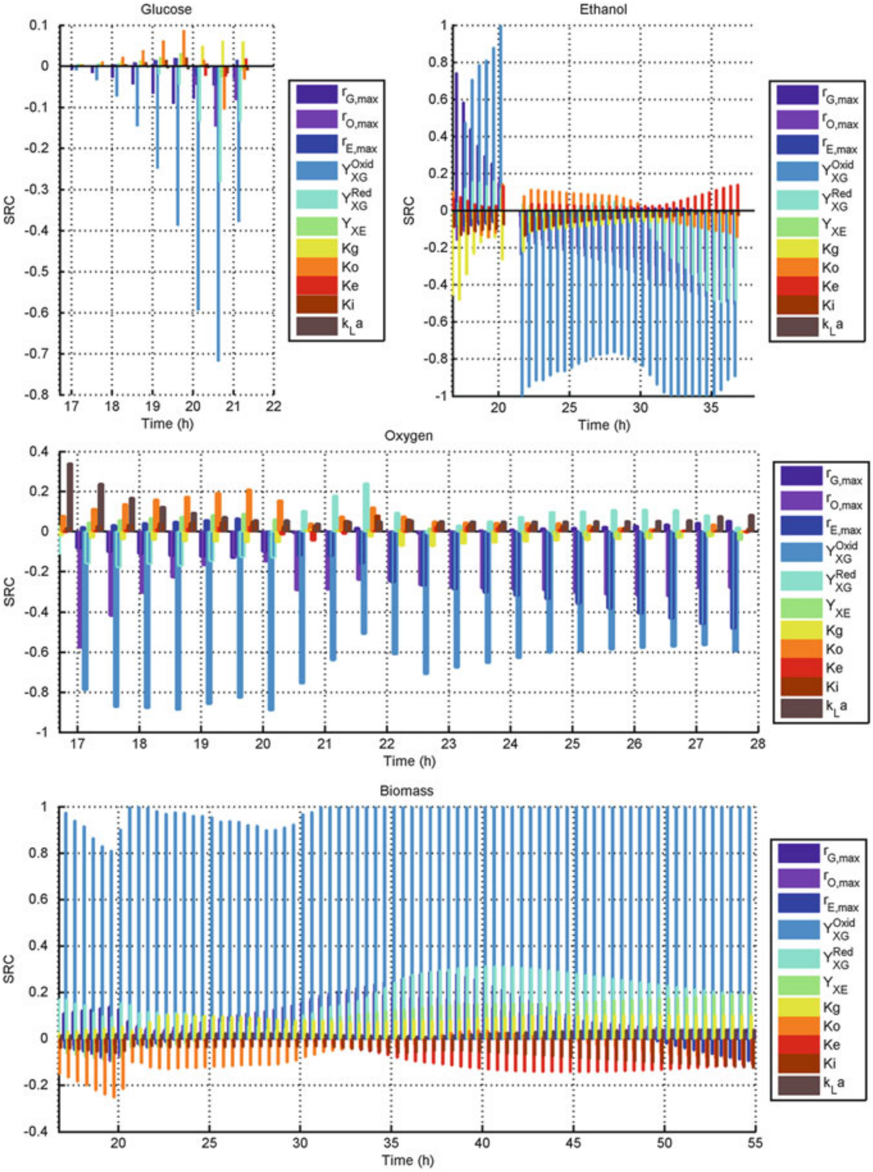


Fig. 8 Standardized regression coefficients (SRC) for the four model outputs as a function of time. Only the time points for which $R^2 > 0.7$ was observed are presented. Each color corresponds to a model parameter

