STOCHASTIC GREY-BOX MODELLING AS A TOOL FOR IMPROVING THE QUALITY OF FIRST ENGINEERING PRINCIPLES MODELS

Niels Rode Kristensen [*,1] **Henrik Madsen** [**]
Sten Bay Jørgensen [*]

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

Abstract: A systematic framework for improving the quality of first engineering principles models using experimental data is presented. The framework is based on stochastic grey-box modelling and incorporates statistical tests and nonparametric regression in a manner that permits systematic iterative model improvement. More specifically, the proposed framework provides features that allow model deficiencies to be pinpointed and their structural origin to be uncovered through estimation of unknown functional relations. The performance of the proposed framework is illustrated through a case study involving a model of a fed-batch bioreactor, where it is shown how an incorrectly modelled biomass growth rate can be uncovered and a more appropriate functional relation inferred.

Keywords: Stochastic modelling, parameter estimation, statistical inference, nonparametric regression, iterative improvement.

1. INTRODUCTION

Dynamic model development is an inherently purpose-driven act in the sense that the required accuracy of a model depends on its intended application, and developing a suitable model for a given purpose involves a fundamental trade-off between model accuracy and model simplicity (Raisch, 2000). For models intended for simulation and optimisation purposes, which must be valid over wide ranges of state space, the required model accuracy and hence the necessary model complexity is high, which means that developing such models is potentially time-consuming.

Ordinary differential equation (ODE) models developed from first engineering principles and physical insights are typically used for such purposes and a common problem with the development of such models is that only the basic structure of the model can be determined directly from first engineering principles, whereas a number of constitutive equations describing e.g. reaction kinetics often remain to be determined from experimental data, which may be difficult. Furthermore, if the quality of a model of this type proves to be too low, few systematic methods are available for determining how to improve the model. Altogether, this often renders the development of first engineering principles models very time-consuming.

[1] Corresponding author.

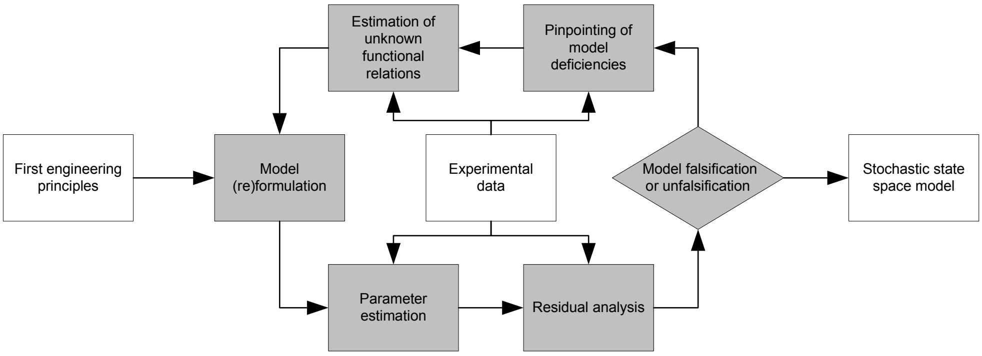


Fig. 1. The proposed grey-box modelling cycle. The boxes in grey illustrate tasks and the boxes in white illustrate inputs to and outputs from the modelling cycle.

In the present paper stochastic grey-box modelling is proposed as a tool for systematic improvement of first engineering principles models, as this approach resolves some of the issues mentioned above. In particular, the proposed framework facilitates pinpointing of model deficiencies and provides means to subsequently uncover the structural origin of these deficiencies through estimation of unknown functional relations. To obtain these estimates nonparametric modelling is applied, and the integration of nonparametric modelling with conventional stochastic grey-box modelling into a systematic framework for improving the quality of first engineering principles models is the key new contribution of the paper.

The remainder of the paper is organized as follows: In Section 2 the proposed framework is outlined and in Section 3 a case study demonstrating its performance is given. Finally, in Section 4, the conclusions of the paper are presented.

2. METHODOLOGY

A diagram of the proposed framework is shown in Figure 1 in the form of a modelling cycle, which shows the individual steps of the corresponding iterative model development procedure. These steps are explained in more detail in the following.

