Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers and Chemical Engineering 28 (2004) 1431–1449
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

A method for systematic improvement of stochastic grey-box models

Niels Rode Kristensen [a,*], Henrik Madsen [b], Sten Bay Jørgensen [a]

[a] *Department of Chemical Engineering, Technical University of Denmark, Building 229, Lyngby DK-2800, Denmark*
[b] *Informatics and Mathematical Modelling, Technical University of Denmark, Building 321, Lyngby DK-2800, Denmark*

Received 25 November 2002; received in revised form 17 October 2003; accepted 17 October 2003

Abstract

A systematic framework for improving the quality of continuous time models of dynamic systems based on experimental data is presented. The framework is based on an interplay between stochastic differential equation modelling, statistical tests and nonparametric modelling and provides features that allow model deficiencies to be pinpointed and their structural origin to be uncovered. More specifically, the proposed framework can be used to obtain estimates of unknown functional relations, in turn allowing unknown or inappropriately modelled phenomena to be uncovered. In this manner the framework permits systematic iterative model improvement. The performance of the proposed framework is illustrated through a case study involving a dynamic model of a fed-batch bioreactor, where it is shown how an inappropriately modelled biomass growth rate can be uncovered and a proper functional relation inferred. A key point illustrated through this case study is that functional relations involving unmeasured variables can also be uncovered.

Keywords: Model improvement; Stochastic differential equations; Parameter estimation; Statistical tests; Nonparametric modelling; Bioreactor modelling

1. Introduction

Dynamic process models are used in many areas of chemical engineering and for many different purposes. Dynamic model development is therefore inherently purpose-driven in the sense that the required accuracy of a model, in terms of prediction capabilities, depends on its intended application. More specifically, models intended for open-loop applications such as process simulation and optimisation, where long-term prediction capabilities are important, must be more accurate than models intended for closed-loop applications such as standard feedback control, where only short-term prediction capabilities are needed. However, to be more accurate, a model must be more complex, which means that it will be more difficult and time-consuming to develop. Finding a suitable model for a given purpose thus involves a trade-off between required model accuracy and affordable model complexity (Raisch, 2000).

For open-loop applications, ordinary differential equation (ODE) models or *white-box* models developed from first engineering principles and physical insights are typically used. Such models are often very detailed, because they must be able to capture nonlinear effects in order to be valid over wide ranges of state space, and, as a consequence, developing such models may be difficult and time-consuming. Indeed, the corresponding model development procedure is by no means guaranteed to converge, and few tools for making inferences about the structure of such models are available.

For closed-loop applications, much simpler input–output models or *black-box* models developed from experimental data with methods for time series analysis and system identification can often be used (Box & Jenkins, 1976; Ljung, 1987; Söderström & Stoica, 1989). Such models only have to be valid for a small range of state space, typically close to a constant operating point, which means that nonlinear effects can be neglected, making model development much faster. Furthermore, well-developed tools for structural identification of such linear models are available and the corresponding model development procedure is guaranteed to converge if certain conditions of identifiability of parameters and persistency of excitation of inputs are fulfilled.

Model-based optimizing control of batch and fed-batch processes, e.g. by means of nonlinear model predictive control (MPC) (Allgöwer & Zheng, 2000), presents a borderline case between open-loop and closed-loop applications,

* Corresponding author. Tel.: +45-44428446; fax: +45-44428300.
E-mail address: nikr@novonordisk.com (N.R. Kristensen).

0098-1354/$ – see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2003.10.003

where neither of the above modelling approaches is ideal. On one hand, a model is needed, which is sufficiently accurate to be used for long-term prediction over wide ranges of state space, but on the other hand, the affordable model complexity is low due to the importance of time-to-market issues in the biochemical, pharmaceutical and specialty chemicals industries, where batch and fed-batch processes are common.

A methodology that provides an appealing trade-off between the white-box and black-box approaches is *grey-box* modelling, where mechanistic and empirical model components are combined, which may be done in a deterministic as well as a stochastic setting. Not disregarding the importance of deterministic grey-box modelling, the remainder of the present paper will be concerned with stochastic grey-box modelling (Madsen & Melgaard, 1991; Melgaard & Madsen, 1993; Bohlin & Graebe, 1995; Bohlin, 2001), the key idea of which is to find the simplest model for a given purpose, which is consistent with prior physical knowledge and not falsified by available experimental data. In the approach by Bohlin and Graebe (1995) and Bohlin (2001) this is done by formulating a sequence of hypothetical model structures of increasing complexity and systematically expanding the model by falsifying incorrect hypotheses through statistical tests based on the experimental data. This way models can be developed, which have almost the same validity range as white-box models, but it can be done in a less time-consuming manner and the models are guaranteed not to be overly complex.

Stochastic grey-box models are stochastic state space models consisting of a set of stochastic differential equations (SDEs) (Øksendal, 1998) describing the dynamics of the system in continuous time and a set of discrete time measurement equations. A considerable advantage of such models as opposed to white-box models is that they are designed to accommodate random effects. In particular, they allow for a decomposition of the noise affecting the system into a process noise term and a measurement noise term. As a consequence of this *prediction error decomposition* (PED), unknown parameters of stochastic grey-box models can be estimated from experimental data in a *prediction error* (PE) setting (Young, 1981), whereas for white-box models it can only be done in an *output error* (OE) setting (Young, 1981), which tends to give biased and less reproducible results, because random effects are absorbed into the parameter estimates, particularly if the model structure is incorrect. Furthermore, PE estimation allows a number of powerful statistical tools to be applied to give indications for possible improvements to the model structure.

Stochastic grey-box modelling as presented by Bohlin and Graebe (1995) and Bohlin (2001) is an iterative and inherently interactive procedure, because it relies on the model maker to formulate the specific hypothetical model structures to be tested to improve the model. As pointed out by Bohlin (2001) this poses the problem that the model maker may run out of ideas for improvement before a sufficiently accurate model is obtained, which means that he or she may have to resort to using black-box models for filling the gaps. In the present paper a stochastic grey-box modelling framework is proposed, which relies less on the model maker. Within this framework specific model deficiencies can be pinpointed and their structural origin can be uncovered, which provides the model maker with valuable information about how to formulate new hypotheses to improve the model. This clearly speeds up the iterative model development procedure, and, as an additional benefit, also prevents the model maker from having to resort to using black-box models for filling the gaps, when all prior physical knowledge is exhausted. The key to obtaining information about how to improve the model is the ability of the proposed framework to provide estimates of unknown functional relations, allowing unknown or inappropriately modelled phenomena to be uncovered. These estimates are obtained by making use of the PED and other properties of stochastic state space models along with nonparametric modelling. The integration of nonparametric modelling with conventional stochastic grey-box modelling into a systematic framework for model improvement is the key result of the paper. The remainder of the paper is organized as follows: In Section 2 the details of the proposed framework are outlined and in Section 3 a case study illustrating its performance is presented. In Section 4 a discussion of important results is given and in Section 5 the conclusions of the paper are presented.

2. Methodology

In this section the details of the proposed stochastic grey-box modelling framework are outlined. The framework is shown in Fig. 1 in the form of a modelling cycle comprising the individual steps of the model development procedure. A key idea of stochastic grey-box modelling is to use all relevant prior physical knowledge, for which reason the first step within the modelling cycle is *model (re)formulation* based on first engineering principles, where the idea is to formulate an initial model structure (first modelling cycle iteration) or make modifications to this structure (subsequent iterations). The second step within the modelling cycle is *parameter estimation*, where the idea is to estimate unknown parameters of the model from available experimental data, and the third step is *residual analysis*, where the idea is to evaluate the quality of the resulting model by means of cross-validation. The fourth step within the modelling cycle is the important step of *model falsification or unfalsification*, which deals with whether or not, based on the available information, the model is sufficiently accurate to serve its intended purpose. If the model is unfalsified, the model development procedure can be terminated, but if the model is falsified, the modelling cycle must be repeated by re-formulating the model. A key feature of the proposed framework is that, in the latter case, the PED and

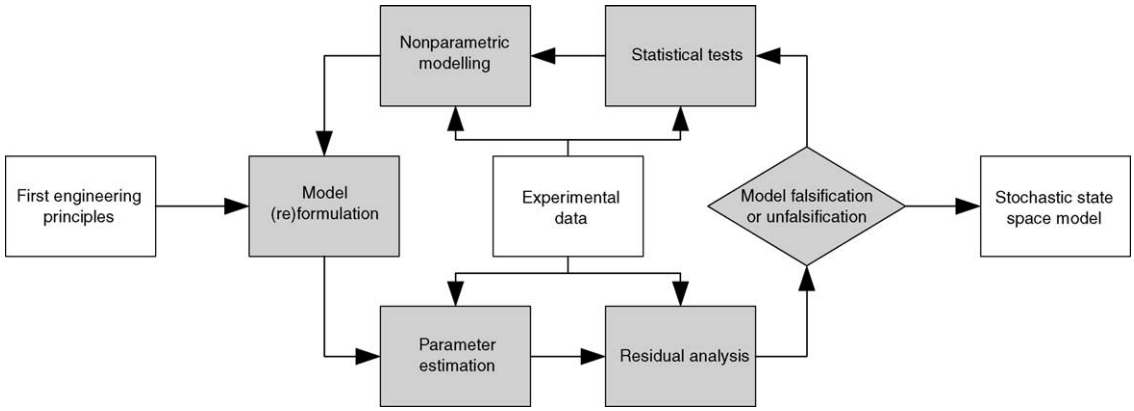


Fig. 1. The proposed modelling cycle. The boxes in grey illustrate tasks and the boxes in white illustrate inputs to and outputs from the modelling cycle.

other properties of stochastic state space models can be exploited to facilitate the task at hand. More specifically, the *statistical tests* of the fifth step within the modelling cycle can be applied to provide indications of which parts of the model that are deficient, and the *nonparametric modelling* techniques of the sixth step can be applied to provide estimates of the functional relations needed to repair these deficiencies in order to improve the model. In the remainder of this section the individual steps are described in more detail and an algorithm for systematic model improvement based on the proposed modelling cycle is presented.

2.1. Model (re)formulation

In the first step of the proposed modelling cycle, the idea is to formulate an initial model structure. This is a two-step procedure, because it involves derivation of a standard ODE model from first engineering principles and translation of the ODE model into a stochastic state space model consisting of a set of SDEs and a set of discrete time measurement equations. Deriving an ODE model from first engineering principles is a standard discipline for most chemical engineers and yields a model of the following type:

$$\frac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) \tag{1}$$

where $t \in \mathbb{R}$ is time, $\boldsymbol{x}_t \in \mathbb{R}^n$ is a vector of balanced quantities or state variables, $\boldsymbol{u}_t \in \mathbb{R}^m$ is a vector of input variables and $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of possibly unknown parameters, and where $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$ is a nonlinear function. Translating the ODE model into a stochastic state space model is also straightforward, as it can simply be done by replacing the ODEs with SDEs and adding a set of algebraic equations describing how measurements are obtained at discrete time instants. This yields a model of the following type:

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})\,\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{u}_t, t, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\omega}_t \tag{2}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{3}$$

where $t \in \mathbb{R}$ is time (t_k, $k = 0, \ldots, N$ are sampling instants), $\boldsymbol{x}_t \in \mathbb{R}^n$ is a vector of state variables, $\boldsymbol{u}_t \in \mathbb{R}^m$ is a vector of input variables, $\boldsymbol{y}_k \in \mathbb{R}^l$ is a vector of measured output variables, $\boldsymbol{\theta} \in \mathbb{R}^p$ is a vector of possibly unknown parameters, $\boldsymbol{f}(\cdot) \in \mathbb{R}^n$, $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{h}(\cdot) \in \mathbb{R}^l$ are nonlinear functions, $\{\boldsymbol{\omega}_t\}$ is an n-dimensional standard Wiener process and $\{\boldsymbol{e}_k\}$ is an l-dimensional white noise process with $\boldsymbol{e}_k \in \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta}))$. The first term on the right-hand side of (2) is called the *drift* term and is a deterministic term equivalent to the term on the right-hand side of (1), whereas the second term on the right-hand side of (2) is called the *diffusion* term and is a stochastic term included to accommodate random effects due to, e.g. approximation errors or unmodelled phenomena. A detailed account of the theory behind SDEs is given by Øksendal (1998).

