Promega

Don’t settle for just a supplier. Find a custom manufacturing partner.

Your specifications. Your format.

Our scientists waiting to help.

Let’s TALK CUSTOM

40+ years CUSTOM EXPERIENCE

5 GLOBAL Mfg Sites

PRIMARY MANUFACTURER

Selecting a supplier for your biotechnology and biopharma products can be a challenge—especially one who can adapt to your specific needs. Don’t settle for just a supplier. Instead, partner with Promega and work with a custom manufacturer willing to provide you with the scientific expertise, ongoing technical support and quality standards that support your success.

Learn more with our video:
promega.com/CustomProcess

© 2019 Promega Corporation. All Rights Reserved. 53350832

Received: 4 April 2019 | Revised: 20 June 2019 | Accepted: 19 July 2019

DOI: 10.1002/bit.27125

ARTICLE

BIOTECHNOLOGY BIOENGINEERING WILEY

Predicting industrial-scale cell culture seed trains–A Bayesian framework for model fitting and parameter estimation, dealing with uncertainty in measurements and model parameters, applied to a nonlinear kinetic cell culture model, using an MCMC method

Tanja Hernández Rodríguez[1] | Christoph Posch[2] | Julia Schmutzhard[2] | Josef Stettner[2] | Claus Weihs[3] | Ralf Pörtner[4] | Björn Frahm[1]

[1]Biotechnology & Bioprocess Engineering, Ostwestfalen-Lippe University of Applied Sciences and Arts, Lemgo, Germany

[2]Novartis Technical Research & Development, Sandoz GmbH, Langkampfen, Austria

[3]Faculty of Statistics, TU Dortmund University, Dortmund, Germany

[4]Institute for Bioprocess- and Biosystems Engineering, Hamburg University of Technology, Germany

Correspondence
Björn Frahm, Biotechnology & Bioprocess Engineering, Department of Life Science Technologies, Ostwestfalen-Lippe University of Applied Sciences and Arts, Campusallee 12, 32657 Lemgo, Germany.
Email: bjoern.frahm@th-owl.de

Abstract

For production of biopharmaceuticals in suspension cell culture, seed trains are required to increase cell number from cell thawing up to production scale. Because cultivation conditions during the seed train have a significant impact on cell performance in production scale, seed train design, monitoring, and development of optimization strategies is important. This can be facilitated by model-assisted prediction methods, whereby the performance depends on the prediction accuracy, which can be improved by inclusion of prior process knowledge, especially when only few high-quality data is available, and description of inference uncertainty, providing, apart from a "best fit"-prediction, information about the probable deviation in form of a prediction interval. This contribution illustrates the application of Bayesian parameter estimation and Bayesian updating for seed train prediction to an industrial Chinese hamster ovarian cell culture process, coppled with a mechanistic model. It is shown in which way prior knowledge as well as input uncertainty (e.g., concerning measurements) can be included and be propagated to predictive uncertainty. The impact of available information on prediction accuracy was investigated. It has been shown that through integration of new data by the Bayesian updating method, process variability (i.e., batch-to-batch) could be considered. The implementation was realized using a Markov chain Monte Carlo method.

KEYWORDS

Bayes, CHO cell culture, Markov chain Monte Carlo (MCMC), seed train prediction, uncertainty

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2019 The Authors. *Biotechnology and Bioengineering* Published by Wiley Periodicals, Inc.

1 | INTRODUCTION

In bioprocessing, mathematical modeling, statistical data analysis, and IT-supported tools have become important instruments within the framework of process design, optimization, and control. They are also part of the process analytical technology (PAT) regulatory initiative for building in quality to pharmaceutical manufacturing, defined by the United States Food and Drug Administration. PAT methods are playing an important role, for example, in cell culture upstream processes for the production of biopharmaceuticals (Glassey et al., 2011). While optimization of the production scale has been in the focus for a long time, it turned out, that the cost- and time-intensive cell proliferation process (the so-called seed train) also has an impact on the success rate in production (Brunner, Fricke, Kroll, & Herwig, 2017). There are various factors that influence the seed train (Le et al., 2012). Examples are selection of vessel and filling volumes of the seed train scales, differences in bioprocess engineering parameters between scales, inoculation cell densities, ratio of fresh medium to passaged medium, substrate and metabolite concentrations, point in time for cell passaging and corresponding viable cell density, apparent growth rate, and viability.

To maintain cell growth and product formation attributes within the seed train, monitoring and optimization strategies are required (Frahm, 2014). Temporal or longer lasting changes in cell behavior can occur, so that the seed train protocol has to be adapted. Also, for new cell lines or new products or the transfer of the process to another production plant, seed train protocols also have to be developed or adapted, keeping in mind the reduction of time and costs during development of those protocols. Another application is to support the selection of the optimal clone for a new process and the development of a suitable seed train protocol.

Model building of dynamic bioprocesses such as cell culture seed trains faces a lot of challenges due to different factors, like limited amount of high-quality experimental data (measurement uncertainty, offline data and large time steps between measurements, etc.), process nonlinearity and the necessity of various model parameters characterizing the bioprocess. As already described in Liu and Gunawan (2017), these factors lead to significant uncertainty in the process model. Furthermore, prediction performance depends on the accuracy of the model, the variability of the biological process, and identifiability of model parameters. Nonidentifiability arises if many different combinations of model parameters can explain the experimental data equally well. The reasons could be that the model contains too many parameters (overparameterization) leading to the problem that noise or random variations in the training data is interpreted and learned as concepts. Consequently, these concepts do not apply to new data, which impede the models' ability to generalize to new data. Different approaches addressing this problem can be found in literature (Ashyraliyev, Fomekong-Nanfack, Kaandorp, & Blom, 2009; Liu & Gunawan, 2017; Sin, Gernaey, & Lantz, 2009).

Often, estimation methods identifying only one set of model parameters based on the available dataset, like the Nelder–Mead simplex algorithm (Press, 1996), are applied. This type of optimization algorithm is a "best fit" estimator, a point estimator, meaning that only one value for each model parameter is identified, leading to one predicted value of the quantity of interest at each time step. No information about the output uncertainty is given this way, and most of these types of optimization algorithms could get stuck in a local minimum. Nevertheless, these methods have turned out to be useful tools, sometimes resulting in fast solutions and they are already implemented in functions (e.g., in Matlab or R), which are easy to apply. They can be combined with statistical methods like Monte Carlo (MC) simulation, sensitivity, uncertainty, and/or identifiability analysis, in order simulate output uncertainty and to gain more information of the process (Hines, 2015; Price, Nordblad, Woodley, & Huusom, 2013; Raue et al., 2009; Sin et al., 2009).

However, in many cases, there are only few data available for model building and parameter estimation (e.g., when planing a new production), but frequently there is some knowledge about the organism or process from literature or expert knowledge. It is desirable to quantify this information and include it in the model building process. Within a Bayesian context, this kind of knowledge is expressed by probability statements and it is combined with the available data, leading to a whole set of probable model parameter values for each model parameter. This procedure is also called Bayesian parameter estimation and the numerical implementation can be performed through a Markov chain Monte Carlo (MCMC) procedure. There are many ways of numerically implementing this method and some examples in the field of biochemical engineering can be found in Galagali and Marzouk (2015); Liu and Gunawan (2017); Vrugt (2016); and Xing, Bishop, Leister, and Li (2010). Another similar technique to simulate process dynamics under uncertainty are Gaussian processes, but they were not subject to investigation in this work.

In this work, a Bayesian approach, facing the above-mentioned challenges in model building and parameter estimation dealing with uncertainties and eventually lack of data, is applied to seed train prediction of an industrial Chinese hamster ovarian (CHO) cell culture process. It is shown in which way sources of uncertainty as well as prior knowledge (from experts or literature) can be considered leading to predictions including inference uncertainty. Bayesian parameter estimation also provides a framework for detection of nonidentifiabilites but as this is not the focus of this work, we refer to Raue, Kreutz, Theis, and Timmer (2013). Numerical implementation of the Bayesian approach was carried out via an MCMC procedure using an adaptive single component metropolis algorithm.

A mechanistic model, similar to Frahm (2014) and Kern, Platas-Barradas, Pörtner, and Frahm (2016) is applied for this approach. These types of models have gained renewed attention because they can be considered as a structured representation of the available process knowledge (Glassey et al., 2011; Kroll, Hofer, Stelzer, & Herwig, 2017; Möller & Pörtner, 2017; Sanderson, Phillips, & Barford, 1996). They have been used for development of predictive control strategies, for example, to ensure high batch-to-batch reproducibility in animal suspension cell cultures (Aehle et al., 2012) and for the design of cell culture fed-batch control (Frahm et al., 2002).

The industrial CHO cell culture process is used to illustrate in which way inference uncertainty can be derived and with which accuracy individual seed train scales and the whole seed train can be predicted, depending on the available information.

2 | MATERIALS AND METHODS

2.1 | Investigated suspension cell culture process

In this contribution, the subject of investigation is an industrial CHO cell culture process containing a seed train comprising five shake flask scales and three bioreactor scales as well as the production scale, whereby the focus lies on the the bioreactor part of the seed train, which is composed of bioreactor 1 (N-3, 40 L), bioreactor 2 (N-2, 320 L) and bioreactor 3 (N-1, 2,160 L). From experimental data (offline measurements), taken once a day, time profiles for viable cell density X_v, viability Via, concentrations of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm} have been used. In this work, data from 20 cultivations from six campaigns with cultivation times between 72 and 96 hours per scale (meaning 4–5 measurement time points per scale) were divided into 10 seed trains for training and 10 seed trains for testing, choosing randomly one or two cultivations for training and one or two for testing per campaign. Additional datasets have been generated for modeling purposes. Therefore, 12 cultivations in four flask scales (three cultivations each) having filling volumes of 40, 70, 300, and 1,500 ml were provided. They cover cultivation time spans of 264 hr (11 days) each, meaning that the stationary and death phases were also included. All datasets are labeled and listed in Table 1 to assign them correctly in this work.