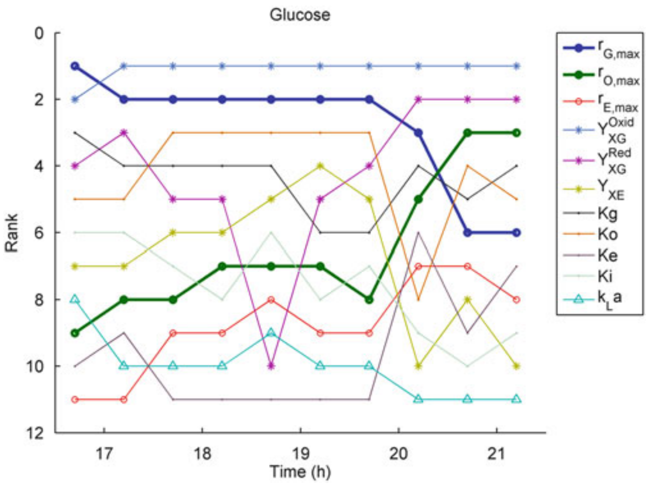


Fig. 9 Ranking of each model parameter according to the magnitude of the SRC for each model output: a rank of 1 indicates that the model output is most sensitive to that parameter, while a rank of 11 indicates that the parameter contributes the least to the variance of the model output

during the second growth phase following the depletion of glucose. A similar pattern was observed with regard to the model predictions for oxygen.

To analyze the sensitivity of the outputs to the parameters in more detail, two time points during the exponential growth phase on glucose (t = 17 h) and on ethanol (t = 27 h) were selected. The SRC and corresponding rank position for these time points are provided in Table 9a and b, respectively. As could be expected, during the growth on glucose, the parameters that most influence the prediction of glucose are the biomass yield parameters (for the two pathways) and the maximum uptake rate. The two yield coefficients have, however, a different effect on the glucose prediction: while an increase in the oxidative yield will lead to a lower predicted concentration, an increase in the reductive yield seems to imply an increase in the predicted concentration. This may reflect the fact that the oxidative pathway is the most effective way of transforming glucose into biomass.

The maximum glucose uptake rate is also the most influential parameter for the prediction of the ethanol concentration (produced by reduction of glucose), during this first growth phase. The glucose saturation rate plays an important role, however not as significant as the maximum uptake rate ($r_{G,max}$: SRC = 0.74; K_G: SRC = −0.48).

Obviously, the results of the global sensitivity analysis (SRC) should be compared with the results of local sensitivity analysis (Fig. 2). It can be seen that both methods rank the biomass yield on glucose (oxidation) as the most influential parameter. For the ranking of the other parameters, there are quite some differences between the results obtained by the two methods.

2.5.1 Morris Screening

As discussed by Sin et al. [11], an alternative to the linear regression method, especially when low R^2 values are observed, is Morris screening. Similarly to the linear regression method, a sampling-based approach is used. The method is based on Morris sampling, which is an efficient sampling strategy for performing randomized calculation of one-factor-at-a-time (OAT) sensitivity analysis. The parameters are assigned uniform distributions with lower and upper bounds defined by the confidence intervals for estimated parameters and by 30 % variability for the remaining ones (as done previously for the Latin hypercube sampling). The number of repetitions (r) was set to 90, corresponding to a sampling matrix with 1,080 [90 × (11 + 1)] different parameter combinations. The model was simulated for all the parameter combinations, and the results are summarized in Fig. 10.

The elementary effects (EE) were estimated as described by Sin et al. [12]. These EEs are described as random observations of a certain distribution function F, and are defined by Eq. 13, where Δ is a predetermined perturbation factor of θ_j, $sy_k(\theta_1, \theta_2, \theta_j,\ldots, \theta_M)$ is the scalar model output evaluated at input parameters $(\theta_1, \theta_2, \theta_j,\ldots, \theta_M)$, whereas $sy_k(\theta_1, \theta_2, \theta_j + \Delta,\ldots, \theta_M)$ is the scalar model output corresponding to a Δ change in θ_j.