The diffusion term is the key to the proposed procedure for systematic model improvement, because estimation of the parameters of this term from experimental data provides a measure of model uncertainty.

The translation of the ODE model into a stochastic state space model does not affect the parameters of the drift term, which means that their physical interpretability is preserved.

Remark 1. The standard Wiener process $\{\boldsymbol{\omega}_t\}$ driving the SDEs in (2) is a continuous stochastic process with stationary and independent Gaussian time increments, which have zero mean and a covariance that is equal to the size of the time increment (Jazwinski, 1970).

Remark 2. The notation used in (2) is shorthand for the corresponding integral interpretation and is therefore ambiguous unless a specific integral interpretation is given. SDEs may be interpreted both in the sense of Stratonovich and in the sense of Itô (Jazwinski, 1970), but since the Stratonovich interpretation is unsuitable for parameter estimation (Åström, 1970), the Itô interpretation is adapted in the following.

2.2. Parameter estimation

In the second step of the proposed modelling cycle the idea is to estimate the unknown parameters of the stochastic

state space model (2) and (3) from experimental data. The solution to (2) is a Markov process, and hence an estimation scheme based on probabilistic methods can be applied. A brief outline of the scheme used within the proposed framework is given in the following. A more detailed account is given by Kristensen, Madsen, and Jørgensen (2003).

2.2.1. Maximum likelihood (ML) estimation

Given a sequence of measurements $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k, \ldots, \boldsymbol{y}_N$, ML estimates of the unknown parameters in (2) and (3) can be determined by finding the parameters $\boldsymbol{\theta}$ that maximize the likelihood function, i.e. the joint probability density:

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = p(\mathcal{Y}_N|\boldsymbol{\theta}) = p(\boldsymbol{y}_N, \boldsymbol{y}_{N-1}, \ldots, \boldsymbol{y}_1, \boldsymbol{y}_0|\boldsymbol{\theta}) \tag{4}$$

or equivalently:

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = \left(\prod_{k=1}^{N} p(\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}) \right) p(\boldsymbol{y}_0|\boldsymbol{\theta}) \tag{5}$$

where the rule $P(A \cap B) = P(A|B)P(B)$ has been applied to form a product of conditional probability densities.

In order to obtain an exact evaluation of the likelihood function, a general nonlinear filtering problem must be solved (Jazwinski, 1970), but this is computationally infeasible in practice. However, since the increments of the standard Wiener process $\{\boldsymbol{\omega}_t\}$ driving the SDEs in (2) are Gaussian, it is reasonable to assume that the conditional probability densities in (5) can be well approximated by Gaussian densities. Thus a method based on the much simpler extended Kalman filter (EKF) can be applied.

Remark 3. The validity of the Gaussianity assumption can be checked subsequent to the estimation, and a number of different methods are available for this purpose (Holst, Holst, Madsen, & Melgaard, 1992; Bak, Madsen, & Nielsen, 1999). However, the assumption is only likely to hold if the structure of the model is appropriate, and it may therefore not be strictly correct in the initial iterations of the modelling cycle. Nevertheless, the corresponding estimation results can be used to provide indications for model improvement as shown in the next sections.

The Gaussian density is completely characterized by its mean and covariance, so by introducing the notation:

$$\hat{\boldsymbol{y}}_{k|k-1} = E\{\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\} \tag{6}$$

$$\boldsymbol{R}_{k|k-1} = V\{\boldsymbol{y}_k|\mathcal{Y}_{k-1}, \boldsymbol{\theta}\} \tag{7}$$

$$\boldsymbol{\epsilon}_k = \boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1} \tag{8}$$

the likelihood function can be rewritten:

$$L(\boldsymbol{\theta}; \mathcal{Y}_N) = \left(\prod_{k=1}^{N} \frac{\exp(-(1/2)\boldsymbol{\epsilon}_k^{\mathrm{T}} \boldsymbol{R}_{k|k-1}^{-1} \boldsymbol{\epsilon}_k)}{\sqrt{\det(\boldsymbol{R}_{k|k-1})}(\sqrt{2\pi})^l} \right) p(\boldsymbol{y}_0|\boldsymbol{\theta}) \tag{9}$$

and the parameter estimates can be determined by conditioning on $\boldsymbol{y}_0$ and solving the nonlinear optimisation problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \{-\ln(L(\boldsymbol{\theta}; \mathcal{Y}_N|\boldsymbol{y}_0))\} \tag{10}$$

where, for each set of parameters $\boldsymbol{\theta}$ in the optimisation, $\boldsymbol{\epsilon}_k$ and $\boldsymbol{R}_{k|k-1}$ are computed recursively by means of the EKF.

2.2.2. Maximum a posteriori (MAP) estimation

If prior information about the parameters is available and given in the form of a prior probability density $p(\boldsymbol{\theta})$ for the parameters, Bayes' rule can be applied to give an improved estimate by forming the posterior probability density:

$$p(\boldsymbol{\theta}|\mathcal{Y}_N) = \frac{p(\mathcal{Y}_N|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y}_N)} \propto p(\mathcal{Y}_N|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{11}$$

and subsequently finding the parameters that maximize this function, i.e. by performing MAP estimation. By assuming that the prior probability density of the parameters is Gaussian, and by introducing the notation:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = E\{\boldsymbol{\theta}\} \tag{12}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = V\{\boldsymbol{\theta}\} \tag{13}$$

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}} \tag{14}$$

the posterior probability density can be rewritten:

$$p(\boldsymbol{\theta}|\mathcal{Y}_N) \propto \left(\prod_{k=1}^{N} \frac{\exp(-(1/2)\boldsymbol{\epsilon}_k^{\mathrm{T}} \boldsymbol{R}_{k|k-1}^{-1} \boldsymbol{\epsilon}_k)}{\sqrt{\det(\boldsymbol{R}_{k|k-1})}(\sqrt{2\pi})^l} \right) p(\boldsymbol{y}_0|\boldsymbol{\theta}) \times \frac{\exp(-(1/2)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\epsilon}_{\boldsymbol{\theta}})}{\sqrt{\det(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})}(\sqrt{2\pi})^p} \tag{15}$$

and the parameter estimates can now be determined by conditioning on $\boldsymbol{y}_0$ and solving the nonlinear optimisation problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \{-\ln(p(\boldsymbol{\theta}|\mathcal{Y}_N, \boldsymbol{y}_0))\} \tag{16}$$

Remark 4. If no prior information is available ($p(\boldsymbol{\theta})$ uniform), this formulation reduces to the ML formulation in (10). Thus, it can be seen as a generalization of the ML formulation. In fact, this formulation also allows for MAP estimation on a subset of the parameters ($p(\boldsymbol{\theta})$ partly uniform).

2.2.3. Using multiple independent data sets

If multiple consecutive, but stochastically independent, sequences of measurements $\mathcal{Y}_{N_1}^1, \mathcal{Y}_{N_2}^2, \ldots, \mathcal{Y}_{N_i}^i, \ldots, \mathcal{Y}_{N_S}^S$, are available, a similar estimation method can be applied by expanding the posterior probability density to:

$$p(\boldsymbol{\theta}|\boldsymbol{Y}) = p(\boldsymbol{\theta}|\mathcal{Y}_{N_1}^1, \mathcal{Y}_{N_2}^2, \ldots, \mathcal{Y}_{N_i}^i, \ldots, \mathcal{Y}_{N_S}^S]) \propto \left(\prod_{i=1}^{S} \left(\prod_{k=1}^{N_i} \frac{\exp(-(1/2)(\boldsymbol{\epsilon}_k^i)^{\mathrm{T}} (\boldsymbol{R}_{k|k-1}^i)^{-1} \boldsymbol{\epsilon}_k^i)}{\sqrt{\det(\boldsymbol{R}_{k|k-1}^i)}(\sqrt{2\pi})^l} \right) p(\boldsymbol{y}_0^i|\boldsymbol{\theta}) \right) \times \frac{\exp(-(1/2)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\epsilon}_{\boldsymbol{\theta}})}{\sqrt{\det(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})}(\sqrt{2\pi})^p} \tag{17}$$

and the parameter estimates can now be determined by conditioning on $\mathbf{y}_0 = [\boldsymbol{y}_0^1, \boldsymbol{y}_0^2, \ldots, \boldsymbol{y}_0^i, \ldots, \boldsymbol{y}_0^S]$ and solving the nonlinear optimisation problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \{-\ln(p(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{y}_0))\} \tag{18}$$

Remark 5. If only one sequence of measurements is available ($S = 1$), this formulation reduces to the MAP formulation in (16). Thus, it can be seen as a generalization of the MAP formulation for multiple independent data sets.

In the estimation scheme used within the proposed framework the nonlinear optimisation problem (18) is solved by means of a quasi-Newton method incorporating the BFGS updating formula and a soft line search algorithm. More details about this method and about how robustness towards outliers and missing observations has been incorporated into the estimation scheme are given by Kristensen et al. (2003), who also demonstrate the general efficiency and consistency of the scheme, especially with respect to the parameters of the diffusion term, which is of key importance within the proposed framework.

2.3. Residual analysis

In the third step of the proposed modelling cycle, the idea is to evaluate the quality of the model once the unknown parameters have been estimated.

An important aspect in assessing the quality of the model is to investigate its predictive capabilities by performing cross-validation and examining the corresponding residuals. Depending on the intended application of the model this should be done in either a one-step-ahead prediction setting (closed-loop applications) or in a pure simulation setting (open-loop applications). In either case a number of different methods can be applied (Holst et al., 1992).

One of the most powerful of these methods is to compute and inspect the *sample autocorrelation function* (SACF) and the *sample partial autocorrelation function* (SPACF) (Brockwell & Davis, 1991) of the residuals to detect if they can be regarded as white noise or if there are significant lag dependencies, i.e. correlations between current and lagged values of the residuals, as this indicates that the predictive capabilities of the model are not perfect.

Nielsen and Madsen (2001) recently presented extensions of these linear tools to nonlinear systems in the form of the *lag-dependence function* (LDF) and the *partial lag-dependence function* (PLDF), which are based on a close relation between correlation coefficients and the coefficients of determination for regression models. This relation allows for an extension to nonlinear systems by incorporating various nonparametric regression models.

Remark 6. Being an extension of the SACF, the LDF can be interpreted as being, for each lag k, the part of the overall variation in the observations of X_t from a stochastic process $\{X_t\}$, which can be explained by the observations of X_{t-k}. Likewise, being an extension of the SPACF, the PLDF can be interpreted as being, for each lag k, the relative decrease in one-step-ahead prediction variation when including X_{t-k} as an extra predictor.

Unlike the SACF and the SPACF, the LDF and the PLDF can also detect certain nonlinear lag dependencies and are therefore extremely useful for residual analysis within the proposed framework. More details about these and other similar tools are given by Nielsen and Madsen (2001).