2.2 | Cultivation conditions and analytics

Cell cultivation was carried out using a CHO cell line for the production of a therapeutic recombinant protein (cell line and product are not further specified due to confidential reasons). Process conditions, which were the same for all investigated seed train cultivations, are listed in Table 1. Samples were taken once a day. Viable cell concentration and viability were measured using the Vi-CELL cell viability analyzer from Beckman Coulter. Glucose, glutamine, lactate, and ammonia were determined by a Nova Bioprofile 100+ Analyzer.

2.3 | Data cleansing/preparation

Data cleansing and preparation was performed by handling missing data. Within the parameter estimation process initial concentration values are required for solving the ordinary differential equation system (the model), but in some cultivation datasets, there are one or two missing initial concentrations. Concerning datasets of 20 seed trains, with three bioreactor scales and six state variables each, 16% of the initial concentrations are missing in total (viable cell concentration 0%, viability 0%, glucose 0%, glutamine 23%, lactate 72%, and ammonia 0%). If initial concentrations are missing at the beginning of bioreactor scale 1, the relevant quantity is replaced by the mean of initial concentration values of training datasets. This decision is based on the fact that the same cultivation conditions are intended for each cultivation. If initial concentrations of bioreactor scale 2 or 3 are missing, then they are calculated based on the concentrations at the end of the previous scales and the volumes of the previous and the current scale.

TABLE 1 List of available data, some used for training, and other for testing, containing the following abbreviations: Systems (SF, shake flask; BR, bioreactor; ST, seed train), initial filling volumes (Volume) and cultivation labels (Label). Since process data from 6 campaigns were considered, they were labeled by C1 (campaign 1) to C6 (campaign 6)

System	Volume (L)	Cultivation labels	Controlled process parameters	Usage
Training data from development				
SF	0.04	SF1.1, SF1.2, SF1.3	Temperature, CO_2, humidity, seeding VCD	Training
SF	0.07	SF2.1, SF2.2, SF2.3		Training
SF	0.3	SF3.1, SF3.2, SF3.3		Training
SF	1.5	SF4.1, SF4.2, SF4.3		Training
Training and test data from the process (from 6 campaigns)				
BR	40	R1.1, … , R1.10 (10 sets)	Temperature, pH, DO (disolved oxygen)	Training
BR	40	R1.11, … , R1.20 (10 sets)	total gas flow stirrer speed, seeding/	Testing
BR	320	R2.1, … , R2.10 (10 sets)	transfer VCD, pressure	Training
BR	320	R2.11, … , R2.20 (10 sets)		Testing
BR	2160	R3.1, … , R3.10 (10 sets)		Training
BR	2160	R3.11, … , R3.20 (10 sets)		Testing
Cultivations ending on … belong to campaign (C.): .1, 0.2, 0.11, 0.12 (C1), 0.3, 0.4, 0.13, 0.14 (C2), 0.5, 0.6, 0.15, 0.16 (C3), 0.7, 0.8, 0.17, 0.18 (C4), 0.9, 0.19 (C5), 0.10, 0.20 (C6) Seed trains (comprising BR1 [40 L], BR2 [320 L], and BR3 [2,160 L]):				
ST1, ST2 (C1), ST3, ST4 (C2), ST5, ST6 (C3), ST7, ST8 (C4), ST9 (C5), ST10 (C6)				Training
ST11, ST12 (C1), ST13, ST14 (C2), ST15, ST16 (C3), ST17, ST18 (C4), ST19 (C5), ST20 (C6)				Testing
(In brackets the campaigns the seed trains belong to)				

2.4 | Cell culture model

The applied kinetic model is based on modifications of previous model variations published in Frahm et al. (2002) and Kern et al., 2016. Differential algebraic equations (Equation 1), containing six mostly Monod-type algebraic equations (description of growth rate, death rate, substrate uptake, and metabolite production kinetics) and 17 model parameters are describing cell culture dynamics of total and viable cell density, X_t and X_v, as well as concentrations of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm}. All these variables and model parameters are listed in Table 2 including unit and description. Volume changes because of sampling were considered by the sampling flow rate F_{Sample} (describing the sample volume expressed as an effluent flow rate [with negative values] during the sampling period of time). The titer dynamics are not considered in this example as it was measured only in production scale, but not during the seed train.

$$
\begin{aligned}
\frac{dX_v}{dt} &= Xv\cdot(\mu - \mu_d)\\
\frac{dX_t}{dt} &= Xv\cdot\mu - K_{Lys}\cdot(X_t - X_v)\\
\frac{dc_{Glc}}{dt} &= -X_v\cdot q_{Glc}\\
\frac{dc_{Gln}}{dt} &= -X_v\cdot q_{Gln}\\
\frac{dc_{Lac}}{dt} &= X_v\cdot q_{Lac}\\
\frac{dc_{Amm}}{dt} &= X_v\cdot q_{Amm}\\
\frac{dV}{dt} &= F_{Sample}\\
\mu &= \mu_{max}\cdot\frac{c_{Glc}}{(c_{Glc}+K_{S,Glc})}\cdot\frac{c_{Gln}}{(c_{Gln}+K_{S,Gln})}\cdot\left(1-\frac{t}{t_{Lag}}\right)\cdot a_{Lag}\cdot\mu_{max}\\
\mu_d &= \mu_{d,min} + \mu_{d,max}\cdot\frac{K_{S,Glc}}{(K_{S,Glc}+c_{Glc})}\cdot\frac{K_{S,Gln}}{(K_{S,Gln}+c_{Gln})}\\
q_{Glc} &= q_{Glc,max}\cdot\frac{c_{Glc}}{(c_{Glc}+k_{Glc})}\\
q_{Gln} &= q_{Gln,max}\cdot\frac{c_{Gln}}{(c_{Gln}+k_{Gln})}\\
q_{Lac} &= Y_{Lac/Glc}\cdot q_{Glc}\cdot\frac{c_{Glc}}{c_{Lac}} - q_{Lac,uptake}\cdot\frac{(\mu_{max}-\mu)}{\mu_{max}}\\
q_{Amm} &= Y_{Amm/Gln}\cdot q_{Gln}\cdot\frac{c_{Gln}}{c_{Amm}} - K_{Amm}\cdot q_{Amm,uptake,max}\cdot\frac{(\mu_{max}-\mu)}{\mu_{max}}
\end{aligned}
\tag{1}
$$

The specific growth rate includes a term for lag phase description where t_{Lag} stands for the duration of the lag phase. $a_{Lag} \in [0, 1]$ describes by which percentage growth rate is decreased in the beginning of the lag phase and for $t > t_{Lag}$. The specific death rate contains constant minimum and maximum death rates as well as dependencies on glucose and glutamine concentration (similar to Frahm, 2014). Substrate uptake rates are expressed similar to substrate uptake rates presented in Frahm (2014) and Kern et al. (2016), describing a high glucose uptake at high glucose concentrations and low glucose uptake at low glucose concentrations and analogously for glutamine. Also metabolic production rates are expressed similar to substrate uptake rates presented in Frahm (2014) but with additional terms for metabolite uptake, followed by renewed production (in case of ammonia) at the end of the death phase, with

if $c_{Glc} > c_{Lac}$: $q_{Lac,uptake} = 0$
else: $q_{Lac,uptake} = q_{Lac,uptake,max}$ (constant)
if ($c_{Gln} > c_{Amm}$): $K_{Amm} = 0$
else if ($c_{Gln} \le c_{Amm}$) and ($(\mu - \mu_d) > 0.001$): $K_{Amm} = 1$
else: $K_{Amm} = -k_{Amm}$ (constant)

The ode23 function of MATLAB version 2017b (Matlab, 2017) was used for numerical computation.

TABLE 2 Modeled variables and model parameters included in the underlying model (symbols, units, and descriptions)

Variable/ parameter	Unit	Description
X_t	cell/L	Total cell density
X_v	cell/L	Viable cell density
c_{Glc}	mmol/L	Glucose concentration
c_{Gln}	mmol/L	Glutamine concentration
c_{Lac}	mmol/L	Lactate concentration
c_{Amm}	mmol/L	Ammonia concentration
V	L	Volume
μ_{max}	hr^{-1}	Maximum cell-specific growth rate
$K_{S,Glc}$	mmol/L	Monod kinetic constant for glucose
$K_{S,Gln}$	mmol/L	Monod kinetic constant for glutamine
a_{Lag}	–	Correction factor for lag phase
t_{Lag}	hr	Duration of la phase
$\mu_{d,min}$	hr^{-1}	Minimum cell-specific death rate
$\mu_{d,max}$	hr^{-1}	Maximum cell-specific death rate
K_{Lys}	hr^{-1}	Cell lysis constant
$q_{Glc,max}$	$mmol\cdot cell^{-1}\cdot hr^{-1}$	Maximum cell-specific glucose uptake rate
k_{Glc}	mmol/L	Monod kinetic constant for glucose uptake
$q_{Gln,max}$	$mmol\cdot cell^{-1}\cdot hr^{-1}$	Maximum cell-specific glutamine uptake rate
k_{Gln}	mmol/L	Monod kinetic constant for glutamine uptake
$Y_{Lac/Glc}$	mmol/mmol	Kinetic production constant for lactate
$q_{Lac,uptake,max}$	$mmol\cdot cell^{-1}\cdot hr^{-1}$	Cell-specific maximum lactate uptake rate
$Y_{Amm/Gln}$	mmol/mmol	Kinetic production constant for ammonia
$q_{Amm,uptake,max}$	$mmol\cdot cell^{-1}\cdot hr^{-1}$	Cell-specific maximum ammonia uptake rate
k_{Amm}	–	Correction factor for ammonia uptake

2.5 | Bayesian parameter estimation and inference

The goal of Bayesian parameter estimation is to compute a maximum a posteriori point (MAP) estimate of each unknown model parameter (e.g., maximum growth rate μ_{max}) as well as the corresponding probability distribution (posterior distribution), describing how probable it is that certain parameter values are adopted, based on the measured data and prior knowledge. These estimates and distributions can then be used for prediction of new observations (e.g., viable cell density X_v; Gelman et al., 2013, chap. 1). Bayesian parameter estimation and prediction can be divided into the following main tasks, which will be explained afterward:

- Step 1: Quantification of prior knowledge including uncertainties
- Step 2: Bayesian parameter estimation/determination of posterior distributions
- Step 3: Prediction including credible intervals
- Step 4: Bayesian updating (if desired and additional data is available)

2.5.1 | Quantification of prior knowledge

In a *first step*, quantification of prior knowledge including uncertainties through probability distributions is required. These probability distributions are called *prior distributions*. There are different types of prior distributions which can be chosen, according to the available prior information. In the present work, there are mainly two sources of uncertainties considered, the uncertainty in model parameters and the uncertainty in initial concentration values, resulting from measurement deviations (see Figure 1). A gamma distribution was chosen to describe the existing prior knowledge including the above-mentioned uncertainties. This assumption is based on the fact that the considered random variables can only adopt positive values, and, furthermore, the gamma distribution is well suited for representing the realistic range based on the available priori knowledge. It is defined by the parameters α (shape) and λ (rate). The expected value and variance are calculated by $E(Y) = \frac{\alpha}{\lambda}$, $\mathrm{Var}(Y) = \frac{\alpha}{\lambda^2}$. For estimating expected value and variance, the two equations can be solved for α and λ (method of moments) and consequently, measures of location and variation of the distribution can be computed by

$$\alpha = \frac{E(Y)^2}{\mathrm{Var}(Y)} \quad \text{and} \quad \lambda = \frac{\mathrm{Var}(Y)}{E(Y)}. \tag{2}$$

More details on different types of prior distributions are given in the Supporting Information Material.

2.5.2 | Bayesian parameter estimation/determination of posterior distributions

In a *second step*, Bayesian parameter estimation using prior probabilities and experimental data has to be performed, obtaining *posterior parameter distributions*. The key element is the Bayes theorem, which is a theorem for the computation of conditional probabilities. Since in practice the applied mathematical models are complex and high dimensional, the calculation of the posteriori parameter distributions turns out to be a nontrivial tasks. But numerical solutions can be computed by application of MCMC methods. The concept of MCMC

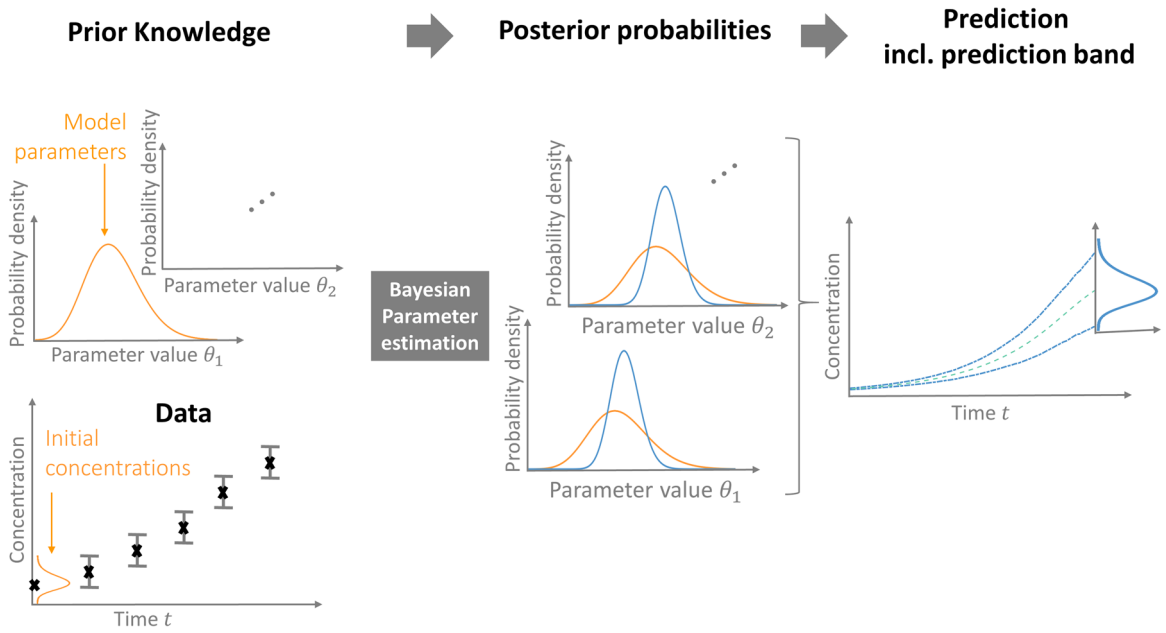

FIGURE 1 Propagation of uncertainty. Uncertainty in model parameters and uncertainty resulting from measurement deviations are considered and a Bayesian approach, having the Bayes theorem as a key element, was applied to propagate these uncertainties, to estimate model parameters, and to include the information of uncertainty in the prediction of the interesting quantities in form of prediction intervals forming a prediction band [Color figure can be viewed at wileyonlinelibrary.com]

simulation is to create a random process whose stationary distribution is the specified target distribution and to run the simulation long enough that the distribution of the current draws is close enough to this stationary distribution (Gelman et al., 2013, chap. 11). Different types of algorithms realizing this principle exist, whereby the single component Metropolis-Hastings algorithm was applied in this work (Gilks et al., 1998), which has the posterior samples of the model parameters as its output, which represent the posterior distributions of the model parameters to be estimated. The sample size should be chosen large enough so that the Markov Chain (MC) standard error is less than 5% (more details on convergence diagnostics are given in the Supporting Information Material).

2.5.3 | Prediction including credible intervals

In a *third step*, predictions based on the obtained posterior parameter distributions can be performed, using MC simulations. As a result the *posterior predictive distributions* of the variables included in the model (or even of functions built with these variables) are obtained. In the case of a dynamic process model over time, a posterior prediction distribution is obtained for each variable at each point in time within the defined time span. *Credible intervals* (also called *prediction intervals* or *prognostic intervals*) of nonobserved values can be computed using the posterior predictive distributions. A credible interval can be described as a coverage interval that contains the set of true values of a quantity with a given probability, based on available information. For example, they can be computed using quantiles of the posterior predictive distributions at different points in time, leading to prediction bands over the considered time span (see Figure 1).

2.5.4 | Bayesian updating

As a *fourth step*, (if additional data is provided), Bayesian updating can be executed, which is an important characteristic of Bayesian statistics. It is the ability to learn from new data through adding information to the present knowledge and thus, to update the current state of information.

This is realized by repetition of Steps 2 and 3 using the current posterior distributions as new prior distributions and executing MCMC simulation to obtain new posterior parameter distributions. Simply spoken Bayesian updating is performed by "*taking the posterior from today as prior from tomorrow*." This is described in literature by terms like *Bayesian updating*, *Bayesian learning*, or the *sequential nature of Bayes* (Luce, Anthony, & Dennis, 2003; O'Hagan, 2008).

More detailed information, such as formulas and implementation of the adaptive single based Metropolis-Hastings algorithm, is provided as Supporting Information Material.

2.6 | Evaluation

To evaluate the prediction performance regarding accuracy and precision, three criteria are presented in this work. The first criterion quantifies the amount of predictive uncertainty, based on the available information and is expressed in this work by the relative half bandwidth of the prediction interval at a specified point in time. Supposing that y_{pred} is the posterior predictive sample and $q_{0.975}$ its 97.5% quantile and $q_{0.025}$ its 0.25% quantile, then:

$$\text{Half bandwidth} = \frac{1}{2}\frac{q_{0.975} - q_{0.025}}{\bar{y}_{\text{pred}}} \cdot 100. \tag{3}$$

Simply spoken, this measure describes how well the prediction can be bounded (precision) or how much deviation from the predicted value is expected, based on the available information. This score is computed before test data are used for evaluation and can adopt values between 0 and 100, where a low value is desired because it stands for a high prediction precision. The other two criteria, within band score and relative error describe how good the prediction performed, thus describing the prediction accuracy. They are measures for evaluation of the prediction after adding the test data.

Accuracy score 1, within band score: Percentage of subsequently added test data falling within prediction band in relation to the total number k of test data. Values between 0 and 100 can be adopted, where a high value stands for a high prediction accuracy.

$$\text{Within band score} = \left(1 - \frac{\text{No. of predicted values out of band}}{\text{k}}\right) \cdot 100. \tag{4}$$

Accuracy score 2, the relative error: The relative deviation between predicted values (values with maximum posterior probability for model parameters (MAP estimate)) and subsequently added test data is described by the relative error in % with values between 0 and 100, where a low value stands for a high accuracy.

$$\text{Rel. error} = \frac{1}{k} \cdot \sum_{i=1}^{k} \left| \frac{(y_{i,\text{exp}} - y_{i,\text{pred}})}{y_{i,\text{exp}}} \right| \cdot 100, \tag{5}$$

with k number of measurements.

All criteria can be computed for one or more quantities. Furthermore, the coefficient of variation is used for the quantification of uncertainty. The coefficient of variation of a sample y is calculated by

$$cv = \frac{\text{var}(y)}{\text{mean}(y)} \cdot 100. \tag{6}$$

3 | RESULTS AND DISCUSSION

This section shows how an industrial cell culture seed train, that is described by a mechanistic model, is predicted (simulated) using Bayesian parameter estimation. The prediction is complemented by corresponding prediction intervals describing the expected deviation from the predicted values based on the available information. Furthermore, the performance of the predictions will be analyzed concerning prediction precision and accuracy. Moreover, it is investigated, how taking additional data into account improves the prediction.

First, it is demonstrated in Section 3.1 how prior knowledge was quantified in form of probability distributions. Afterwards, Bayesian parameter estimation by determination of posterior distributions for the model parameters as well as a Bayesian updating step is explained and illustrated (see Section 3.2).

Thereafter, the model parameter distributions were used for prediction and the prediction performance based on available information/knowledge was evaluated using test data. First, these investigation were realized for single bioreactor scales (see Section 3.3). Afterwards the whole seed train comprising the mentioned three consecutive bioreactor scales (before production) was predicted based on available information. The corresponding results are presented and discussed in Section 3.4.