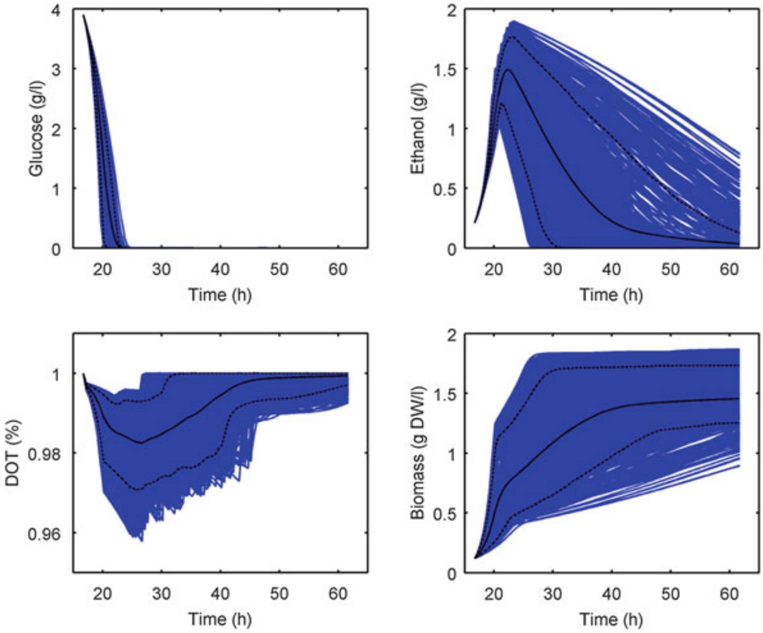


Fig. 10 Model simulation results using Morris sampling of parameter space: model simulations for glucose, ethanol, dissolved oxygen, and biomass showing simulations (*blue*), mean, and the 10th and 90th percentile of the simulations (*black*) (not to be confused with uncertainty analysis)

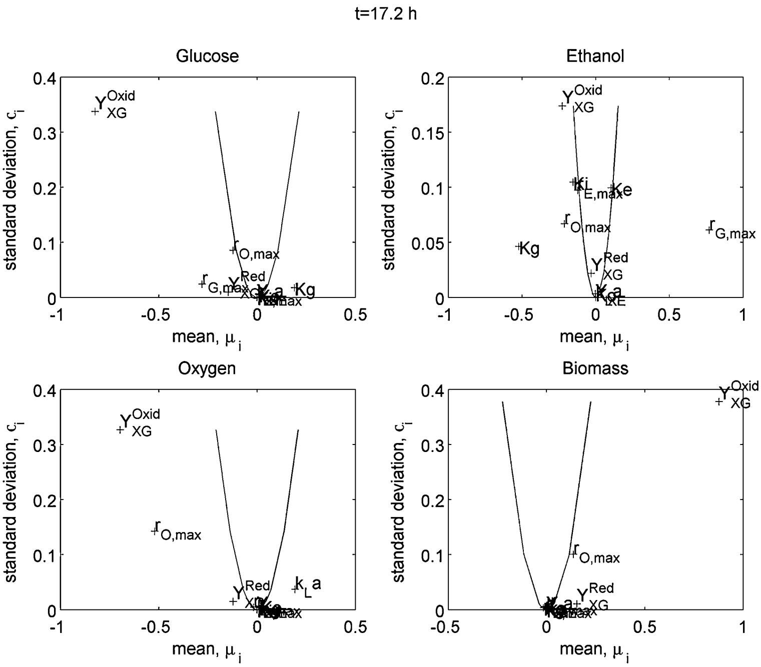


Fig. 11 Elementary effects during growth phase on glucose: estimated mean and standard deviation of the distributions of elementary effects of the 11 parameters on the model outputs. The two lines drawn in each subplot correspond to Mean$_i$ ± 2sem$_i$ (see text)

$$\begin{aligned} EE_{jk} &= \frac{\partial sy_k}{\partial \theta_j} \\ &= \frac{sy_k(\theta_1, \theta_2, \theta_j + \Delta, \ldots, \theta_M) - sy_k(\theta_1, \theta_2, \theta_j, \ldots, \theta_M)}{\Delta} \end{aligned} \tag{13}$$

The results obtained are compared with the mean and the standard deviation of this distribution. Often, the EEs obtained for each parameter are plotted together with two lines defined by Mean$_i$ ± sem$_i$, where Mean$_i$ is the mean effect for output i and sem$_\mathrm{i}$ is the standard error of the mean $(\mathrm{sem}_i = \mathrm{std\ deviation}_i/\sqrt{r})$. The EEs are scaled, and thus a comparison across parameters is possible.

Also this analysis has to be performed for a selected time point, or using a time-series average. As the cultivation has distinct phases, several time points were selected. The results for the growth phase on glucose (t = 17.2 h) and the growth phase on ethanol (t = 27.2 h) are presented in Figs. 11 and 12, respectively.