Remark 7. If the Gaussianity assumption mentioned in Section 2.2 holds, which is only likely to be the case in the final iterations of the modelling cycle, i.e. when an appropriate model structure has been obtained, the statistical tests described in Section 2.5 can also be applied in the evaluation of the quality of the model. More specifically, it can be determined if some of the parameters of the model are insignificant, indicating that the model is overparameterized and that these parameters may be eliminated.

2.4. Model falsification or unfalsification

In the fourth step of the proposed modelling cycle, the idea is to determine whether or not, based on the information obtained in the previous step, the model is sufficiently accurate to serve its intended purpose. This essentially involves a completely subjective decision by the model maker, addressing the trade-off between required model accuracy and affordable model complexity for the particular application. Nevertheless, a few guidelines can be given.

For models intended for closed-loop applications such as standard feedback control, where only short-term prediction capabilities are important, whiteness of cross-validation residuals obtained in a one-step-ahead prediction setting is a good indication of sufficient model accuracy. On the other hand, for models intended for open-loop applications such as process simulation and optimisation, where long-term prediction capabilities are important, whiteness of cross-validation residuals obtained in a pure simulation setting is a very good such indication. However, sufficient information may not be available to achieve this, and the model maker may have to settle for less.

If, with respect to the available information, the model is unfalsified for its intended purpose, the model development procedure can be terminated. If, on the other hand, the model is falsified, the modelling cycle must be repeated by re-formulating the model. In the latter case, the properties of the model in (2) and (3) facilitate the task at hand as shown in the following.

2.5. Statistical tests

In the fifth step of the proposed modelling cycle, which is only needed if the model has been falsified and needs to

be improved, the idea is to apply statistical tests to provide indications of which parts of the model that are deficient. The key statistical tests needed for this purpose are tests for significance of the individual parameters, particularly the parameters of the diffusion term.

Remark 8. The residual analysis tools mentioned in Section 2.3 can also be applied in the analysis of possibilities for model improvement, at least if it holds that the residuals can be regarded as a realization from a stationary stochastic process. More specifically, like the SACF and the SPACF, the LDF and the PLDF can be applied for structural identification (Nielsen & Madsen, 2001), e.g. to determine if more state variables are needed.

An estimate of the uncertainty of the individual parameter estimates can be obtained by using the fact that by the central limit theorem the estimator in (18) is asymptotically Gaussian with mean $\boldsymbol{\theta}$ and covariance:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{H}^{-1} \tag{19}$$

where the matrix $\boldsymbol{H}$ is given by:

$$\{h_{ij}\} = -E\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln(p(\boldsymbol{\theta}|\boldsymbol{Y},\mathbf{y}_0))\right\},\quad i,j=1,\ldots,p \tag{20}$$

and where an estimate of $\boldsymbol{H}$ can be obtained from:

$$\{h_{ij}\} \approx -\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln(p(\boldsymbol{\theta}|\boldsymbol{Y},\mathbf{y}_0))\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},\quad i,j=1,\ldots,p \tag{21}$$

which is simply the Hessian evaluated at the minimum of the objective function in (18). To obtain a measure of the uncertainty of the individual parameter estimates, the covariance matrix can be decomposed:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{\sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{R}\boldsymbol{\sigma}_{\hat{\boldsymbol{\theta}}} \tag{22}$$

into $\boldsymbol{\sigma}_{\hat{\boldsymbol{\theta}}}$, which is a diagonal matrix of the standard deviations of the parameter estimates, and $\boldsymbol{R}$, which is the corresponding correlation matrix. The asymptotic Gaussianity of the estimator in (18) also allows marginal t-tests to be performed to test the hypothesis:

$$H_0 : \theta_j = 0 \tag{23}$$

against the corresponding alternative:

$$H_1 : \theta_j \neq 0 \tag{24}$$

i.e. to test whether a given parameter θ_j is insignificant or not. The test quantity is the value of the parameter estimate $\hat{\theta}_j$ divided by the standard deviation of the estimate $\sigma_{\hat{\theta}_j}$ and under H_0 this quantity is asymptotically t-distributed with a number of degrees of freedom that equals the number of data points minus the number of estimated parameters, i.e.:

$$z(\hat{\theta}_j) = \frac{\hat{\theta}_j}{\sigma_{\hat{\theta}_j}} \in t\left(\sum_{i=1}^{S} N_i - p\right) \tag{25}$$

Due to correlations between the individual parameter estimates, a series of such marginal tests cannot be used to test the hypothesis that a subset of the parameters, $\boldsymbol{\theta}_* \subset \boldsymbol{\theta}$, are simultaneously insignificant:

$$H_0 : \boldsymbol{\theta}_* = \mathbf{0} \tag{26}$$

against the alternative that they are not:

$$H_1 : \boldsymbol{\theta}_* \neq \mathbf{0} \tag{27}$$

Hence a test that takes correlations into account must be used instead, e.g. a likelihood ratio test, a Lagrange multiplier test or a test based on Wald's W-statistic (Holst et al., 1992). Under H_0 the test quantities for these tests all have the same asymptotic χ^2-distribution with a number of degrees of freedom that equals the number of parameters subjected to the test (Holst et al., 1992). However, in the context of the proposed framework the test based on Wald's W-statistic has an advantage in that no re-estimation is required, because it can simply be computed as follows:

$$W(\hat{\boldsymbol{\theta}}_*) = \hat{\boldsymbol{\theta}}_*^{\mathrm{T}}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_*}^{-1}\hat{\boldsymbol{\theta}}_* \in \chi^2(\dim(\hat{\boldsymbol{\theta}}_*)) \tag{28}$$

where $\hat{\boldsymbol{\theta}}_* \subset \hat{\boldsymbol{\theta}}$ is the subset of the parameter estimates subjected to the test and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_*}$ is the corresponding covariance matrix, which can be computed as follows:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_*} = \boldsymbol{E}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{E}^{\mathrm{T}} \tag{29}$$

where $\boldsymbol{E}$ is a permutation matrix, which can be constructed from a unit matrix by eliminating the rows corresponding to parameter estimates not subjected to the test.

Remark 9. Strictly speaking, these tests can only be applied if the Gaussianity assumption mentioned in Section 2.2 holds, which is only likely to be the case if the structure of the model is appropriate, i.e. in the final iterations of the modelling cycle. Nevertheless, the corresponding test results can be used to provide indications for model improvement as shown in the following.

The above tests for insignificance provide the necessary framework for obtaining indications of which parts of the model that are deficient. In principle, *insignificant* parameters are parameters that may be eliminated, and the presence of such parameters is therefore an indication that the model is overparameterized. On the other hand, because of the particular nature of the model in (2) and (3), where the diffusion term is included to account for random effects due to, e.g. approximation errors or unmodelled phenomena, the presence of *significant* parameters in the diffusion term is an indication that the corresponding drift term may be incorrect, which in turn provides an uncertainty measure that allows model deficiencies to be detected. If, instead of the general parameterization of the diffusion term indicated in (2), a diagonal parameterization is used, this also allows the deficiencies to be pinpointed in the sense that deficiencies in specific elements of the drift term can be detected.

2.5.1. Pinpointing model deficiencies

If a diagonal parameterization of the diffusion term in (2) is used, the presence of significant parameters in a given diagonal element is an indication that the corresponding element of the drift term may be incorrect. This is valuable information for the model maker, as it indicates that some of the inherent phenomena of this term may be inappropriately modelled. If, by using physical insights, the model maker is able to subsequently select a specific phenomena model for further analysis, the proposed framework also provides means to confirm the suspicion that this model is inappropriate, if it is in fact true.

Typical suspect phenomena models include models of reaction rates, heat and mass transfer rates and similar complex dynamic phenomena, all of which can usually be described using functions of the state and input variables, i.e.:

$$r_t = \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}) \tag{30}$$

where r_t is a phenomenon of interest and $\varphi(\cdot) \in \mathbb{R}$ is the nonlinear function used by the model maker to describe it. To confirm the suspicion that $\varphi(\cdot)$ is inappropriate, the parameter estimation step must be repeated with a re-formulated version of the model in (2) and (3) to give new information.

More specifically, if r_t is isolated by including it in the re-formulated model as an additional state variable, i.e.:

$$\mathrm{d}\boldsymbol{x}_t^* = \boldsymbol{f}^*(\boldsymbol{x}_t^*, \boldsymbol{u}_t, t, \boldsymbol{\theta})\,\mathrm{d}t + \boldsymbol{\sigma}^*(\boldsymbol{u}_t, t, \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\omega}_t^* \tag{31}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k^*, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{32}$$

where $\boldsymbol{x}_t^* = [\boldsymbol{x}_t^{\mathrm{T}} \;\; r_t]^{\mathrm{T}}$, $\boldsymbol{\sigma}^*(\cdot) \in \mathbb{R}^{(n+1)\times(n+1)}$ and $\{\boldsymbol{\omega}_t^*\}$ is an $(n+1)$-dimensional standard Wiener process and where:

$$\boldsymbol{f}^*(\boldsymbol{x}_t^*, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) \\ \dfrac{\partial \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta})}{\partial \boldsymbol{x}_t}\dfrac{\mathrm{d}\boldsymbol{x}_t}{\mathrm{d}t} + \dfrac{\partial \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta})}{\partial \boldsymbol{u}_t}\dfrac{\mathrm{d}\boldsymbol{u}_t}{\mathrm{d}t} \end{pmatrix} \tag{33}$$

the presence of significant parameters in the corresponding diagonal element of the expanded diffusion term is a strong indication that $\varphi(\cdot)$ is inappropriate.

Remark 10. A particularly simple and very important special case of the above formulation is obtained if $\varphi(\cdot)$ is assumed to be constant, in which case the partial derivatives in (33) are both zero and any variation in r_t must be explained by the corresponding diagonal element of the expanded diffusion term, which in turn means that if the parameters of this element are significant, this is an indication that $\varphi(\cdot)$ is not constant.

2.6. Nonparametric modelling

In the sixth step of the proposed modelling cycle, which can only be used if specific model deficiencies have been pinpointed as described above, the idea is to uncover the structural origin of these deficiencies. The procedure for accomplishing this is based on a combination of the applicability of stochastic state space models for state estimation and the ability of nonparametric regression methods to provide visualizable estimates of unknown functional relations.

2.6.1. Estimating unknown functional relations

Using the re-formulated model in (31) and (32) and the corresponding parameter estimates, state estimates $\hat{\boldsymbol{x}}_{k|k}^*$, $k = 0, \ldots, N$, can be obtained for a given set of experimental data by applying the EKF. In particular, since the inappropriately modelled phenomenon r_t is included as an additional state variable in this model, estimates $\hat{r}_{k|k}$, $k = 0, \ldots, N$, can be obtained, which in turn facilitates application of nonparametric regression to provide estimates of possible functional relations between r_t and the state and input variables.

Several nonparametric regression techniques are available (Hastie, Tibshirani, & Friedman, 2001), but in the context of the proposed framework, *additive models* (Hastie & Tibshirani, 1990) are preferred, because fitting such models circumvents the curse of dimensionality, which tends to render nonparametric regression infeasible in higher dimensions, and because results obtained with such models are particularly easy to visualize, which is important.

Remark 11. Additive models are nonparametric extensions of linear regression models and are fitted by using a training data set of observations of several predictor variables $X_1, \ldots, X_n$ and a single response variable Y to compute a smoothed estimate of the response variable for a given set of values of the predictor variables. This is done by assuming that the contributions from each of the predictor variables are additive and can be fitted nonparametrically using the *backfitting algorithm* (Hastie & Tibshirani, 1990).