3.1 | Prior knowledge

For all quantities, which are assumed as random variables (meaning that they were considered including uncertainties), a prior probability distribution, expressing the prior knowledge about the possible model parameter values, had to be defined. In which way prior knowledge (based on literature/expert knowledge/historical or training data analysis; before parameter estimation) of model parameters as well as prior information about starting concentration values were quantified is described below.

3.1.1 | Prior knowledge of model parameters

To characterize the prior distributions of model parameters, means and coefficients of variation, representing the uncertainty, were defined using information from literature (Kern et al. 2016), information from analysis of training data and information from expert knowledge (from industry concerning the investigated process and from academia relying on experiences with cultivation of similar cell lines). Training data of four flask scale cultivations, one from 40 (SF 1.1), 70 (SF 2.1), 300 (SF 3.1), and 1,500 ml (SF 4.1) (labels according to Table 1), respectively were analyzed using offline measurements of viable cell density X_v and viability as well as concentrations of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm}. Measurements were taken once a day (except on weekends) at cultivation days 0, 1, 2, 3, 6, 7, 8, 9, 10, 11. This time period covered lag phase, exponential phase, stationary phase, and death phase.

Furthermore, the 17 model parameters were divided into model parameters with fixed values and model parameters to be estimated ("free" parameters) according to the identifiability of model parameters based on the available data. This was carried out because in some cases the available data is not sufficient for identifying every model parameter unambiguously concerning the applied discrepancy function (which is optimized during parameter estimation). Nevertheless, some parameters can be estimated combining training data and expert knowledge. Thus, considering the equation for growth rate μ both parameters describing lag phase, the correction factor a_{Lag} and the duration of lag phase t_{Lag}, were kept as fixed parameter values, while μ_{max}, $K_{S,Glc}$, and $K_{S,Gln}$ were set as "free" parameters. Concerning death rate, the minimum death rate $\mu_{d,min}$ were kept fix, while the maximum death rate $\mu_{d,max}$ were kept "free". Moreover the cell lysis constant K_{Lys} were kept fix, based on Kern et al. (2016). Concerning production and uptake of lactate and ammonia the parameters $q_{Lac,uptake}$ and $q_{Amm,uptake}$ were kept fix because they describe lactate and ammonia uptake at the end of the death rate, which is not relevant during the process. Because they would increase the complexity of parameter estimation, we decided to estimate them once from training data and keep them fix later on whereas k_{Amm} was set "free" as well as kinetic production constants $Y_{Lac/Glc}$ and $Y_{Amm/Gln}$.

In a next step these quantities were used to adopt a gamma distribution for each free model parameter following the methods explained in Section 2.5, meaning that the distribution were calculated according to Equation (2). This form of probability distribution was chosen because of the range of the model parameters (only positive values) and the flexibility of the gamma distribution (more information on gamma distributions can be found in the Supporting Information Material). The assumed means and coefficients of variation, expressing the amount of uncertainty, are listed in Table 3A.

3.1.2 | Prior knowledge of starting concentrations

The starting concentration values (of the modeled time courses, e.g., initial viable cell density $X_{v,0}$, glucose concentration $c_{Glc,0}$, ...) are also set as random variables because it is assumed that the measurement errors have a significant impact on prediction performance. The prior distributions of the measurement error were derived from trend chart data of Vi-Cell in case of viable cell density (sample size $N = 236$, gathered from two instruments) and Nova BioProfile 100+ analyzers for glucose, glutamine, lactate, and ammonia concentrations (sample size $N = 1{,}065$, gathered from three instruments) using resampling techniques (bootstrapping). The corresponding means and standard deviations of coefficients of variation are listed in Table 3B.

In case of viable cell density X_v, the coefficient of variation (cv) from Table 3B was composed of a cv due to uncertainty caused by practical reasons and by a Poisson distributed cv because of uncertainty in cell count data (under the assumption of independent Poisson random variables). Thus, $cv_{total} = cv_{pract} + cv_{pois}$, where $cv_{pois} = \frac{1}{\sqrt{N}}$ depends on the number of cells in the sample volume (here 0.001 ml) and cv_{pract} is computed from $cv_{pract} = cv_{total} - cv_{pois} = (0.047 - \frac{1}{\sqrt{1000}}) \cdot 100 = 1.5$.

3.2 | Bayesian parameter estimation/ determination of posterior distributions and Bayesian updating

After quantification of prior knowledge, Bayesian parameter estimation using the MCMC method described in Section 2.5 was performed, based on the remaining eight datasets from shake flask

TABLE 3 A) Prior knowledge of model parameters expressed by prior means and coefficients of variation (*cv*) in % as well as posterior knowledge expressed by posterior means (based on eight specific experiments performed for modeling; after parameter estimation) and posterior coefficients of variation (*cv*) in %. Prior distributions were also used for sampling starting values of model parameters. B, Prior knowledge concerning measurement errors of initial concentrations expressed by mean and standard deviation (sd) of coefficient of variation

A) Model parameters

		Prior		Posterior	
Parameter	**Unit**	**Mean**	***cv* (%)**	**Mean**	***cv* (%)**
μ_{max}	hr^{-1}	.028	20	.029	9
$K_{S,Glc}$	mmol/L	.03	30	.025	32.2
$K_{S,Gln}$	mmol/L	.03	30	.025	32.8
a_{Lag}	–	.01	30	–	0
t_{Lag}	hr	24	30	–	0
$\mu_{d,min}$	hr^{-1}	.0005		–	0
$\mu_{d,max}$	hr^{-1}	.005	50	.003	63.9
K_{Lys}	hr^{-1}	.001		–	0
$q_{Glc,max}$	$mmol{\cdot}cell^{-1}{\cdot}hr^{-1}$	1.8×10^{-10}	30	1.5×10^{-10}	30.4
k_{Glc}	mmol/L	10	30	8.2	32
$q_{Gln,max}$	$mmol{\cdot}cell^{-1}{\cdot}hr^{-1}$	$.8 \times 10^{-10}$	30	$.6 \times 10^{-10}$	20.7
k_{Gln}	mmol/L	2.5	30	2.4	27
$Y_{Lac/Glc}$	mmol/mmol	.3	30	.2	28.6
$q_{Lac,uptake}$	$mmol{\cdot}cell^{-1}{\cdot}hr^{-1}$	1.2×10^{-11}	30	–	0
$Y_{Amm/Gln}$	mmol/mmol	.7	30	.5	29.4
$q_{Amm,uptake}$	$mmol{\cdot}cell^{-1}{\cdot}hr^{-1}$	4×10^{-12}	30	–	0
k_{Amm}	–	.5	30	.4	32.2

B) Measurement error of initial concentrations

			Prior	
Variable	**Unit**	**Mean of *cv***	**Mean of *cv* (%)**	**SD of *cv***
X_v	cells/L	.047	4.7	.002
c_{Glc}	mmol/L	.097	9.7	.004
c_{Gln}	mmol/L	.084	8.4	.003
c_{Lac}	mmol/L	.079	7.9	.005
c_{Amm}	mmol/L	.068	6.8	.004

cultivations (SF1.2, SF1.3, SF2.2, SF2.3, SF3.2, SF3.3, SF4.2, and SF4.3; labels according to Table 1), for each dataset individually. The starting parameter values were also sampled randomly from these prior distributions.

Every MCMC run results in posterior samples representing the posterior parameter distributions of each free model parameter and then the quantities *mean*, *variance* and *coefficient of variation* (*cv*) were computed for each posterior sample. For each of these quantities the mean was computed to derive one distribution representing all shake flask datasets. The resulting mean and *cv* for every free parameter are listed in Table 3A. The corresponding distributions are illustrated in Figure 2, where the prior distributions "prior 1" are shown in light gray dashed lines and the posterior distributions "posterior 1" from shake flask data are shown in gray dotted lines. (The values on the *y*-axis depend on the parameter values on the *x*-axis and have to fulfill that the integral of a density function integrates to one.)

As soon as new cultivation data are collected, a Bayesian update can be performed using the old posterior "posterior 1" as new prior "prior 2," running a new MCMC run using the new dataset. As an example, cultivation data R1.3 from the smallest bioreactor seed train scale (40 L) was used for this update step, getting a new posterior "posterior 2" (see dark gray solid lines in Figure 2). All three mentioned distributions are shown exemplarily for nine model parameters in Figure 2.

It can be seen that the steps from prior 1 to posterior 1 and from prior 2 (= posterior 1) to posterior 2 lead to more narrow distributions → less uncertainty, more precision) in case of the maximum growth rate μ_{max}. In case of the other parameters $K_{S,Glc}$, $K_{S,Gln}$, k_{Glc}, k_{Gln}, $q_{Glc,max}$, $q_{Gln,max}$, Y_{Lac}, and the mean moved slightly to the left (smaller values) without significant changes in the variance. This can also be conducted from Table 3 by comparing prior and posterior coefficients of variation. For most parameters, the coefficient of variation changed only slightly except for the parameter concerning maximum cell growth, μ_{max}, and the parameter describing the maximum uptake rate of glutamine $q_{Gln,max}$.

It should be mentioned that stronger deviations between prior and posterior means could have been obtained if the data used for Bayesian parameter estimation would have clearly indicated this deviation. But it turned out that the information contained in the prior distributions mostly coincided with the information coming along with the additional data. This is not surprisingly, since a lot of information (prior knowledge from literature and experiences regarding similar cell lines as well as training data from the four shake flask cultivation scales) were already available.

In each MCMC run, convergence diagnostic was applied using visual methods such as history plots and autocorrelation plots. Furthermore, the MC error was controlled and an MC error less than 5% was satisfied in each run, which indicates convergence. This procedure was applied during every parameter estimation (update step) via MCMC.

3.3 | Prediction of a single seed train bioreactor scale based on available information

Following the above presented procedure, a comprehensive study concerning prediction accuracy and prediction precision based on available information from cultivation training data is presented.