Parameters that lie in the area in between the two curves (inside the wedge) are said to have an insignificant effect on the output, while parameters outside the wedge have a significant effect. Moreover, nonzero standard deviations indicate nonlinear effects, implying that parameters with zero standard deviation and nonzero mean have a linear effect on the outputs.

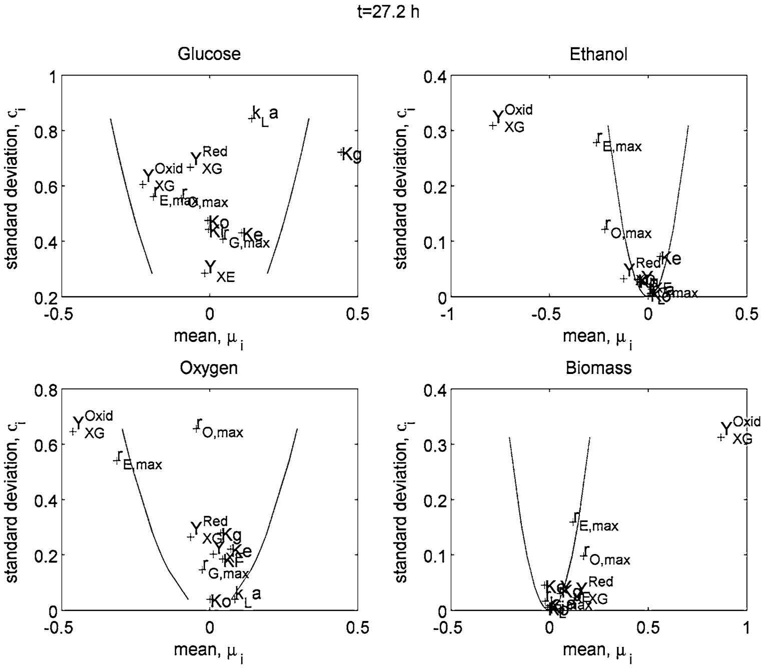


Fig. 12 Elementary effects during growth phase on ethanol: estimated mean and standard deviation of the distributions of elementary effects of the 11 parameters on the model outputs. The two lines drawn in each subplot correspond to Mean$_i$ ± 2sem$_i$ (see text)

During growth on glucose (Fig. 11) only a few parameters show a significant effect on the model outputs. While Y_{XG}^{Oxid} seems to have a nonlinear effect on the glucose prediction, $r_{G,max}$ has a linear one. The effects of other parameters are mostly nonlinear, as expected given the structure of the model used in the example. The former parameter has also a significant effect on oxygen and biomass, while the latter parameter has a significant effect on ethanol.

With regard to results for a time point during growth on ethanol, it is important to note that Y_{XG}^{Oxid} appears to have a significant effect on the ethanol, oxygen, and biomass predictions, although the glucose has been depleted. This may reflect the impact of the biomass concentration (originated during the prior growth on glucose) on the total amount of ethanol produced, as well as its consumption and the consumption of oxygen for the observed time point.

There is good agreement of the results of the Morris analysis with the previously presented SRC ranking obtained for the linear regression method. In Figs. 11 and 12, the parameters most distant from the wedge are the parameters ranked as the most influential on the model outputs (Table 9a, b).

Table 9a Ranking and SRC value of the model parameters for each model output, for a time point during the exponential growth phase on glucose

t = 17.2 h	Glucose		Ethanol		Oxygen		Biomass	
	SRC	Rank	SRC	Rank	SRC	Rank	SRC	Rank
$r_{G,\max}$	−0.0089	2	0.7423	1	−0.0858	6	0.1111	4
$r_{O,\max}$	0.0006	8	−0.1591	3	−0.5768	2	−0.0129	10
$r_{E,\max}$	0.0002	11	−0.0837	5	0.0210	10	−0.0400	7
Y_{xg}^{Oxid}	−0.0107	1	−0.0777	6	−0.7884	1	0.9746	1
Y_{xg}^{Red}	0.0060	3	0.0467	8	−0.1599	4	0.1697	2
Y_{Xe}	0.0008	7	0.0070	10	0.0452	7	−0.0643	5
K_G	0.0058	4	−0.4819	2	−0.0301	8	0.0328	8
K_O	0.0042	5	0.0279	9	0.1142	5	−0.1664	3
K_E	−0.0005	9	0.0756	7	0.0027	11	−0.0103	11
K_i	0.0013	6	−0.1324	4	0.0273	9	−0.0420	6
k_La	−0.0005	10	−0.0070	11	0.2380	3	0.0170	9