Using additive models, the variation in r_t can be decomposed into the variation that can be attributed to each of the state and input variables in turn, and the result can be visualized by means of partial dependence plots with associated bootstrap confidence intervals (Hastie et al., 2001). In this manner, it may be possible to reveal the true structure of the function describing r_t, i.e.:

$$r_t = \varphi_{\mathrm{true}}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}) \tag{34}$$

which in turn provides the model maker with valuable information about how to re-formulate the model for the next modelling cycle iteration. Needless to say, this should be done in accordance with physical insights.

Remark 12. The assumption of additive contributions does not necessarily limit the ability of additive models to reveal non-additive functional relations involving more than one predictor variable. By proper processing of the training data set, functions of more than one predictor variable, e.g. X_1X_2,

can be included as predictor variables as well (Hastie & Tibshirani, 1990).

2.7. An overall algorithm for systematic model improvement

In the following the methodologies from the various steps of the proposed modelling cycle are summarized in the form of an algorithm for systematic model improvement given a pre-specified purpose of the model:

(1) Use first engineering principles and physical insights to derive an initial model structure in the form of an ODE model (see Section 2.1).
(2) Translate the ODE model into a stochastic state space model using a diagonal parameterization of the diffusion term (see Section 2.1).
(3) Estimate the parameters of the model from available experimental data using ML or MAP estimation (see Section 2.2).
(4) Evaluate the quality of the model by performing residual analysis on cross-validation data (see Section 2.3).
(5) Determine if the model is sufficiently accurate to serve its intended purpose. If unfalsified, terminate model development. If falsified, proceed with model development (see Section 2.4).
(6) Try to pinpoint specific model deficiencies by applying statistical tests and by re-formulating the model with additional state variables and repeating the estimation and test procedures (see Section 2.5).
(7) If specific model deficiencies can be pinpointed, use state estimation and nonparametric modelling to uncover their structural origin by obtaining appropriate estimates of functional relations (see Section 2.6).
(8) Reformulate the model according to the estimated functional relations and physical insights and repeat from (3) (see Section 2.6).

This algorithm can be applied to develop new as well as to improve existing models of dynamic systems for a variety of purposes. More specifically, models can be developed with emphasis on short-term as well as long-term prediction capabilities, i.e. models intended for closed-loop as well as open-loop applications. However, as discussed in Section 4, the algorithm is not guaranteed to converge, especially not if insufficient prior information is available or if the quality and amount of available experimental data is limited. In particular, a situation may occur, where the model is falsified, but where none of the parameters of the diffusion term appear to be significant and pinpointing a specific model deficiency is impossible. A situation may also occur, where the model is falsified and the significance of certain parameters of the diffusion term have allowed a specific deficiency to be pinpointed, but where the structural origin of the deficiency cannot be uncovered. In the context of the proposed framework, both situations imply that a point has been reached, where the model cannot be further improved with the available information.

Remark 13. The estimation methods described in Section 2.2 (estimation in a PE setting) tend to emphasize the one-step-ahead prediction capabilities of the model and are therefore not ideal for models intended for open-loop applications. Nevertheless, these methods should be used in the development of such models as well, because of the possibility of using the tools described above for improving the structure of the model, if necessary, which would otherwise not be possible. Once an appropriate model structure has been obtained (ultimately corresponding to an insignificant diffusion term), the parameters can then be re-calibrated with an estimation method that emphasizes the pure simulation capabilities of the model (estimation in an OE setting).

3. Case study: modelling a fed-batch bioreactor

To illustrate the performance of the proposed framework in terms of improving the quality of an existing model, a simple simulation example is considered in the following. The process considered is a fed-batch bioreactor, where the true model used to simulate the process is given as follows:

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \mu(S)X - \frac{FX}{V} \tag{35}$$

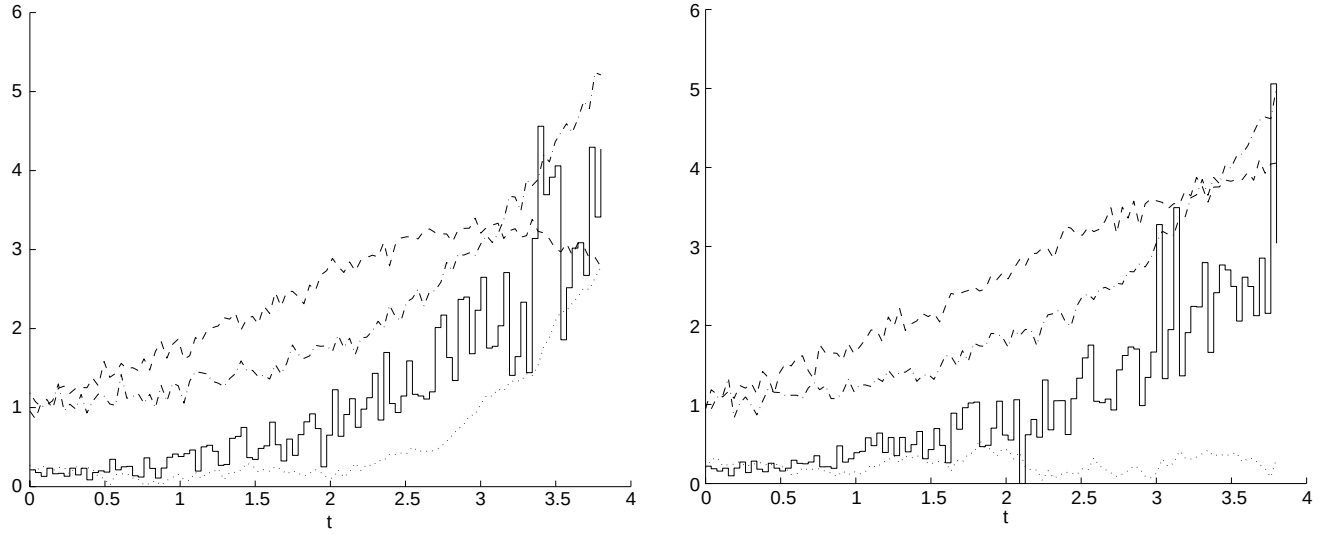
$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\mu(S)X}{Y} + \frac{F(S_\mathrm{F} - S)}{V} \tag{36}$$

$$\frac{\mathrm{d}V}{\mathrm{d}t} = F \tag{37}$$

where X is the biomass concentration, S the substrate concentration, V the volume, F the feed flow rate, $Y = 0.5$ the yield coefficient of biomass, $S_\mathrm{F} = 10$ the feed concentration of substrate, and $\mu(S)$ is the biomass growth rate, which is described by Monod kinetics with substrate inhibition, i.e.:

$$\mu(S) = \mu_{\max}\frac{S}{K_2S^2 + S + K_1} \tag{38}$$

where $\mu_{\max} = 1$, $K_1 = 0.03$ and $K_2 = 0.5$. Using $(X_0, S_0, V_0) = (1, 0.2449, 1)$ as initial states, simulated data sets from two batch runs (101 samples each) are generated by perturbing the feed flow rate along a pre-determined trajectory and subsequently adding Gaussian measurement noise to the appropriate variables using the noise levels mentioned beneath Fig. 2. In the following it is assumed that the model to be developed is to be used for an open-loop application, where long-term prediction capabilities are important, and that the model maker has been able to set up an initial model structure corresponding to (35)–(37) but is unaware of the true structure of $\mu(S)$ given in (38). In terms



(a) Batch no. 1. (b) Batch no. 2.

Fig. 2. The two batch data sets available for case 1. Solid staircase: feed flow rate F; dashed lines: biomass measurements y_1 (with $N(0, 0.01)$ noise); dotted lines: substrate measurements y_2 (with $N(0, 0.001)$ noise); dash-dotted lines: volume measurements y_3 (with $N(0, 0.01)$ noise)).

of available measurements, two different cases are considered: A full state information case, where it is assumed that all state variables can be measured, and a partial state information case, where it is assumed that only the biomass and the volume can be measured.

3.1. Case 1: full state information

The available sets of experimental data for the full state information case are shown in Fig. 2. Using these data sets it will now be illustrated how the proposed modelling cycle can be used to improve the initial model set up by the model maker. In this particular case only two iterations of the modelling cycle are needed. In the general case more iterations may be needed.

3.1.1. First modelling cycle iteration

Model formulation. The first iteration of the modelling cycle starts with the model formulation step, where it is assumed that the model maker has been able to set up an initial model structure corresponding to (35)–(37), which is then translated into a stochastic state space model with the following system equation:

$$\mathrm{d}\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu X - \dfrac{FX}{V} \\ -\dfrac{\mu X}{Y} + \dfrac{F(S_\mathrm{F} - S)}{V} \\ F \end{pmatrix} \mathrm{d}t + \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} \mathrm{d}\boldsymbol{\omega}_t \tag{39}$$

and the following measurement equation:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}_k = \begin{pmatrix} X \\ S \\ V \end{pmatrix}_k + \boldsymbol{e}_k, \quad \boldsymbol{e}_k \in \mathcal{N}(\mathbf{0}, \boldsymbol{S}),$$
$$\boldsymbol{S} = \begin{bmatrix} S_{11} & 0 & 0 \\ 0 & S_{22} & 0 \\ 0 & 0 & S_{33} \end{bmatrix} \tag{40}$$

where, because the true structure of $\mu(S)$ given in (38) is unknown, a constant growth rate μ has been assumed. As recommended above, a diagonal parameterization of the diffusion term in the system equation has been used to allow model deficiencies to be pinpointed if the model is falsified.

Parameter estimation. As the next step, the unknown parameters of the model in (39) and (40) are estimated with the ML method using the data from batch no. 1 (Fig. 2a), which gives the results shown in Table 1.

Residual analysis. Evaluating the quality of the resulting model is the next step. Pure simulation residual analysis is therefore performed as shown in Fig. 3, and the results of this show that the model does a poor job, particularly for y_1 and y_2.

Model falsification or unfalsification. Moving to the model falsification or unfalsification step, the poor pure simulation capabilities falsify the model for its intended purpose, which means that the modelling cycle must be repeated by re-formulating the model.