For the prediction of one cultivation scale at a time (reactor scale 1 (N-3) = 40 L, 2 (N-2) = 320 L, or 3 (N-1) = 2,160 L), the initial concentrations of a scale are known and the following two sources of

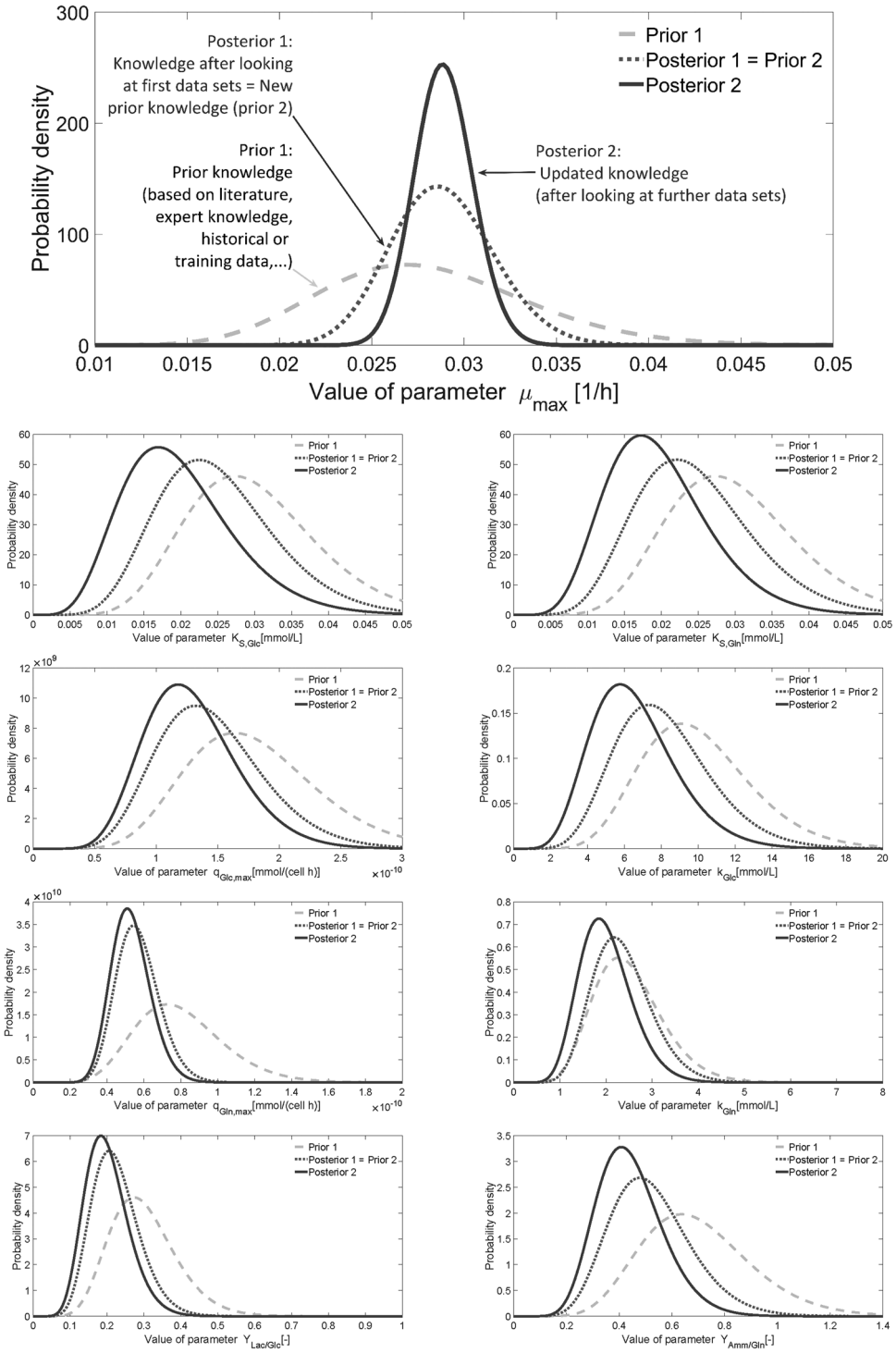

FIGURE 2 Probability distributions describing knowledge about possible values of nine selected model parameters (17 model parameters in total, 6 fixed and 11 to be identified) before and after two updating steps. Prior 1: Based on literature, expert knowledge and previous data analysis, before parameter estimation. Posterior 1 (=Prior 2): Knowledge based on additional specific experiments performed for modeling in shake flasks, after (posterior to) parameter estimation. Posterior 2: Knowledge based on additional data from smallest bioreactor seed train scale (40 L)

uncertainty and their propagation on the predictions are considered: uncertainty of model parameters and uncertainty of measurements.

At first, an example is presented in Subsection 3.3.1, where prediction results based on the information from shake flask data are compared with the prediction results, where also information from another reactor scale training dataset from the same campaign was considered.

Thereafter, the investigation results of prediction performance for a single seed train bioreactor, depending on the available information, evaluated for 10 seed train cultivations are

presented. The labels of the cultivation data used for training and testing correspond to the labels listed in Table 1.

It should be mentioned that evaluation scores, half bandwidth, within band score and rel. error, were first calculated comparing one test dataset at a time (e.g., concerning viable cell density often only four measurements were compared with the predictions at the same points in time, meaning if one measurement falls outside the prediction band, the within band score is reduced to 75%). Afterwards, the average over 10 cultivations was calculated.

3.3.1 | Example: Prediction based on shake flask data versus prediction based on shake flask data and one bioreactor scale

As an example, Figure 3 shows the predicted temporal courses (solid lines) as well as the corresponding 90% prediction bands (dashed lines) for all six observed state variables (viable and total cell concentration, X_v and X_t, concentration of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm},) for the smallest seed train bioreactor scale (R1.13, 40 L). The six diagrams of Figure 3 (above) show the prediction, only based on shake flask

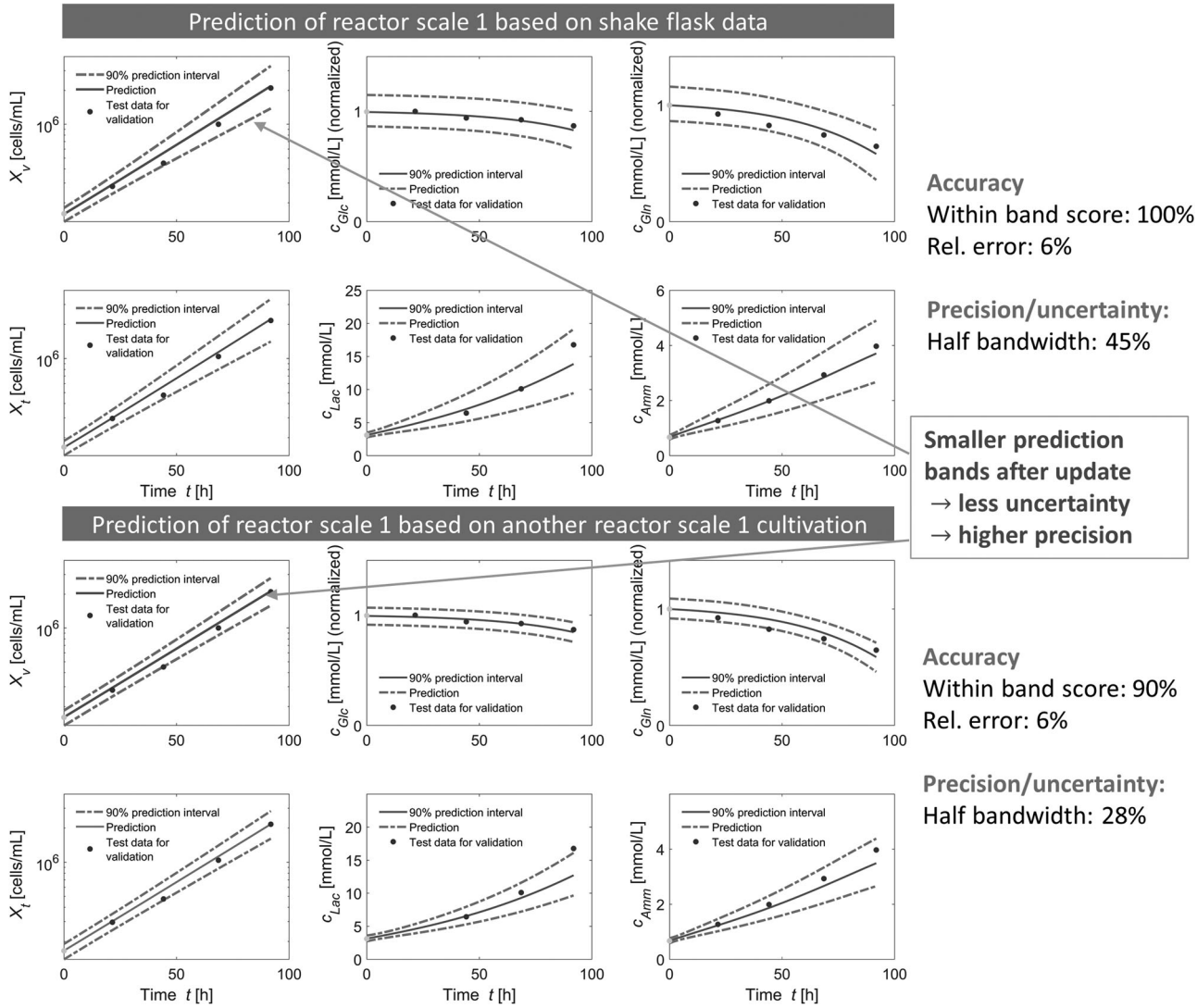

FIGURE 3 Predicted time courses of six state variables, viable and total cell concentration X_v and X_t, concentration of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm} as well as performance measures of prediction, for the smallest seed train bioreactor scale (40 L filling volume). As measures of accuracy the within band score (percentage of test data falling within prediction band) and the rel. error (the relative deviation between predicted values and subsequently added test data) are presented. The amount of uncertainty (only presented by numbers on the right) is expressed by the relative half bandwidth, that is, half width of prediction interval of viable cell density at the last time point of each scale, describing how many deviation from the predicted value is expected. The prediction was performed given data from shake flask scales (six diagrams above) and from data of another cultivation from the same campaign in the same bioreactor scale (six diagrams below)

scale data (SF1.1-SF4.3). This means that the posterior 1 distributions of the model parameters (compare with Figure 2) were used for prediction. The six diagrams in Figure 3 (below) show the prediction based on the updated information, based on shake flask data and another cultivation from the same campaign in the same bioreactor scale (here R1.3, 40 L). This means that the posterior 2 distributions (compare with Figure 2) were used for prediction. In both cases the starting concentration values were varied according to the coefficients of variation, presented in Table 3.