Table 9b Ranking and SRC value of the model parameters for each model output, for a time point during the exponential growth phase on ethanol

t = 27.2 h	Glucose		Ethanol		Oxygen		Biomass	
	SRC	Rank	SRC	Rank	SRC	Rank	SRC	Rank
$r_{G,\max}$			−0.0253	11	−0.0003	11	0.0356	8
$r_{O,\max}$			−0.1310	3	−0.2484	3	0.0409	5
$r_{E,\max}$			−0.1700	2	−0.2534	2	−0.0208	9
Y_{xg}^{Oxid}			−0.9504	1	−0.6101	1	0.9812	1
Y_{xg}^{Red}			−0.0653	7	0.0936	4	0.1157	3
Y_{xe}			0.0281	10	0.0105	9	−0.0357	6
K_G			−0.1195	4	−0.0710	6	0.0997	4
K_O			0.1150	5	0.0748	5	−0.1279	2
K_E			0.0394	8	−0.0010	10	−0.0120	11
K_i			−0.1110	6	0.0596	7	−0.0356	7
k_Lan			−0.0284	9	0.0560	8	0.0202	10

Values corresponding to the prediction of glucose are not shown, as the linear regression was found not to be suitable for this time point and model output ($R^2 < 0.7$)

3 Discussion

A mechanistic model of glucose oxidation by *Saccharomyces cerevisiae* has been taken as an example and has been analyzed rigorously with a number of methods. The chosen case study is purposely kept relatively simple in order to better illuminate how the different methods work and what kind of information is gained in each step. In practice, the presented analysis methods are generic and can be applied to a wide range of process models to assess their reliability. Each step of the analysis has been commented in detail already. However, one thing that cannot be emphasized enough is the importance of collecting proper datasets: biological replicates (duplicate/triplicate fermentations) but also sample replicates are needed to know the error of the measurements. If the quality of the collected data is not sufficiently high, this might later raise severe questions about the reliability of the resulting model.

Assuming that a decision has been taken to develop a mechanistic model of a pharmaceutical production process, or one of its unit operations, one could, of course, wonder how such a model can be established, and how it can support PAT objectives. In general, construction of a mechanistic model is considered time-consuming, which may explain why data-driven models and chemometrics have been more popular than mechanistic approaches, despite the PAT guidance. However, during the past 5 years, this situation has already changed considerably for small-molecule drug substances [4]. According to us, the tools presented here can be helpful in setting up and structuring the model equations in an efficient way, for example, by making use of matrix notation, which can facilitate transfer of the model equations between different users. Such sharing of modeling knowledge is essential in multidisciplinary process development. As discussed by Sin et al [14], a significant part of such a model matrix can be transferred from one system to a second or a third, which undoubtedly makes the whole model-building exercise more efficient.

Finally, we would also like to emphasize that one should move ahead in small steps when constructing a mechanistic model of a process or unit operation. One should rather start with a smaller model with limited scope, for example, an unstructured model [21]. Such a model could then be gradually extended with more detail, while the development of the production process at laboratory and pilot scale is ongoing. The model analysis tools presented here can then be used in the different stages of the model-building as continuous quality checks of the model.

Once a model is considered ready for use, a first application that is relevant for such a model is to use simulations to propose more informative experiments leading to more accurate estimation of the model parameters, for example, by applying optimal experimental design (OED) [22]. Furthermore, the mechanistic model can be helpful in process design, optimization, and in development of suitable control strategies [23]. The latter applications of the model are essential for implementing PAT principles, and can potentially contribute to more efficient process development, replacing data collection and experiments by simulations whenever possible.

4 Conclusions

Mechanistic models form an attractive alternative for structuring and representing process knowledge, also for production processes in biotechnology. The reliability of such models can be confirmed by performing identifiability, uncertainty, and sensitivity analyses on the resulting model. Tools for performing such analyses can be considered as standard engineering tools and are increasingly available on different software platforms. Once it can be documented that the model is reliable, it can be used for design of experiments, for process optimization and design, and for investigating the usefulness of novel control strategies.