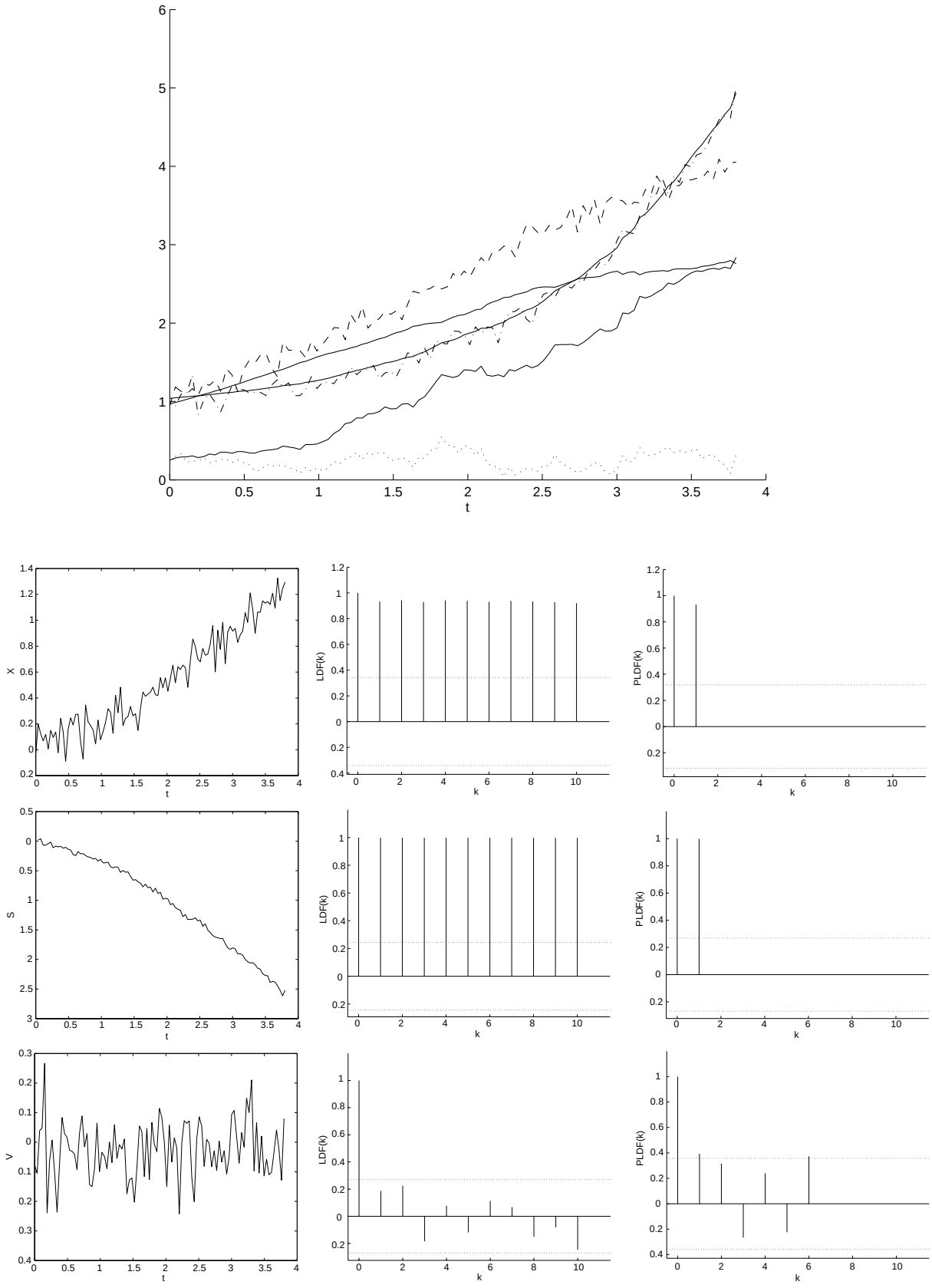


Fig. 3. Pure simulation residual analysis for the model in (39) and (40) with parameters in Table 1 using data from batch no. 2 (Fig. 2b). Top: comparison (solid lines are simulated values); bottom: residuals, LDF and PLDF for y_1, y_2 and y_3.

Table 1
Estimation results. Model in (39)–(40). Data from batch no. 1

Parameter	Estimate	S.D.	t-Score	Significant?
X_0	9.6973E−01	3.4150E−02	28.3962	Yes
S_0	2.5155E−01	3.1938E−02	7.8761	Yes
V_0	1.0384E+00	1.8238E−02	56.9359	Yes
μ	6.8548E−01	2.2932E−02	29.8921	Yes
σ_{11}	1.8411E−01	2.5570E−02	7.2000	Yes
σ_{22}	2.2206E−01	3.4209E−02	6.4912	Yes
σ_{33}	2.7979E−02	1.7943E−02	1.5594	No
S_{11}	6.7468E−03	1.3888E−03	4.8580	Yes
S_{22}	3.9131E−04	2.4722E−04	1.5828	No
S_{33}	1.0884E−02	1.5409E−03	7.0633	Yes

Statistical tests. To obtain information about how to re-formulate the model in an intelligent way, model deficiencies should be pinpointed, if possible. Table 1 also includes t-scores for performing marginal tests for significance of the individual parameters, which show that, on a 5% level, only one of the parameters of the diffusion term is insignificant, viz. σ_{33}, whereas σ_{11} and σ_{22} are both significant, which indicates that the first two elements of the drift term may be incorrect. These elements both depend on μ and a skilled model maker, who knows how difficult it is to model complex dynamic phenomena such as growth rates, would immediately suspect μ to be deficient. To avoid jumping to conclusions, the suspicion should be confirmed, which is done by first re-formulating the model with μ as an additional state variable, which yields the system equation:

$$\mathrm{d}\begin{pmatrix} X \\ S \\ V \\ \mu \end{pmatrix} = \begin{pmatrix} \mu X - \dfrac{FX}{V} \\ -\dfrac{\mu X}{Y} + \dfrac{F(S_\mathrm{F} - S)}{V} \\ F \\ 0 \end{pmatrix} \mathrm{d}t + \begin{bmatrix} \sigma_{11} & 0 & 0 & 0 \\ 0 & \sigma_{22} & 0 & 0 \\ 0 & 0 & \sigma_{33} & 0 \\ 0 & 0 & 0 & \sigma_{44} \end{bmatrix} \mathrm{d}\boldsymbol{\omega}_t \tag{41}$$

where, because μ has been assumed to be constant, the last element of the drift term is zero. The measurement equation is the same as in (40).

Estimating the parameters of this model, using the same data set as before, gives the results shown in Table 2, and inspection of the t-scores for marginal tests for insignificance now show that, of the parameters of the diffusion term, only σ_{44} is significant on a 5% level. This in turn indicates that there is substantial variation in μ and thus confirms the suspicion that μ is deficient.

Nonparametric modelling. Having pinpointed μ as being deficient, nonparametric modelling can be applied as the next step to uncover the structural origin of the deficiency.

Table 2
Estimation results. Model in (41) and (40). Data from batch no. 1

Parameter	Estimate	S.D.	t-Score	Significant?
X_0	1.0239E+00	4.9566E−03	206.5723	Yes
S_0	2.3282E−01	1.1735E−02	19.8405	Yes
V_0	1.0099E+00	3.8148E−03	264.7290	Yes
μ_0	7.8658E−01	2.4653E−02	31.9061	Yes
σ_{11}	2.0791E−18	1.4367E−17	0.1447	No
σ_{22}	1.1811E−30	1.6162E−29	0.0731	No
σ_{33}	3.1429E−04	2.0546E−04	1.5297	No
σ_{44}	1.2276E−01	2.5751E−02	4.7674	Yes
S_{11}	7.5085E−03	9.9625E−04	7.5368	Yes
S_{22}	1.1743E−03	1.6803E−04	6.9887	Yes
S_{33}	1.1317E−02	1.3637E−03	8.2990	Yes

Using the re-formulated model in (40) and (41) and the parameter estimates in Table 2, state estimates $\hat{X}_{k|k}$, $\hat{S}_{k|k}$, $\hat{V}_{k|k}$, $\hat{\mu}_{k|k}$, $k = 0, \ldots, N$, are obtained by means of the EKF and an additive model is fitted to reveal the true structure of the function describing μ by means of estimates of functional relations between μ and the state and input variables.

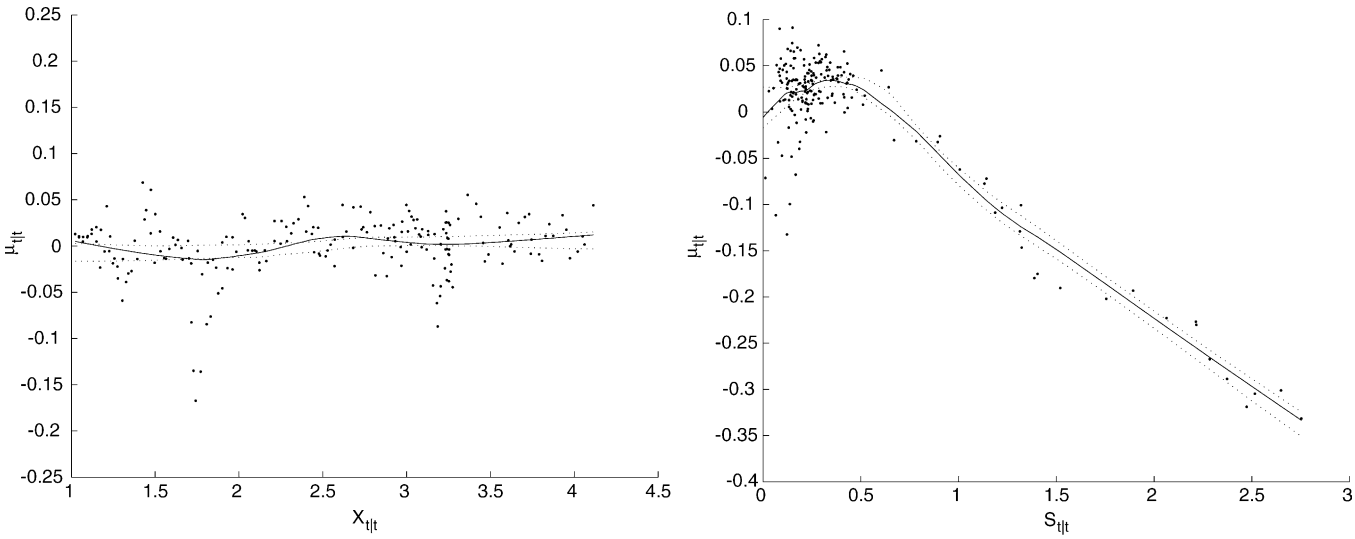
It is reasonable to assume that μ does not depend on V and F, so only functional relations between $\hat{\mu}_{k|k}$ and $\hat{X}_{k|k}$ and $\hat{S}_{k|k}$ are estimated, giving the results shown in Fig. 4 in the form of partial dependence plots with associated bootstrap confidence intervals. These plots indicate that $\hat{\mu}_{k|k}$ does not depend on $\hat{X}_{k|k}$, but is highly dependent on $\hat{S}_{k|k}$, which in turn suggests to replace the assumption of constant μ with an assumption of μ being a function of S when the model is re-formulated for the next iteration of the modelling cycle. More specifically, this function should comply with the functional relation revealed in the partial dependence plot between $\hat{\mu}_{k|k}$ and $\hat{S}_{k|k}$.

3.1.2. Second modelling cycle iteration

Model re-formulation. To a skilled model maker with experience in bioreactor modelling, the functional relation revealed in the partial dependence plot between $\hat{\mu}_{k|k}$ and $\hat{S}_{k|k}$ in Fig. 4 is a clear indication that the growth of biomass is governed by Monod kinetics and inhibited by substrate, which in the first step of the second iteration of the modelling cycle makes it possible to re-formulate the model in (39) and (40) accordingly to yield the system equation:

$$\mathrm{d}\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu(S)X - \dfrac{FX}{V} \\ -\dfrac{\mu(S)X}{Y} + \dfrac{F(S_\mathrm{F} - S)}{V} \\ F \end{pmatrix} \mathrm{d}t + \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} \mathrm{d}\boldsymbol{\omega}_t \tag{42}$$

where $\mu(S)$ is given by the true structure in (38). The measurement equation remains the same as in (40).



(a) $\hat{\mu}_{k|k}$ vs. $\hat{X}_{k|k}$ (b) $\hat{\mu}_{k|k}$ vs. $\hat{S}_{k|k}$

Fig. 4. Partial dependence plots of $\hat{\mu}_{k|k}$ vs. $\hat{X}_{k|k}$ and $\hat{S}_{k|k}$. Solid lines: estimates; dotted lines: 95% bootstrap confidence intervals (1000 replicates).