2.1 Model (re)formulation

A basic assumption of the proposed framework is that an initial ODE model, derived from first engineering principles, is available, which needs to be improved to serve its intended purpose. The first step of the modelling cycle then deals with model (re)formulation, which essentially means translation of the ODE model into a stochastic grey-box model (or modification of this model in subsequent modelling cycle iterations).

Stochastic grey-box models are state space models consisting of a set of stochastic differential equations (SDE's) describing the dynamics of the system in continuous time and a set of discrete time measurement equations, i.e.:

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t \quad (1)$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \quad (2)$$

where $t \in \mathbb{R}$ is time, $\boldsymbol{x}_t \in \mathbb{R}^n$ 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 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 $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S}(\boldsymbol{u}_k, t_k, \boldsymbol{\theta}))$.

A considerable advantage of models of this type is that they are designed to accommodate random effects due to e.g. approximation errors or unmodelled phenomena through the diffusion term of the SDE's in (1), which means that estimation of the parameters of this term from experimental data provides a measure of model uncertainty. This is a key point and forms the basis of the proposed framework for systematic model improvement.

2.2 Parameter estimation

In the second step of the modelling cycle the idea therefore is to estimate the unknown parameters of the model in (1)-(2) from experimental data, including the parameters of the diffusion term.

Stochastic grey-box models allow for a decomposition of the noise affecting the system into a process noise term (the diffusion term) and a measurement noise term. As a result unknown parameters of such models can be estimated from experimental data in a *prediction error* (PE) setting, whereas for standard ODE models it can only be done in an *output error* (OE) setting, which tends to give biased and less reproducable results, because random effects are absorbed into the parameter estimates (Young, 1981). Furthermore, since the solution to (1) is a Markov process, an estimation scheme based on probabilistic methods can be applied, e.g. *maximum likelihood* (ML) or *maximum a posteriori* (MAP). An efficient such scheme, based on the extended Kalman filter (EKF), is available (Kristensen *et al.*, 2002*b*).

2.3 Residual analysis

In the third step of the modelling cycle the idea is to evaluate the quality of the model once the unknown parameters have been estimated. The most important aspect in this regard is to investigate the predictive capabilities of the model by performing cross-validation residual analysis, and various methods are available for this purpose.

2.4 Model falsification or unfalsification

The fourth step of the modelling cycle is the important step of *model falsification or unfalsification*, which deals with whether or not, based on the information obtained in the previous step, the model is sufficiently accurate to serve its intended purpose. In practice, this is a subjective decision, as it involves addressing the trade-off between necessary model accuracy and affordable model complexity with respect to the specific intended purpose of the model. If, based on this decision, 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. In the latter case, the properties of the model in (1)-(2) facilitate the task at hand as shown in the following.

2.5 Pinpointing of model deficiencies

In the fifth step of the modelling cycle, which is only needed if the model has been falsified, 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, and as it turns out, the properties of the ML and MAP estimators mentioned above allow t-tests to be applied for this purpose.

These tests 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 (1)-(2), 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 is 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 (1), 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, which in turn provides an error indicator for the constitutive equations or phenomena models influencing this term. If, by using physical insights, it is subsequently possible to select a specific phenomena model for further investigation, the proposed framework also provides means to confirm if the suspicion that this model is incorrect is 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.:

$$r_t = \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}) \tag{3}$$

where r_t is a phenomenon of interest and $\varphi(\cdot) \in \mathbb{R}$ is the nonlinear function used to describe it. To confirm if the suspicion that $\varphi(\cdot)$ is incorrect is true, the parameter estimation step must be repeated with a re-formulated version of the model in (1)-(2), where r_t is isolated by including it as an additional state variable, i.e.:

$$d\boldsymbol{x}_t^* = \boldsymbol{f}^*(\boldsymbol{x}_t^*, \boldsymbol{u}_t, t, \boldsymbol{\theta})dt + \boldsymbol{\sigma}^*(\boldsymbol{u}_t, t, \boldsymbol{\theta})d\boldsymbol{\omega}_t^* \tag{4}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k^*, \boldsymbol{u}_k, t_k, \boldsymbol{\theta}) + \boldsymbol{e}_k \tag{5}$$