Acknowledgments The Danish Council for Strategic Research is gratefully acknowledged for financial support in the frame of project number 09-065160.

References

1. US Food and Drug Administration (FDA) (2004) PAT guidance
2. Bhatia T, Biegler LT (1996) Dynamic optimization in the design and scheduling of multiproduct batch plants. Ind Eng Chem Res 35:2234–2246
3. Gernaey KV, Cervera-Padrell AE, Woodley JM (2012) A perspective on PSE in pharmaceutical process development and innovation. Comput Chem Eng 42:15–29
4. Nielsen J, Villadsen J (1992) Modeling of microbial kinetics. Chem Eng Sci 47:4225–4270
5. Teusink B, Smid EJ (2006) Modelling strategies for the industrial exploitation of lactic acid bacteria. Nat Rev Microbiol 4:46–56
6. Gernaey KV, Eliasson Lantz A, Tufvesson P, Woodley JM, Sin G (2010) Application of mechanistic models to fermentation and biocatalysis for next generation processes. Trends Biotechnol 28:346–354
7. Villadsen J, Nielsen J, Lidén G (2011) Bioreaction engineering principles (3rd ed). Springer, New York, 561 p, ISBN 978-1-4419-9687-9
8. Sin G, Woodley JM, Gernaey KV (2009) Application of modeling and simulation tools for the evaluation of biocatalytic processes: a future perspective. Biotechnol Prog 25:1529–1538
9. Vasić-Rački D, Findrik Z, Vrsalović Presečki A (2011) Modelling as a tool of enzyme reaction engineering for enzyme reactor development. Appl Microbiol Biotechnol 91:845–856
10. Sidoli FR, Mantalaris A, Asprey SP (2004) Modelling of mammalian cells and cell culture processes. Cytotechnology 44:27–46
11. Sin G, Gernaey KV, Eliasson Lantz A (2009) Good modelling practice (GMoP) for PAT applications: propagation of input uncertainty and sensitivity analysis. Biotechnol Prog 25:1043–1053
12. Sonnleitner B, Käppeli O (1986) Growth of *Saccharomyces cerevisiae* is controlled by its limited respiratory capacity: formulation and verification of a hypothesis. Biotechnol Bioeng 28:927–937
13. Ferrer-Miralles N, Domingo-Espín J, Corchero JL, Vázquez E, Villaverde A (2009) Microbial factories for recombinant pharmaceuticals. Microb Cell Factories 8:17
14. Sin G, Ödman P, Petersen N, Eliasson Lantz A, Gernaey KV (2008) Matrix notation for efficient development of first-principles models within PAT applications: integrated modeling of antibiotic production with Streptomyces coelicolor. Biotechnol Bioeng 101:153–171
15. Roels JA (1980) Application of macroscopic principles to microbial metabolism. Biotechnol Bioeng 22:2457–2514
16. Esener AA, Roels J, Kossen NWF (1983) Theory and applications of unstructured growth models: kinetic and energetic aspects. Biotechnol Bioeng 25:2803–2841
17. Holmberg A (1982) On the practical identifiability of microbial growth models incorporating Michaelis–Menten type nonlinearities. Math Biosci 62:23–43
18. Brun R, Kuhni M, Siegrist H, Gujer W, Reichert P (2002) Practical identifiability of ASM2d parameters—systematic selection and tuning of parameter subsets. Water Res 36:4113–4127
19. Carlquist M, Lencastre Fernandes R, Helmark S, Heins A-L, Lundin L, Sørensen SJ, Gernaey KV, Eliasson Lantz A (2012) Physiological heterogeneities in microbial populations and implications for physical stress tolerance. Microb Cell Factories 11:94

20. McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245
21. Bailey JE (1998) Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. Biotechnol Prog 14:8–20
22. Baltes M, Schneider R, Sturm C, Reuss M (1994) Optimal experimental design for parameter estimation in unstructured growth models. Biotechnol Prog 10:480–488
23. Singh R, Gernaey KV, Gani R (2009) Model-based computer aided framework for design of process monitoring and analysis systems. Comput Chem Eng 33:22–42