Parameter estimation. As the next step, estimation of the unknown parameters of the re-formulated model using the same data set as before gives the results shown in Table 3.

Residual analysis. Evaluating the quality of the resulting model as the next step, pure simulation residual analysis is performed as shown in Fig. 5, and the results of this show that the re-formulated model does a very good job.

Model falsification or unfalsification. Moving to the model falsification or unfalsification step, the re-formulated model is thus unfalsified for its intended purpose with respect to the available information, and the model development procedure can now be terminated. As the intended purpose of the model is to use it for an open-loop application, the parameters should ideally be re-calibrated at this point[1] with an estimation method that emphasizes the pure simulation capabilities of the model, but this is outside the scope of the present paper. This therefore concludes the full state information case.

3.2. Case 2: partial state information

To illustrate that the proposed modelling cycle can also be used when only a subset of the state variables can be measured, the previous example is repeated with the assumption that only the biomass and the volume can be measured. The available sets of experimental data for this partial state information case are shown in Fig. 6, and, otherwise, the same assumptions apply with respect to the intended purpose of the model and the availability of an initial model structure, where the growth rate is unknown.

Table 3
Estimation results. Model in (42) and (40). Data from batch no. 1

Parameter	Estimate	S.D.	t-Score	Significant?
X_0	1.0148E+00	1.0813E−02	93.8515	Yes
S_0	2.4127E−01	9.4924E−03	25.4177	Yes
V_0	1.0072E+00	8.7723E−03	114.8168	Yes
$\mu_{\max}$	1.0305E+00	1.7254E−02	59.7225	Yes
K_1	3.7929E−02	4.1638E−03	9.1092	Yes
K_2	5.4211E−01	2.4949E−02	21.7286	Yes
σ_{11}	2.3250E−10	2.1044E−07	0.0011	No
σ_{22}	1.4486E−07	7.9348E−05	0.0018	No
σ_{33}	3.2842E−12	3.6604E−09	0.0009	No
S_{11}	7.4828E−03	1.0114E−03	7.3982	Yes
S_{22}	1.0433E−03	1.4331E−04	7.2804	Yes
S_{33}	1.1359E−02	1.6028E−03	7.0867	Yes

3.2.1. First modelling cycle iteration

Model formulation. The first iteration of the modelling cycle again starts with the model formulation step, where it is assumed that the model maker has been able to set up an initial model structure corresponding to (35)–(37), which is translated into a stochastic state space model with the following system equation:

$$\mathrm{d}\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu X - \dfrac{FX}{V} \\ -\dfrac{\mu X}{Y} + \dfrac{F(S_{\mathrm{F}} - S)}{V} \\ F \end{pmatrix} \mathrm{d}t + \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} \mathrm{d}\boldsymbol{\omega}_t \quad (43)$$

[1] Inspection of the t-scores for marginal tests for insignificance (Table 3) suggest that, on a 5% level, there are no significant parameters in the diffusion term, which is confirmed by a test for simultaneous insignificance based on Wald's W-statistic.

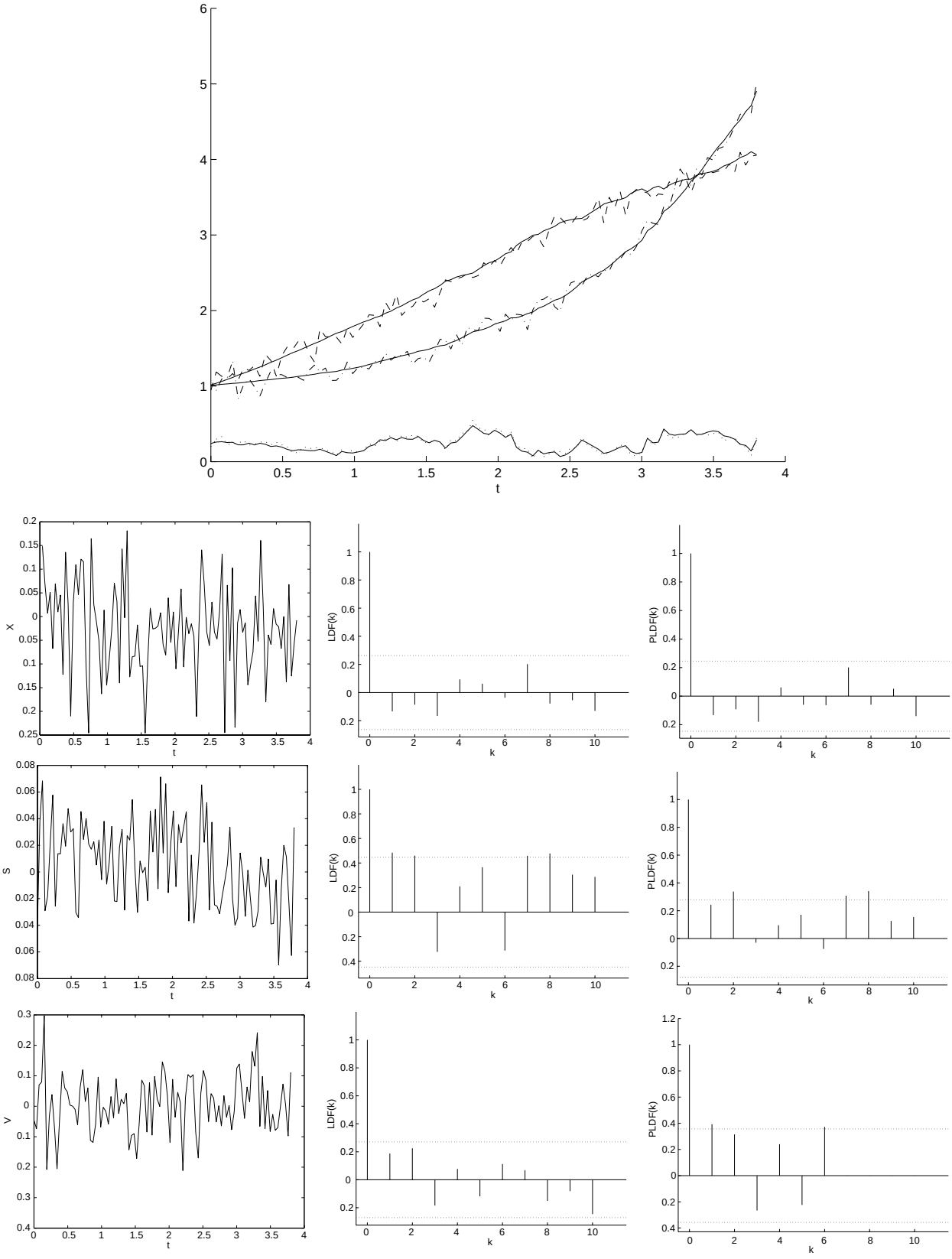
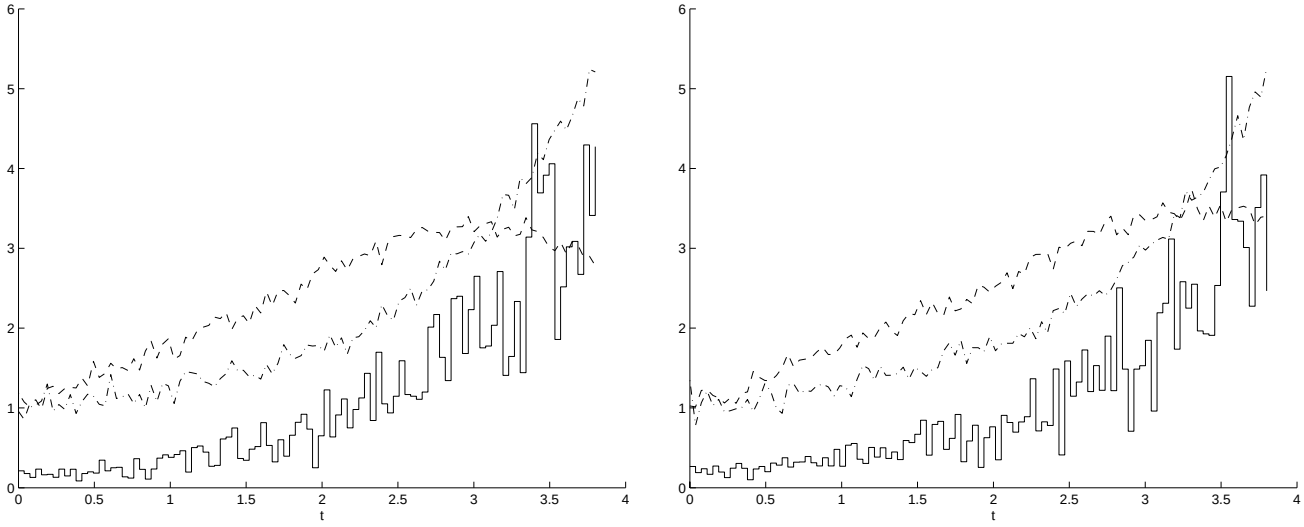


Fig. 5. Pure simulation residual analysis for the model in (40) and (42) with parameters in Table 3 using data from batch no. 2 (Fig. 2b). Top: comparison (solid lines are simulated values); bottom: residuals, LDF and PLDF for y_1, y_2 and y_3.



(a) Batch no. 1. (b) Batch no. 2.

Fig. 6. The two batch data sets available for case 2. Solid staircase: feed flow rate F; dashed lines: biomass measurements y_1 (with $N(0, 0.01)$ noise); dash-dotted lines: volume measurements y_2 (with $N(0, 0.01)$ noise)).

and the following modified measurement equation:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}_k = \begin{pmatrix} X \\ V \end{pmatrix}_k + \boldsymbol{e}_k, \quad \boldsymbol{e}_k \in \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}),$$
$$\boldsymbol{S} = \begin{bmatrix} S_{11} & 0 \\ 0 & S_{22} \end{bmatrix} \tag{44}$$

where, because the true structure of $\mu(S)$ given in (38) is unknown, a constant growth rate μ has again been assumed.

Parameter estimation. Estimating the unknown parameters of the model in (43) and (44) using the data from batch no. 1 (Fig. 6a) gives the results shown in Table 4.

Residual analysis. Evaluating the quality of the resulting model, the pure simulation residual analysis results in Fig. 7 shows that the model does a poor job.

Model falsification or unfalsification. Again the model is thus falsified for its intended purpose, and the modelling cycle must be repeated by re-formulating the model once its deficiencies have been pinpointed, if possible.

Table 4
Estimation results. Model in (43)–(44). Data from batch no. 1

Parameter	Estimate	S.D.	t-Score	Significant?
X_0	9.6230E−01	1.2996E−02	74.0451	Yes
V_0	1.0272E+00	2.1417E−02	47.9641	Yes
μ	6.8730E−01	2.1875E−02	31.4198	Yes
σ_{11}	1.8846E−01	3.9179E−02	4.8104	Yes
σ_{22}	8.7290E−03	1.8577E−03	4.6989	Yes
σ_{33}	1.7391E−02	1.5107E−02	1.1512	No
S_{11}	6.7225E−03	1.0795E−03	6.2273	Yes
S_{22}	1.1078E−02	1.5137E−03	7.3184	Yes

Statistical tests. Table 4 also includes t-scores for performing marginal tests for significance of the individual parameters, and, as in the full state information case, these show that, on a 5% level, only σ_{33} is insignificant, whereas the other parameters of the diffusion term are both significant. This indicates that the first two elements of the drift term may be incorrect, and hence that μ is a possible suspect for being deficient. To confirm this suspicion the model is re-formulated with μ as an additional state variable to yield the system equation:

$$\mathrm{d}\begin{pmatrix} X \\ S \\ V \\ \mu \end{pmatrix} = \begin{pmatrix} \mu X - \dfrac{FX}{V} \\ -\dfrac{\mu X}{Y} + \dfrac{F(S_\mathrm{F} - S)}{V} \\ F \\ 0 \end{pmatrix} \mathrm{d}t + \begin{bmatrix} \sigma_{11} & 0 & 0 & 0 \\ 0 & \sigma_{22} & 0 & 0 \\ 0 & 0 & \sigma_{33} & 0 \\ 0 & 0 & 0 & \sigma_{44} \end{bmatrix} \mathrm{d}\boldsymbol{\omega}_t \tag{45}$$

and the measurement equation in (44). The parameters of this model are estimated using the same data set as before to give the results shown in Table 5, and inspection of the t-scores again show that only σ_{44} is now significant on a 5% level, which in turn indicates that there is substantial variation in μ and thus confirms the suspicion that μ is deficient.