where $\boldsymbol{x}_t^* = [\boldsymbol{x}_t^T \; r_t]^T$ is the extended state vector, $\boldsymbol{\sigma}^*(\cdot) \in \mathbb{R}^{(n+1)\times(n+1)}$ is the extended diffusion term and $\{\boldsymbol{\omega}_t^*\}$ is an $(n+1)$-dimensional standard Wiener process. The extended drift term can be derived from the original drift term as follows:

$$\boldsymbol{f}^*(\boldsymbol{x}_t^*, \boldsymbol{u}_t, t, \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}_t, \boldsymbol{u}_t, t, \boldsymbol{\theta}) \\ \dfrac{\partial \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta})}{\partial \boldsymbol{x}_t} \dfrac{d\boldsymbol{x}_t}{dt} + \dfrac{\partial \varphi(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta})}{\partial \boldsymbol{u}_t} \dfrac{d\boldsymbol{u}_t}{dt} \end{pmatrix} \tag{6}$$

The presence of significant parameters in the corresponding diagonal element of the extended diffusion term is then an indication that $\varphi(\cdot)$ is incorrect and in turn confirms the suspicion.

2.6 Estimation of unknown functional relations

In the sixth step of the 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 corresponding procedure is based on a combination of the applicability of stochastic grey-box models for state estimation and the ability of nonparametric regression methods to provide visualizable estimates of unknown functional relations with associated confidence intervals.

Using the re-formulated model in (4)-(5) 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 incorrectly 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 *et al.*, 2001), but in the context of the proposed framework, *additive models* (Hastie and 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.

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.:

$$r_t = \varphi_{\text{true}}(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{\theta}) \tag{7}$$

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

A more elaborate discussion of the proposed methodology is given by Kristensen *et al.* (2002*a*).

3. CASE STUDY: MODELLING A FED-BATCH BIOREACTOR

To illustrate the performance of the proposed methodology in terms of improving the quality of a 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:

$$\frac{dX}{dt} = \mu(S)X - \frac{FX}{V} \tag{8}$$

$$\frac{dS}{dt} = -\frac{\mu(S)X}{Y} + \frac{F(S_F - S)}{V} \tag{9}$$

$$\frac{dV}{dt} = F \tag{10}$$

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

$$\mu(S) = \mu_{\max} \frac{S}{K_2 S^2 + S + K_1} \tag{11}$$

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. For the present case it is assumed that all state variables can be measured and noise levels corresponding to variances of 0.01, 0.001 and 0.01 (absolute values) are used.

3.1 First modelling cycle iteration

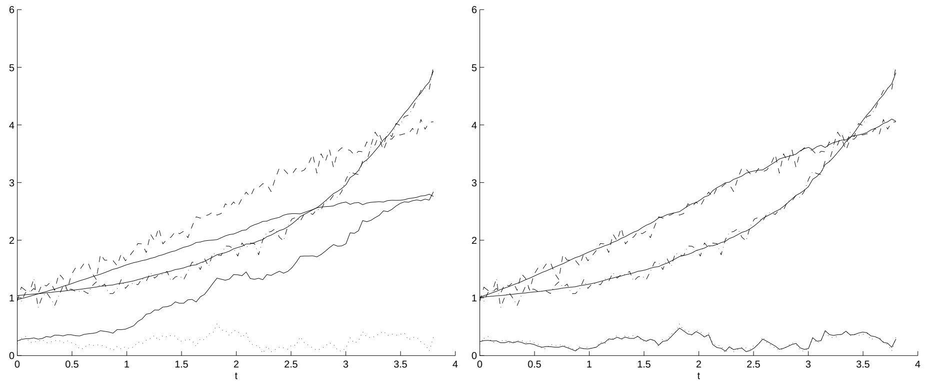
It is assumed that an initial model corresponding to (8)-(10) has been set up, where the true structure of $\mu(S)$ is unknown. As the first step, this model is then translated into a stochastic grey-box model, which has the following system equation:

$$d\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu X - \frac{FX}{V} \\ -\frac{\mu X}{Y} + \frac{F(S_F - S)}{V} \\ F \end{pmatrix} dt + \boldsymbol{\sigma} d\boldsymbol{\omega}_t \tag{12}$$