It can be seen comparing the six diagrams above and below in Figure 3, that especially for viable cell density X_v as well as for total cell density X_t the amount of uncertainty, represented by the width of the prediction band, is reduced significantly meaning that the precision is increased. At the last point in time (hour 92 on the *x*-axis) the relative half bandwidth (compare with Section 2.6) was reduced from 45% to 28% for X_v. It should be noted that only one additional dataset containing five measurements per quantity were considered here. In terms of prediction accuracy both predictions are showing high scores, in both scenarios at least 90% of the test data are falling within the 90%-prediction band and the relative deviation between predicted and experimental data is 6%.

TABLE 4 Results of predictions of single bioreactor scales concerning precision and accuracy

Predicted scale	Training scale used for updating				
	SF	BR 1	BR 2	BR 3	Previous scale
	Half bandwidth (%)				
BR 1	44	26	29	29	–
BR 2	35	22	23	23	21
BR 3	34	21	23	22	23
	Within band score for X_v (in total; %)				
BR 1	92 (94)	94 (92)	78 (87)	87 (89)	–
BR 2	100 (92)	95 (88)	91 (90)	95 (89)	100 (90)
BR 3	88 (91)	66 (79)	78 (83)	83 (85)	88 (93)
	Rel. error for X_v (in total; %)				
BR 1	15 (10)	7 (8)	14 (10)	13 (10)	–
BR 2	7 (9)	5 (8)	7 (8)	6 (8)	5 (8)
BR 3	8 (9)	11 (10)	9 (9)	8 (9)	9 (8)

Note: As a measure for precision the relative half bandwidth for viable cell density before transfer was computed (low percentage, less uncertainty and therefore high precision) and as measures of accuracy the within band score and the rel. error were computed and presented in %, both for X_v and in total, meaning averaged over all six variables (viable and total cell concentration X_v and X_t, concentration of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm}). The following scales were used for training: Shake flasks (SF), shake flasks and another bioreactor scale (BR 1, BR 2 or BR 3), the previous bioreactor scale of the same cultivation (Previous scale).

3.3.2 | Prediction performance of a single bioreactor scale

The impact of available information on prediction performance of three single bioreactor scales (N-3 : 40 L, N-2 : 320 L, and N-1 : 2,160 L) was investigated for 10 seed trains comprising these bioreactor scales. The labels used in this section are listed in Table 1.

Prediction of a single bioreactor based on information from shake flask data, which was expressed by the corresponding model parameter distributions (compare with posterior 1 in Figure 2), lead to the results presented in the first column (SF) of Table 4, whereby the entries correspond to the averaged values, determined from investigation of datasets R1.11, … , R1.20 (reactor scale 1, 40 L), R2.11, … , R2.20 (reactor scale 2, 320 L) and R3.11, … , R3.20 (reactor scale 3, 2,160 L).

To investigate the impact of information from bioreactor scales of other seed trains, 10 seed trains ST1,..., ST10 were used as training data and seed trains ST11, … , ST20 as test data, whereby only one out of 10 seed train bioreactor scales was considered at a time (e.g., information from bioreactor R1.1 [40 L] of ST1 was used for prediction of bioreactor scale R1.11 [40 L] of ST11, then information from R2.1 [320 L] of ST1 was used for prediction of R1.11 [40 L] of ST11 etc.). This way, every combination of training and test data concerning scales was performed and investigated 10 times. The corresponding averaged results concerning precision and accuracy of prediction are shown in the columns 2–4 (BR1, BR2, and BR3) of Table 4.

It should be noted that no scale up parameters were considered within the underlying model, although differences in cell growth at different bioreactor scales were not excluded. In addition, process variability due to biological variability ("batch-to-batch variability") was expected. But such differences or variabilities would be expressed by corresponding changes in model parameter distributions (e.g., by the increase or decrease of the average maximum growth rate), as soon as respective data would be included for updating parameter distributions. These aspects will be discussed later on based on the presented findings.

Several aspects concerning propagation of uncertainty as well as prediction accuracy become apparent from the results presented in Table 4. Prediction of single bioreactor scales, only based on shake flask scale data was possible showing relative errors not exceeding 15% (for X_v) and 10% (in total), concerning predictions based on the Bayes estimator (MAP estimator, see Section 2.5). At least 88% (in case of X_v) and 91% (in total) of the test data are falling within the 90% prediction band. Nevertheless, predictions include between 34% and 44% of uncertainty (represented by the relative half bandwidth).

The inclusion of information from one bioreactor scale of another seed train bioreactor of the same campaign led to a reduction of predictive uncertainty (=increased precision) to 22–29% relative half bandwidth.

In terms of prediction accuracy, what stands out most is that predicting a bioreactor scale 1, a significantly higher accuracy was reached if another bioreactor scale 1 dataset was used for training

(see row 7 [BR 1] in Table 4). This indicates that there are sometimes small effects when cells are passaged from shaken conditions to stirred conditions (here this happens between shake flak scales and bioreactor scale 1). Prediction of a bioreactor scale 2 instead shows a high accuracy (rel. error: 5–8%, within band score 88–95%) independently of which bioreactor scale was used for training or even if information from shake flaks scales was used. Prediction of a bioreactors scale 3 turned out to perform best if another bioreactor scale 3 or even shake flask scales were used for training but also good results were reached if another bioreactor scale 2 was considered. A brief posterior analysis (after analyzing prediction performance) revealed a lower cell growth on average in reactor scale 1 compared with reactor scale 3, which was expressed by corresponding probability distributions.

This predictive performance has been further improved by using the information from the previous scale of the running cultivation to update the posterior distributions of model parameters one more time (see Table 4, last column). Predicting bioreactor scale 2, 100% (in case of X_v) and 90% (in total) of the test data are falling within the 90% prediction band which was reduced to 21% relative half bandwidth and the relative error states 5% (for X_v) and 8% in total. Predicting bioreactor scale 3, 88% (in case of X_v) and 93% (in total) of the test data are falling within the 90% prediction band which was reduced to 23% relative half bandwidth and the relative error states 9% (for X_v) and 8% in total. These results reveal that batch-to-batch variability can be considered by adaption of model parameter distributions through Bayesian updating. It hast to be mentioned that for each Bayesian update, only 4–5 measurements per quantity of a training dataset were used as additional information. It is expected that by adding more process data describing similar cell growth, the amount of predictive uncertainty decreases further. On the other hand, less measurement uncertainty would also lead to less uncertainty in the models outcome, because input uncertainty is propagated to uncertainty in the outcomes.

3.4 | Prediction of seed trains

In the previous sections it has been shown how single bioreactors could be predicted and how these predictions could be updated integrating information from additional data via Bayesian updating. Now, the complete bioreactor part of the seed train, comprising three consecutive bioreactor scales (40, 320, and 2,160 L) before the production bioreactor, is predicted. It should be noted, that in addition to the already considered sources of uncertainty (in model parameters and initial concentrations) uncertainty in the passaging process, which can be caused by different reasons like unknown volume when flushing the sampling valve or deviation of actual substrate concentration in the medium from the intended value (e.g., in case of glutamine in media), must be considered for the prediction of more than one seed train scale. This uncertainty was estimated evaluating the passaging processes of four seed trains (used as training data). In a first step, an exemplary seed train prediction, only based on small shake flask scale data will be illustrated. Afterwards, prediction performance is evaluated, taking further data from bioreactor scales into account.

3.4.1 | Seed train prediction based on shake flask data—Example and performance

A seed train prediction, only based on small scale shake flask data and considering the above-mentioned sources of uncertainty is illustrated as an example in Figure 4 (top left). Predictive time profiles, based on initial concentrations at reactor scale 1 and parameter distributions derived from small scale data, are illustrated as solid lines and 90% prediction bands composed of 90% prediction intervals at each considered point in time are illustrated by dashed lines. After seed train prediction (of ST13) the corresponding experimental data (test data) were considered for evaluation of the prediction concerning accuracy. It has to be mentioned that as points in time for cell passaging the experimentally realized points in time for passaging were applied, due to the comparability. In practice the point in time for cell passaging is often performed according to a specified strategy, for example, based on a minimum transfer cell density.

This seed train prediction leads to high accuracy (100% of the test data fall within the prediction band and the relative deviation between experimental data and predicted values yields 11%), but a low precision (the relative half bandwidth yields 65%). Considering the results of the prediction of single bioreactor scales, this results for seed train prediction can be improved as described below, where information from another seed train cultivation and as soon as available information from the previous scale of the running cultivation is used for the future predictions (see Subsection 3.4.2).

3.4.2 | Optimized seed train prediction through Bayesian updating

While in the last subsection prediction performance of seed train prediction based on the information from shake flask data was presented, here it was investigated how prediction performance changes, when new data (of the current cultivation) are collected and used to update model parameter distributions. By this, prediction for seed trains composed of three large bioreactor scales was performed stepwise:

I) Time profiles for reactor scale 1 (40 L), 2 (320 L), and 3 (2,160 L) were predicted based on the initial concentrations of reactor scale 1 and parameter distributions of a training dataset (data from reactor scales 1, 2, and 3 of another seed train of the same campaign). II) After running reactor scale 1, time profiles for reactor scale 2 and 3 were updated using the initial concentrations of reactor scale 2 and updated posterior parameter distributions using information from the previous (reactor) scale 1. III) After running reactor scale 2, time profiles for bioreactor scale 3 were updated, using the initial concentrations of reactor scale 3 and updated posterior distributions of the previous reactor scale 2.