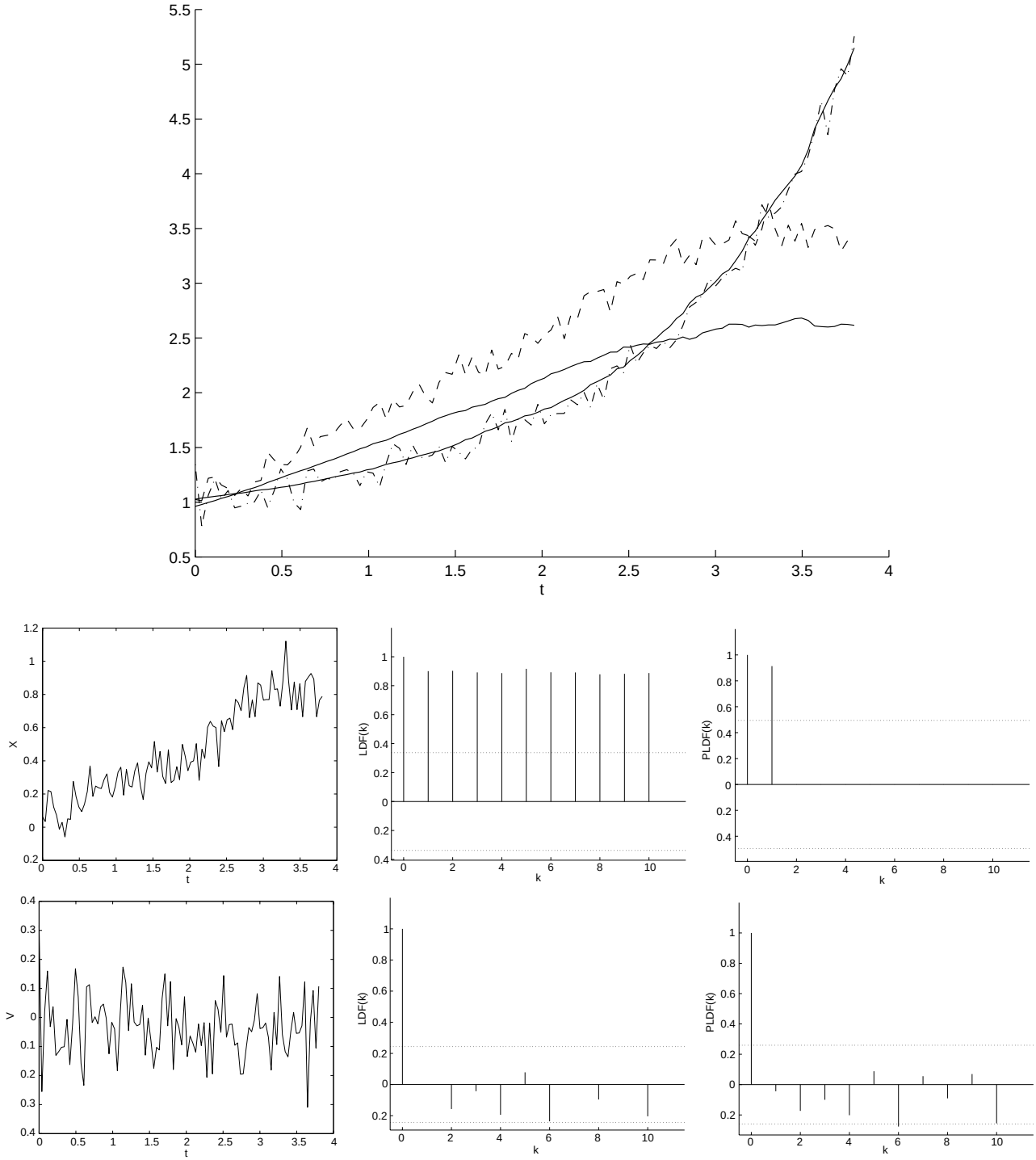


Fig. 7. Pure simulation residual analysis for the model in (43) and (44) with parameters in Table 4 using data from batch no. 2 (Fig. 6b). Top: comparison (solid lines are simulated values); bottom: residuals, LDF and PLDF for y_1 and y_2.

Nonparametric modelling. The structural origin of the deficiency can again be uncovered by using the re-formulated model in (44) and (45) and the parameter estimates in Table 5 to obtain state estimates $\hat{X}_{k|k}$, $\hat{S}_{k|k}$, $\hat{V}_{k|k}$, $\hat{\mu}_{k|k}$, $k = 0, \ldots, N$, and by fitting an additive model to reveal the true structure of the function describing μ.

Assuming again that μ does not depend on V and F, the partial dependence plots shown in Fig. 8 are obtained. In this case there seems to be a dependence between $\hat{\mu}_{k|k}$ and both $\hat{X}_{k|k}$ and $\hat{S}_{k|k}$. However, since the dependence on $\hat{S}_{k|k}$ is much stronger than the dependence on $\hat{X}_{k|k}$, this again suggests to replace the assumption of constant μ with an

Table 5
Estimation results. Model in (45) and (44). Data from batch no. 1

Parameter	Estimate	S.D.	t-Score	Significant?
X_0	1.0069E+00	2.1105E−02	47.7095	Yes
V_0	1.0250E+00	2.7800E−02	36.8687	Yes
μ_0	8.1305E−01	1.2223E−01	6.6516	Yes
σ_{11}	8.5637E−05	5.5485E−05	1.5434	No
σ_{22}	8.2654E−03	8.5005E−03	0.9723	No
σ_{33}	1.5241E−02	2.4948E−02	0.6109	No
σ_{44}	1.4751E−01	4.5181E−02	3.2648	Yes
S_{11}	7.7509E−03	1.1338E−03	6.8362	Yes
S_{22}	1.1118E−02	1.5652E−03	7.1033	Yes

assumption of μ being a function of S when the model is re-formulated for the next iteration.

3.2.2. Second modelling cycle iteration

Model re-formulation. Although less obvious, the functional relation revealed in the partial dependence plot between $\hat{\mu}_{k|k}$ and $\hat{S}_{k|k}$ in Fig. 8, is again an indication to a skilled model maker that the growth rate of biomass can be appropriately described with Monod kinetics and substrate inhibition, which allows the model to be re-formulated to yield the system equation:

$$\mathrm{d}\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu(S)X - \dfrac{FX}{V} \\ -\dfrac{\mu(S)X}{Y} + \dfrac{F(S_\mathrm{F} - S)}{V} \\ F \end{pmatrix} \mathrm{d}t + \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix} \mathrm{d}\boldsymbol{\omega}_t \quad (46)$$

Table 6
Estimation results. Model in (46) and (44). Data from batch no. 1

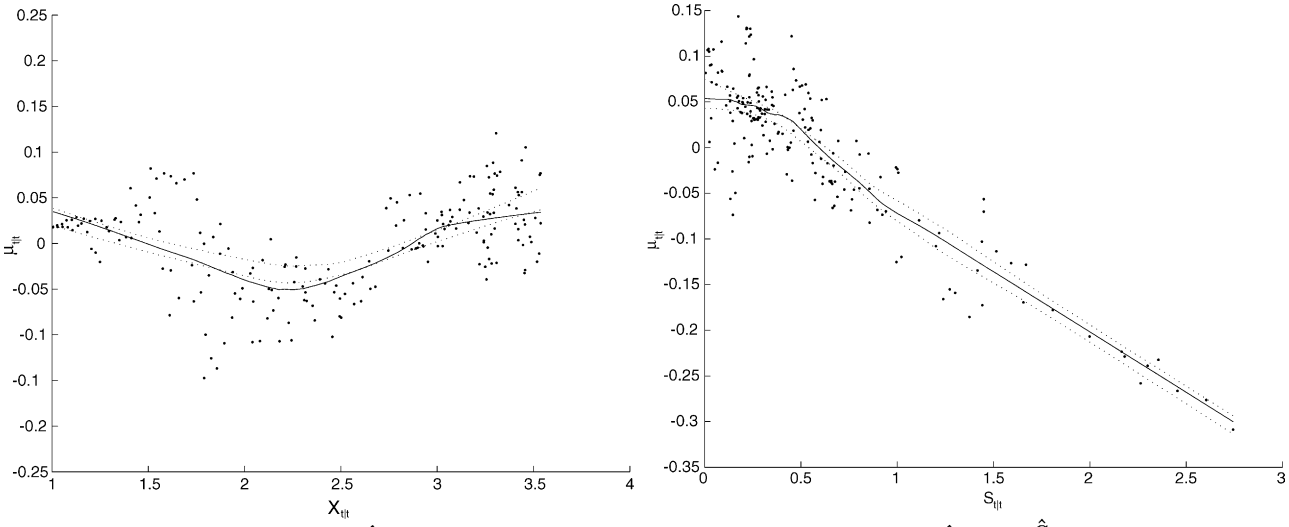
Parameter	Estimate	S.D.	t-Score	Significant?
X_0	1.0137E+00	1.6790E−02	60.3759	Yes
V_0	1.0118E+00	1.1571E−02	87.4443	Yes
$\mu_{\max}$	1.0679E+00	1.4353E−01	7.4405	Yes
K_1	4.1664E−02	3.2800E−02	1.2702	No
K_2	6.3372E−01	1.8116E−01	3.4980	Yes
σ_{11}	6.8577E−11	2.2270E−08	0.0031	No
σ_{22}	7.9677E−06	1.1223E−03	0.0071	No
σ_{33}	1.4241E−07	2.6577E−05	0.0054	No
S_{11}	7.4094E−03	1.0986E−03	6.7447	Yes
S_{22}	1.1364E−02	1.6193E−03	7.0174	Yes

where $\mu(S)$ is given by the true structure in (38), while the measurement equation remains the same as in (44).

Parameter estimation. Estimating the unknown parameters of the re-formulated model using the same data set as before gives the results shown in Table 6.

Residual analysis. Examining the pure simulation residual analysis results shown in Fig. 9, there still seems to be some non-random variation left in the cross-validation data set that is not explained by the model. This may be attributed to the fact that the data set used for parameter estimation and the cross-validation data set cover different regions of state space, which, because only partial state information is available, the model is more sensitive to in this case.

Model falsification or unfalsification. In principle, although the results obtained with the re-formulated model are



(a) $\hat{\mu}_{k|k}$ vs. $\hat{X}_{k|k}$ (b) $\hat{\mu}_{k|k}$ vs. $\hat{S}_{k|k}$

Fig. 8. Partial dependence plots of $\hat{\mu}_{k|k}$ vs. $\hat{X}_{k|k}$ and $\hat{S}_{k|k}$. Solid lines: estimates; dotted lines: 95% bootstrap confidence intervals (1000 replicates).