where $\boldsymbol{\sigma}$ is a diagonal matrix with elements σ_{11}, σ_{22} and σ_{33}. Since the true structure of $\mu(S)$ is unknown, a constant growth rate μ has been assumed, and a diagonal parameterization of the diffusion term has been used to allow possible model deficiencies to be pinpointed. The model also has the following measurement equation:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}_k = \begin{pmatrix} X \\ S \\ V \end{pmatrix}_k + \boldsymbol{e}_k \tag{13}$$

with $\boldsymbol{e}_k \in N(\boldsymbol{0}, \boldsymbol{S})$, where $\boldsymbol{S}$ is a diagonal matrix with elements S_{11}, S_{22} and S_{33}. As the next step, the unknown parameters of the model are estimated using the data from batch no. 1, which gives the results shown in Table 1, and, to evaluate the quality of the resulting model, a pure simulation comparison is performed as shown in Figure 2a. The results of this show that the model does a very poor job, and it is therefore falsified, which means that the modelling cycle must be repeated by re-formulating the model.



(a) Model structure in (12)-(13). (b) Model structure in (15) and (13).

Fig. 2. Pure simulation comparison using cross-validation data from batch no. 2. Dashed lines: y_1, dotted lines: y_2, dash-dotted lines: y_3, solid lines: pure simulation values.

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. This indicates that the first two elements of the drift term may be incorrect. These both depend on μ, which is therefore an obvious deficiency suspect.

To avoid jumping to conclusions, the suspicion should be confirmed, which is done by re-formulating the model with μ as an additional state variable, which yields the system equation:

$$d\begin{pmatrix} X \\ S \\ V \\ \mu \end{pmatrix} = \begin{pmatrix} \mu X - \dfrac{FX}{V} \\ -\dfrac{\mu X}{Y} + \dfrac{F(S_F - S)}{V} \\ F \\ 0 \end{pmatrix} dt + \boldsymbol{\sigma}^* d\boldsymbol{\omega}_t \quad (14)$$

where $\boldsymbol{\sigma}^*$ is a diagonal matrix with elements σ_{11}, σ_{22}, σ_{33} and σ_{44}, and, since μ has been assumed constant, the last element of the drift term is zero. The measurement equation is the same as in (13). 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.

As the next step the re-formulated model in (14) and (13) and the parameter estimates in Table 2 are used to obtain state estimates $\hat{X}_{k|k}$, $\hat{S}_{k|k}$, $\hat{V}_{k|k}$, $\hat{\mu}_{k|k}$, $k = 0, \ldots, N$, by means of the EKF, and an additive model is then fitted to reveal the true structure of the function describing μ by means of estimates of possible functional relations between μ and the state and input variables.

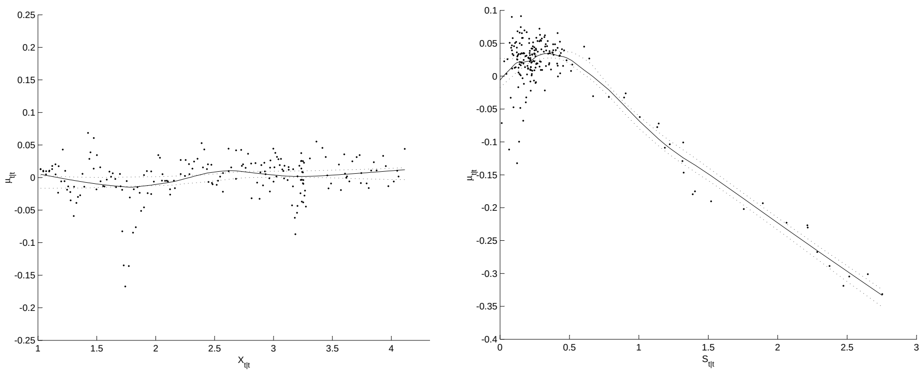
It is reasonable to assume that μ does not depend on V and F, so only functional relations between $\hat{\mu}_{k|k}$ and $\hat{X}_{k|k}$ and $\hat{S}_{k|k}$ are estimated, giving the results shown in Figure 3. These plots indicate that $\hat{\mu}_{k|k}$ does not depend on $\hat{X}_{k|k}$, but is highly 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. More specifically, this function should comply with the functional relation revealed in Figure 3b.