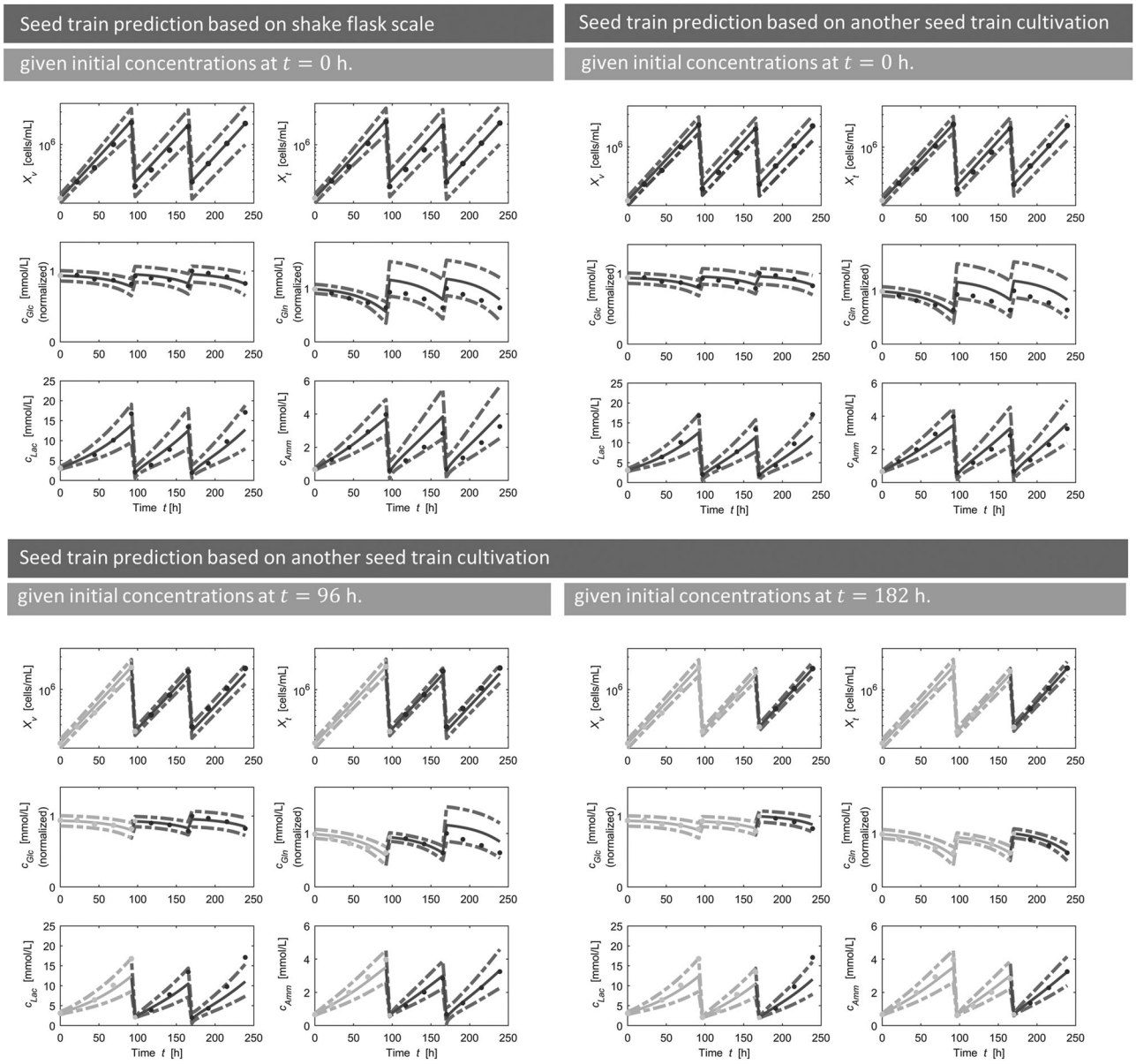

FIGURE 4 Prediction of an exemplary seed train, that is, three consecutive bioreactor scales of 40, 320, and 2,160 L for the six state variables, viable and total cell concentration X_v and X_t, concentration of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm}, for four scenarios: Only based on initial concentrations at reactor scale 1 (40 L) and posterior parameter distributions from shake flask scales (top left); based on initial concentrations at reactor scale 1 and posterior parameter distributions from another seed train (top right); based on initial concentrations at reactor scale 2 or 3 and including parameter distributions from the previous reactor scale (bottom left, bottom right)

The optimized prediction of an exemplary seed train is also illustrated in Figure 4. Time profiles for all six state variables, viable and total cell concentration, X_v and X_t, concentration of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm}, were predicted at the beginning of the seed train (top right), after collecting data from scale 1 (bottom left), and after collecting data from scale 2 (bottom right). Here again, predictive time profiles are shown by solid lines, and 90% prediction bands are illustrated by dashed lines. The corresponding precision and accuracy values are shown in Table 5, row 3 (ST13).

It can be seen that the prediction uncertainty for the remaining "future" time span is reduced (from 42% to 24% half bandwidth, see Table 5, row 3) after each update step indicated by narrow prediction bands (in the Figure, the "past" is shown in light gray, the "future" in the dark gray). Also notable is the fact that the high accuracy (within band score 96% in total [i.e., concerning all six variables] and 92% for X_v [i.e., concerning only viable cell density]) and rel. error of 10% in total and 7% for X_v, achieved for the prediction of bioreactor scale 1, 2, and 3 has been further improved by updating after cultivation of each scale.

After two updating steps the within band score yielded 100% for X_v and 96% in total and the rel. error yielded 1% for X_v and 5% in total (see also Table 5, ST 3).

Following the same procedure, 10 seed trains (from six different campaigns) were predicted, each based on one seed train used as training data (e.g., ST11 was predicted based on information

TABLE 5 Prediction performance of 10 optimized seed train predictions, (three consecutive bioreactor scales of 40 [R1], 320 [R2], and 2,160 L [R3]), based on one seed train for training each

Seed train	Within band score (%) for X_v (in total) (high values desired)			Rel. error (%) for X_v (in total) (low values desired)			Half bandwidth (%) for X_v (low values desired)		
	I) R1-R3	II) R2-R3	III) R3	I) R1-R3	II) R2-R3	III) R3	I) R1-R3	II) R2-R3	III) R3
ST 11	85 (91)	100 (92)	75 (75)	15 (14)	4 (10)	10 (9)	42	31	21
ST 12	100 (87)	100 (88)	100 (83)	5 (13)	11 (15)	8 (12)	41	31	20
ST 13	92 (96)	100 (98)	100 (96)	7 (10)	8 (9)	1 (5)	42	32	24
ST 14	77 (82)	63 (79)	75 (79)	17 (18)	13 (14)	13 (10)	39	29	23
ST 15	100 (90)	100 (94)	100 (100)	8 (15)	2 (11)	3 (5)	42	31	24
ST 16	100 (94)	100 (92)	100 (95.8)	6 (10)	9 (11)	4 (7)	39	32	25
ST 17	69 (86)	75 (79)	75 (83)	33 (22)	17 (17)	17 (12)	44	31	23
ST 18	85 (78)	100 (90)	100 (83)	14 (20)	10 (14)	10 (10)	43	31	23
ST 19	100 (96)	100 (87)	100 (87)	19 (16)	9 (14)	3 (11)	40	30	20
ST 20	100 (100)	100 (94)	75 (83)	7 (10)	6 (9)	8 (8)	38	28	21
Mean	91 (90)	94 (89)	90 (87)	13 (15)	9 (12)	8 (9)	41	31	21

Note: Within band score and rel. error only for X_v and in total (viable and total cell concentration X_v, X_t, concentration of glucose c_{Glc}, glutamine c_{Gln}, lactate c_{Lac}, and ammonia c_{Amm}); relative half bandwidth of prediction interval at last point in time for X_v.
I) Predictions based on initial concentrations at R1 and on posterior parameter distributions from another R1 cultivation of the same campaign.
II) Prediction after running R1 and update using initial concentrations of R2 and posterior parameter distributions of R1 for prediction of R2.
III) Prediction after running R2 and update using initial concentrations of R3 and posterior parameter distributions from previous scales.

from ST1 and so on) and updated as soon as data from one scale of the current cultivation were available. The corresponding results concerning prediction performance are presented in Table 5. Application of Bayesian updating led to significant narrowing of the prediction, meaning a reduction of uncertainty, while reaching or maintaining a high prediction accuracy. The amount of uncertainty concerning relative half bandwidth was reduced from 41% to 31% on average after a Bayesian update step using process data of the previous scale 1 for prediction of scale 2 (see last row "Mean" of Table 5). This uncertainty was further reduced to 21% on average after a Bayesian update step using process data of the previous scale 2. This improvement of precision was achieved without a loss of accuracy, because at least 90% (for X_v) and 87% (in total) of the test data were falling within the prediction band, while the rel. error has been even decreased from 13% to 8% (for X_v) and from 15% to 9% (in total) on average over 10 seed trains.

Nevertheless, the results show that not all seed trains could be predicted well only based on the information (updated parameter distribution) from another randomly sampled seed train as this is the case for seed train 17 (ST17) which was predicted based on information of ST7 (see row 4, ST17, first column of within band score and first column of relative error in Table 5). This can occur due to batch-to-batch variability. A brief posterior analysis (after investigation of prediction performance) revealed that variabilities between the cultivations (batches [here between ST7 and ST17]) are sometimes bigger than the already mentioned variability between reactor scale 1 and the other scales (compare with Section 3.3.2). When this occurred this was taken into account by a Bayesian updating step using data from the ongoing seed train. The presented results show that this way, an improvement of prediction performance was achieved. Considering for example seed trains ST17 and ST18, the rel. error concerning all six variables was reduced from 22% to 12% and from 20% to 10% respectively, after two updating steps.