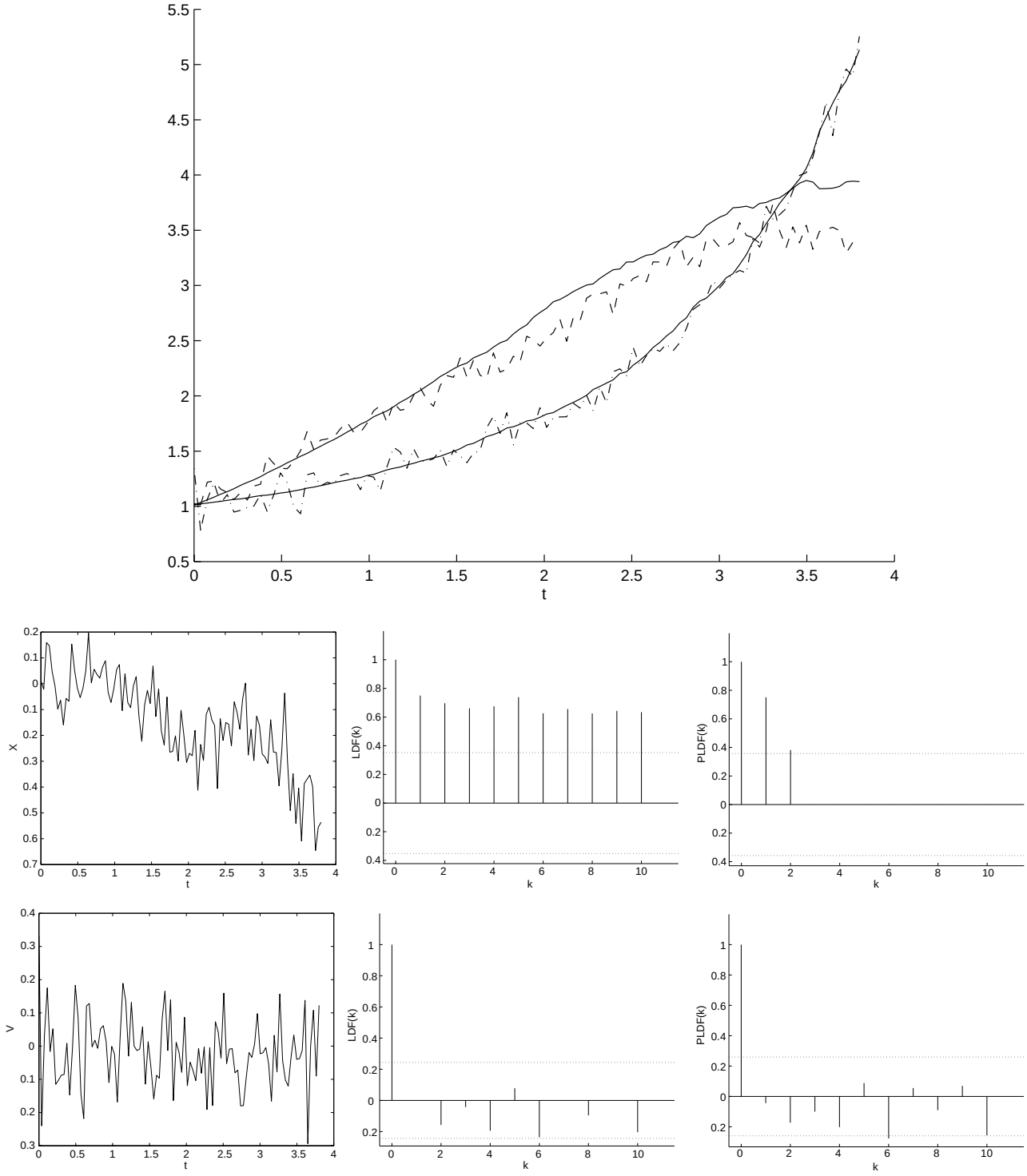


Fig. 9. Pure simulation residual analysis for the model in (44) and (46) with parameters in Table 6 using data from batch no. 2 (Fig. 6b). Top: comparison (solid lines are simulated values); bottom: residuals, LDF and PLDF for y_1 and y_2.

much better than those obtained with the initial model, the re-formulated model is thus falsified for its intended purpose, and the modelling cycle should be repeated by re-formulating the model again. However, in the context of the proposed framework, all information available in the data set used for estimation has been exhausted, because a model has been developed where the diffusion term is insignificant.[2] In other words it is not possible to pinpoint any model deficiencies directly, because these deficiencies

[2] Inspection of the t-scores for marginal tests for insignificance (Table 6) suggest that, on a 5% level, there are no significant parameters in the diffusion term, which is confirmed by a test for simultaneous insignificance based on Wald's W-statistic.

are only revealed by the cross-validation data set and not by the data set used for estimation. Ideally, the parameters of the model should thus be re-estimated using the cross-validation data set as well before re-formulating the model, but this takes away the possibility of easily evaluating the quality of the resulting model through cross-validation, unless more data is obtained. A discussion of possible ways to resolve this issue is outside the scope of the present paper, and this thus concludes the partial state information case.

4. Discussion

The case study presented in the previous section illustrates the strength of the proposed stochastic grey-box modelling framework in terms of facilitating systematic model improvement. A key feature in this regard is the ability to pinpoint and subsequently uncover the structural origin of model deficiencies by means of estimates of unknown functional relations, and another key result is that this is also possible in situations where all process variables cannot be measured. More specifically, the full state information case demonstrates that a high quality estimate of the functional relation between the biomass growth rate, which cannot be measured, and the substrate concentration, which is measured, can easily be obtained, and the partial state information case demonstrates that a similar estimate, of lower quality, can be obtained without measuring the substrate concentration.

The lower quality of the estimate obtained in the partial state information case is due to the fact that the performance of the proposed framework is limited by the quality and amount of available experimental data, in the sense that, if the available data is insufficiently informative, e.g. due to large measurement noise, or if the available measurements render certain subsets of the state variables of the system unobservable, parameter identifiability and hence the reliability of the proposed methods for pinpointing and uncovering the structural origin of model deficiencies is affected. Experimental design and selection of appropriate measurements are thus key issues that must also be addressed in model development, but these are outside the scope of the present paper. The performance of the proposed stochastic grey-box modelling framework is also limited by the quality and amount of available prior information, and if there is insufficient information to establish an initial model structure, it may not be worthwhile to use this approach as opposed to a black-box modelling approach. Furthermore, the model maker must be able to determine the specific phenomenon causing a pinpointed model deficiency in order to uncover its structural origin, and this may not always be possible either. If, however, sufficient prior information and experimental data is available, the proposed framework is very powerful as a tool for systematic model improvement. In particular, it relies less on the model maker than other approaches to stochastic grey-box modelling (Bohlin & Graebe, 1995; Bohlin, 2001) and also prevents him or her from having to resort to using black-box models for filling gaps in the model. This is due to the fact that estimates of unknown functional relations can be obtained and visualized directly.

The proposed framework may be seen as a stochastic grey-box model generalization of the well-developed methodologies for identification of linear black-box models (Box & Jenkins, 1976; Ljung, 1987; Söderström & Stoica, 1989). However, unlike in the linear case, where convergence is guaranteed if certain conditions of identifiability of parameters and persistency of excitation of inputs are fulfilled, no rigorous proof of convergence is available for the framework proposed here. Nevertheless, the case study presented in the previous section has demonstrated that the proposed framework can indeed be used to obtain valuable information to facilitate faster model development.

5. Conclusion

A systematic framework for improving the quality of continuous time models of dynamic systems based on experimental data has been presented. The proposed stochastic grey-box modelling framework is based on an interplay between stochastic differential equation modelling, statistical tests and nonparametric modelling and provides features that allow model deficiencies to be pinpointed and their structural origin to be uncovered to improve the model. A key result in this regard is that the proposed framework can be used to obtain nonparametric estimates of unknown functional relations, which allows unknown or inappropriately modelled phenomena to be uncovered and proper parametric expressions to be inferred from the estimated functional relations. The performance of the proposed framework has been illustrated through a case study involving a dynamic model of a fed-batch bioreactor, where it has been shown how an inappropriately modelled biomass growth rate can be uncovered and a proper parametric expression inferred. A key point illustrated through this case study is that estimates of functional relations involving only unmeasured variables can also be obtained.

References

Allgöwer, F., & Zheng, A. (Eds.). (2000). Nonlinear model predictive control. In *Progress in systems & control theory* (Vol. 26). Switzerland: Birkhauser Verlag.

Åström, K. J. (1970). *Introduction to stochastic control theory*. New York, USA: Academic Press.

Bak, J., Madsen, H., & Nielsen, H. A. (1999). Goodness of fit of stochastic differential equations. In P. Linde, A. Holm (Eds.), *Symposium i Anvendt Statistik*. Copenhagen, Denmark: Copenhagen Business School.

Bohlin, T. (2001). A Grey-Box Process Identification Tool: Theory and Practice. Technical Report IR-S3-REG-0103, Department of Signals, Sensors and Systems, Royal Institute of Technology, Stockholm, Sweden.

Bohlin, T., & Graebe, S. F. (1995). Issues in nonlinear stochastic grey-box identification. *International Journal of Adaptive Control and Signal Processing, 9*, 465–490.

Box, G. E. P., & Jenkins, J. M. (1976). *Time series analysis: forecasting and control*. San Francisco, USA: Holden-Day.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: theory and methods* (2nd ed.). New York, USA: Springer-Verlag.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London, England: Chapman & Hall.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. (2001). *The elements of statistical learning—data mining, inference and prediction*. New York, USA: Springer-Verlag.

Holst, J., Holst, U., Madsen, H., & Melgaard, H. (1992). Validation of grey box models. In L. Dugard, M. M'Saad, & I. D. Landau (Eds.), *Selected papers from the fourth IFAC symposium on adaptive systems in control and signal processing* (pp. 407–414). Oxford: Pergamon Press.

Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. New York, USA: Academic Press.

Kristensen, N. R., Madsen, H., & Jørgensen, S. B. (2003). *Parameter estimation in stochastic grey-box models*. *Automatica*, in press.

Ljung, L. (1987). *System identification: theory for the user*. New York, USA: Prentice-Hall.

Madsen, H., & Melgaard, H. (1991). The Mathematical and Numerical Methods Used in CTLSM. Technical Report 7, IMM, Technical University of Denmark, Lyngby, Denmark.

Melgaard, H., & Madsen, H. (1993). CTLSM—A Program for Parameter Estimation in Stochastic Differential Equations. Technical Report1, IMM, Technical University of Denmark, Lyngby, Denmark.

Nielsen, H. A., & Madsen, H. (2001). A generalization of some classical time series tools. *Computational Statistics and Data Analysis, 37*(1), 13–31.

Øksendal, B. (1998). *Stochastic differential equations—an introduction with applications* (5th ed.). Berlin, Germany: Springer-Verlag.

Raisch, J. (2000). Complex systems—simple models? In L. T. Biegler, A. Brambilla, C. Scali, & G. Marchetti (Eds.), *Proceedings of the IFAC symposium on advanced control of chemical processes* (pp. 275–286). Amsterdam: Elsevier.

Söderström, T., & Stoica, P. (1989). *System identification*. New York, USA: Prentice-Hall.

Young, P. C. (1981). Parameter estimation for continuous-time models—a survey. *Automatica, 17*(1), 23–39.