Table 1. Estimation results - (12)-(13).

Parameter	Estimate	Significant?
X_0	9.6973E-01	Yes
S_0	2.5155E-01	Yes
V_0	1.0384E+00	Yes
μ	6.8548E-01	Yes
σ_{11}	1.8411E-01	Yes
σ_{22}	2.2206E-01	Yes
σ_{33}	2.7979E-02	*No*
S_{11}	6.7468E-03	Yes
S_{22}	3.9131E-04	*No*
S_{33}	1.0884E-02	Yes

Table 2. Estimation results - (14)&(13).

Parameter	Estimate	Significant?
X_0	1.0239E+00	Yes
S_0	2.3282E-01	Yes
V_0	1.0099E+00	Yes
μ_0	7.8658E-01	Yes
σ_{11}	2.0791E-18	*No*
σ_{22}	1.1811E-30	*No*
σ_{33}	3.1429E-04	*No*
σ_{44}	1.2276E-01	Yes
S_{11}	7.5085E-03	Yes
S_{22}	1.1743E-03	Yes
S_{33}	1.1317E-02	Yes



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

Fig. 3. 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 computed from 1000 replicates.

3.2 Second modelling cycle iteration

The functional relation revealed in Figure 3b clearly indicates that the growth of biomass is governed by Monod kinetics and inhibited by substrate, which makes it possible to re-formulate the model in (12)-(13) to yield the system equation

$$d\begin{pmatrix} X \\ S \\ V \end{pmatrix} = \begin{pmatrix} \mu(S)X - \frac{FX}{V} \\ -\frac{\mu(S)X}{Y} + \frac{F(S_F - S)}{V} \\ F \end{pmatrix} dt + \boldsymbol{\sigma} d\boldsymbol{\omega}_t \quad (15)$$

where $\boldsymbol{\sigma}$ is again a diagonal matrix with elements σ_{11}, σ_{22} and σ_{33}, and where $\mu(S)$ is given by (11). The measurement equation remains unchanged and is thus the same as in (13). Estimation of the unknown parameters of this model using the same data set as before gives the results shown in Table 3, and to evaluate the quality of the resulting model, a pure simulation comparison is performed as shown in Figure 2b. The results of this show that the model does a much better job now. It is in fact unfalsified with respect to the available information, and the model development procedure can therefore be terminated.

Table 3. Estimation results - (15)&(13).

Parameter	Estimate	Significant?
X_0	1.0148E+00	Yes
S_0	2.4127E-01	Yes
V_0	1.0072E+00	Yes
$\mu_{\max}$	1.0305E+00	Yes
K_1	3.7929E-02	Yes
K_2	5.4211E-01	Yes
σ_{11}	2.3250E-10	*No*
σ_{22}	1.4486E-07	*No*
σ_{33}	3.2842E-12	*No*
S_{11}	7.4828E-03	Yes
S_{22}	1.0433E-03	Yes
S_{33}	1.1359E-02	Yes

4. CONCLUSION

A systematic framework for improving the quality of first engineering principles models has been presented. The proposed framework is based on stochastic grey-box modelling and incorporates statistical tests and nonparametric regression, which in turn facilitates pinpointing of model deficiencies and subsequent uncovering of their structural origin. A key result is that the proposed framework can be used to obtain estimates of unknown functional relations, which allows unknown or incorrectly modelled phenomena to be uncovered and proper parametric expressions for the associated constitutive equations to be inferred.

REFERENCES

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall. London, England.

Hastie, T. J., R. J. Tibshirani and J. Friedman (2001). *The Elements of Statistical Learning - Data Mining, Inference and Prediction*. Springer-Verlag. New York, USA.

Kristensen, N. R., H. Madsen and S. B. Jørgensen (2002*a*). A method for systematic improvement of stochastic grey-box models. Submitted for publication.

Kristensen, N. R., H. Madsen and S. B. Jørgensen (2002*b*). Parameter estimation in stochastic grey-box models. Submitted for publication.

Raisch, J. (2000). Complex systems - simple models?. In: *Proceedings of the IFAC Symposium on Advanced Control of Chemical Processes* (L. T. Biegler, A. Brambilla, C. Scali and G. Marchetti, Eds.). Elsevier. pp. 275–286.

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