It should be noticed that in this contribution only few datasets were used for the Bayesian updating steps, because it was intended to illustrate every step and to show the corresponding changes to visualize the impact of the available information. Often, there are more datasets available in practice, which could be used for further updating steps, leading possibly to less predictive uncertainty in case of consistent cultivations. Apart from that, the knowledge about predictive uncertainty (which reflects the propagated uncertainty due to input uncertainties) is highly relevant, even though it is not as small as desired. It can help to find out where the process could fail or be improved (e.g., uncertainty would decrease if less uncertainty in measurements or variability in the passaging process could be assured). Using Bayesian parameter estimation and Bayesian updating as presented in this work it could furthermore be investigated, to which amount predictive uncertainty would decrease if input uncertainty would be decreased to a certain amount.

4 | CONCLUSION

In this contribution, the application of a Bayesian approach for parameter estimation and prediction of an industrial cell culture seed train, enabling the integration of prior knowledge and the consideration of uncertainty, is presented. Subject of investigations is the bioreactor part of an industrial CHO cell culture seed train

comprising three consecutive bioreactor scales (40, 320, and 2,160 L filling volume) under consideration of shake flask experiments under equal cultivation conditions.

It has been shown that Bayesian parameter estimation, performed using MCMC simulations, in combination with a mechanistic model, describing the time profiles of viable and total cell density as well as concentrations of glucose, glutamine, lactate, and ammonia, is a suitable statistical method for seed train prediction. It provides the capability of propagating information content (including input uncertainty) provided by prior knowledge and experimental data to prediction uncertainty, expressed by predictions intervals. This way, process relevant decisions can be made based on probabilities of certain events. It should be noted that the same mechanistic model was applied for all scales, from shake flask scales to large bioreactor scales (up to 2,160 L filling volume). It became apparent that despite batch-to-batch variability (e.g., due to biological variability) a high predictive accuracy can be reached, by taking data of the running seed train cultivation into account performing Bayesian updating.

This approach provides various practical advantages concerning applications within the field of bioprocessing. One potential advantage is the capability of the design of robust and optimal seed train protocols, saving experimental work by using prior knowledge (which is currently subject of investigation). Besides, the transfer from one plant to a similar plant can be supported. Furthermore, prediction of running processes can be used for feed-forward control strategies (e.g., prediction of points in time for cell passaging that can be based on viable cell density) or for the development of soft sensors (predicting variables which are difficult to measure).

NOMENCLATURE

α	shape parameter of gamma distribution
λ	rate parameter of gamma distribution
μ	cell-specific growth rate [hr^{-1}]
μ_d	cell-specific death rate [hr^{-1}]
$\mu_{d,min}$	minimum cell-specific growth rate [hr^{-1}]
$\mu_{d,max}$	maximum cell-specific growth rate [hr^{-1}]
σ	standard deviation
a_{Lag}	correction factor for lag phase
c_{Amm}	ammonia concentration [mmol/L]
c_{Glc}	glucose concentration [mmol/L]
c_{Gln}	glutamine concentration [mmol/L]
c_{Lac}	lactate concentration [mmol/L]
cv	coefficient of variation
F_{Glc}	glucose feeding rate [L/hr]
F_{Gln}	glutamine feeding rate [L/hr]
F_{Sample}	sampling rate [L/hr]
k_{Amm}	correction factor for ammonia uptake [–]
k_{Glc}	monod kinetic constant for glucose uptake [mmol/L]
k_{Gln}	monod kinetic constant for glutamine uptake [mmol/L]
K_{Lys}	cell lysis constant [hr^{-1}]
$K_{S,Glc}$	monod kinetic constant for glucose [mmol/L]
$K_{S,Gln}$	monod kinetic constant for glutamine [mmol/L]
PAT	process analytical technology
q_{Amm}	cell-specific ammonia production rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
$q_{Amm,uptake}$	cell-specific ammonia uptake rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
q_{Lac}	cell-specific lactate production rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
q_{Glc}	cell-specific glucose uptake rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
$q_{Glc,max}$	maximum cell-specific glucose uptake rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
q_{Gln}	cell-specific glutamine uptake rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
$q_{Gln,max}$	maximum cell-specific glutamine uptake rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
$q_{Lac,max}$	cell-specific lactate production rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
$q_{Lac,uptake}$	cell-specific lactate uptake rate [$mmol \cdot L^{-1} \cdot hr^{-1}$]
t	time [hr]
t_{Lag}	duration of lag phase [hr]
X_t	total cell density [cells/L]
X_v	viable cell density [cells/L]
y	concentration values in general (data)
y_0	initial concentration values
$Y_{Amm/Gln}$	kinetic production constant (stoichiometric ratio of ammonia production and glutamine uptake) [–]
$Y_{Lac/Glc}$	kinetic production constant (stoichiometric ratio of lactate production and glucose uptake) [–]

ORCID

Tanja Hernández Rodríguez http://orcid.org/0000-0002-7667-8390

REFERENCES

Aehle, M., Bork, K., Schaepe, S., Kuprijanov, A., Horstkorte, R., Simutis, R., & Lübbert, A. (2012). Increasing batch-to-batch reproducibility of cho-cell cultures usinga model predictive control approach. *Cytotechnology*, *64*(6), 623–634.

Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A., & Blom, J. G. (2009). Systems biology: Parameter estimation for biochemical models. *The FEBS Journal*, *276*(4), 886–902.

Brunner, M., Fricke, J., Kroll, P., & Herwig, C. (2017). Investigation of the interactions of critical scale-up parameters (pH, pO_2 and pCO_2) on CHO batch performance and critical quality attributes. *Bioprocess and Biosystems Engineering*, *40*(2), 251–263.

Frahm, B. (2014). Seed train optimization for cell culture. In Pörtner, R. (Ed.), *Animal cell biotechnology* (pp. 355–367). New York, NY: Humana Press.

Frahm, B., Lane, P., Atzert, H., Munack, A., Hoffmann, M., Hass, V. C., & Pörtner, R. (2002). Adaptive, model-based control by the open-loop-feedback-optimal (OLFO) controller for the effective fed-batch cultivation of hybridoma cells. *Biotechnology Progress*, *18*(5), 1095–1103.

Galagali, N., & Marzouk, Y. M. (2015). Bayesian inference of chemical kinetic models from proposed reactions. *Chemical Engineering Science*, *123*, 170–190.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd.). Hoboken, NJ: Chapman & Hall/CRC.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1998). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.

Glassey, J., Gernaey, K. V., Clemens, C., Schulz, T. W., Oliveira, R., Striedner, G., & Mandenius, C.-F. (2011). Process analytical technology (PAT) for biopharmaceuticals. *Biotechnology Journal*, *6*(4), 369–377.

Hines, K. E. (2015). A primer on Bayesian inference for biophysical systems. *Biophysical Journal*, *108*(9), 2103–2113.

Kern, S., Platas-Barradas, O., Pörtner, R., & Frahm, B. (2016). Model-based strategy for cell culture seed train layout verified atlab scale. *Cytotechnology*, *68*(4), 1019–1032.

Kroll, P., Hofer, A., Stelzer, I. V., & Herwig, C. (2017). Workflow to set up substantial target-oriented mechanistic processmodels in bioprocess engineering. *Process Biochemistry*, *62*, 24–36.

Le, H., Kabbur, S., Pollastrini, L., Sun, Z., Mills, K., Johnson, K., & Hu, W.-S. (2012). Multivariate analysis of cell culture bioprocess data-lactate-consumption as process indicator. *Journal of Biotechnology*, *162*(2-3), 210–223.

Liu, Y., & Gunawan, R. (2017). Bioprocess optimization under uncertainty using ensemble modeling. *Journal of Biotechnology*, *244*, 34–44.

Luce, B., Anthony, O. H., & Dennis, F. (2003). *A primer on Bayesian statistics in health economics and outcomes research*. Bethesda, MD: MEDTAP International, Incorporated.

Matlab (2017). Version 9.3.0.713579 (R2017b). Natick, MA: The MathWorks, Inc.

Möller, J., & Pörtner, R. (2017). Model-based design of process strategies for cell culture bioprocesses: State of the art and new perspectives. In Gowder, S. J. T. (Ed.), *New insights into cell culture technology* (pp. 157–172). Rijeka, Croatia: InTech.

O'Hagan, A. (2008). The bayesian approach to statistics. In Rudas, T. (Ed.), *Handbook of probability* (pp. 85–100). Los Angeles, CA: SAGE.

Press, W. H. (1996). *Numerical recipes in C: The art of scientific computing* (2nd.). Cambridge, UK: Cambridge University Press.

Price, J., Nordblad, M., Woodley, J. M., & Huusom, J. K. (2013). Application of uncertainty and sensitivity analysis to a kineticmodel for enzymatic biodiesel production. *IFAC Proceedings Volumes*, *46*(31), 149–156.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partiallyobserved dynamical models by exploiting the profile likelihood. *Bioinformatics*, *25*(15), 1923–1929.

Raue, A., Kreutz, C., Theis, F. J., & Timmer, J. (2013). Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *371*, 1984.

Sanderson, C. S., Phillips, P. J., & Barford, J. P. (1996). Structured modelling of animal cells. *Cytotechnology*, *21*(2), 149–153.

Sin, G., Gernaey, K. V., & Lantz, A. E. (2009). Good modeling practice for pat applications: Propagation of inputuncertainty and sensitivity analysis. *Biotechnology Progress*, *25*(4), 1043–1053.

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, *75*, 273–316.

Xing, Z., Bishop, N., Leister, K., & Li, Z. J. (2010). Modeling kinetics of a large-scale fed-batch CHO cell culture by Markov chain Monte Carlo method. *Biotechnology Progress*, *26*(1), 208–219.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Hernández Rodríguez T, Posch C, Schmutzhard J, et al. Predicting industrial-scale cell culture seed trains–A Bayesian framework for model fitting and parameter estimation, dealing with uncertainty in measurements and model parameters, applied to a nonlinear kinetic cell culture model, using an MCMC method. *Biotechnology and Bioengineering*. 2019;116: 2944–2959. https://doi.org/10.1002/bit.27125