Expert Systems with Applications 38 (2011) 10862–10874

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A novel identification method for hybrid (N)PLS dynamical systems with application to bioprocesses

M. von Stosch[a], R. Oliveira[b], J. Peres[a], S. Feyo de Azevedo[a,*]

[a] *LEPAE, Departamento de Engenharia Quimica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal*

[b] *REQUIMTE, Departamento de Quimica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*

ARTICLE INFO

Keywords:
Hybrid modeling
Dynamic modeling
Nonlinear partial least square (nonlinear projection to latent structures)
Artificial neural networks
Bordetella pertussis

ABSTRACT

This paper presents a method for the identification of nonlinear partial least square (NPLS) models embedded in macroscopic material balance equations with application to bioprocess modeling. The proposed model belongs to the class of hybrid models and consists of a NPLS submodel, which mimics the cellular system, coupled to a set of material balance equations defining the reactor dynamics. The method presented is an analog to the non-iterative partial least square (NIPALS) algorithm where the PLS inner model is trained using the sensitivity method. This strategy avoids the estimation of the target fluxes from measurements of metabolite concentrations, which is rather unrealistic in the case of sparse and noisy off-line measurements.

The method is evaluated with a simulation case study on the fed-batch production of a recombinant protein, and an experimental case study of *Bordetella pertussis* batch cultivations. The results show that the proposed method leads to more consistent models with higher statistical confidence, better calibration properties and reinforced prediction power when compared to other dynamic (N)PLS structures.

1. Introduction

Partial least square (PLS) (also called projection to latent structures) and nonlinear PLS (NPLS) have been shown to be powerful regression methods for static processes when the data is noisy and highly correlated. There are numerous applications of PLS and NPLS in biotechnology (Clementschitsch & Bayer, 2006; Henneke, Hagedorn, Budman, & Legge, 2005; Soons, Streefland, van Straten, & van Boxtel, 2008). The difference between PLS and NPLS lies in the inner models which correlate the latent variables. In PLS the inner model is based on linear regression, whereas in most NPLS the inner model is nonlinear, mimicked by quadratic functions (Wold, Kettaneh-Wold, & Skagerberg, 1989), artificial neural networks (Qin & McAvoy, 1992), radial basis functions (Baffi, Martin, & Morris, 2000) or support vector machines (Wang & Yu, 2004).

Many biotechnological processes are inherently dynamic and the PLS structure cannot be directly applied. Several attempts in the literature were made in order to extend the static PLS models for dynamical systems (Baffi et al., 2000; Lakshminarayanan, Shah, & Nandakumar, 1997; Ljung, 1991; Qin, 1993; Ricker, 1988). In most cases modeling of dynamic systems has been achieved through the augmentation of the inputs with lagged values of input and output data (Baffi et al., 2000; Ljung, 1991; Qin, 1993; Ricker, 1988). One-step-ahead prediction was developed inspired on the series–parallel identification scheme (Eykhoff, 1974) and recurrent training schemes (Qin & McAvoy, 1992; Werbos, 1988) or parallel identification schemes were used (Qin & McAvoy, 1996) for long term predictions. In the paper by Baffi et al. (2000) NPLS with different inner nonlinear models is successfully applied for modeling of nonlinear dynamical systems.

A bioprocess is ruled by a large number of complex physical, chemical and biological constraints, which are associated with both the cellular system and the bioreactor system. The above mentioned PLS models completely disregard such constraints since they are empirical data based techniques.

The dynamic nature of a bioprocess can be established by macroscopic material balances of the compounds with capacity to influence the physiological state of a cell. Thus an alternative way to add dynamics to a PLS model is to combine a static (N)PLS submodel with material balance equations in a hybrid dynamical structure. This type of strategy has been extensively reported in the literature for artificial neural networks (Lee, Vanrolleghem, & Park, 2005; Oliveira, 2004; Peres, Oliveira, & Feyo de Azevedo, 2001; Preusting & Noordover, 1996; Schubert, Simutis, Dors, Havlik, & Luebbert, 1994a; Schubert, Simutis, Dors, Havlfk, & Luebbert, 1994b; Simutis, Oliveira, Manikowski, de Azevedo, & Luebbert, 1997) but very rarely for (N)PLS (Henneke et al., 2005; Lee et al., 2005).

[*] Corresponding author. Tel.: +351 22 508 1694; fax: +351 22 508 1449.
E-mail address: sfeyo@fe.up.pt (S. Feyo de Azevedo).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.02.117

In this paper, a generic nonlinear dynamic PLS approach is developed within the hybrid modeling framework, i.e. by combining a (N)PLS submodel with material balance equations. There are two possible strategies to develop such a model. The probably simplest way is to estimate the reaction rates from the material balance equations and from the concentrations' measurements and then to run a static NPLS model with the rates as target outputs and the state space vector as the inputs (Henneke et al., 2005; Lee et al., 2005). The difficulty of using this method arises when dealing with a limited number of observations and noisy measurements. The conjugation of these two factors is frequent in a real application, leading to very inaccurate estimation of the reaction rates. The second alternative, which is explored in this paper, follows the simultaneous parameter estimation strategy, using the well known sensitivity method (Oliveira, 2004; Peres et al., 2001; Preusting & Noordover, 1996; Schubert et al., 1994a; Schubert et al., 1994b; Simutis et al., 1997).

The paper is organized as follows: in Section 2 the proposed semi-parametric hybrid model, the parameter identification algorithm and model performance criteria are described; Section 3 presents the application, results and discussion of the proposed method for two complementary case studies – one case with simulation data, specifically a model on protein synthesis, also known as the Park Ramirez model (Park & Ramirez, 1988), and another case comprising sparse, infrequent experimental data of *Bordetella pertussiss* cultures; then, in Section 4, the conclusions are drawn.

2. The semi-parametric hybrid model

The semi-parametric hybrid structure here developed can also be referred to as an intrinsically dynamic NPLS model, which consists of two parts, namely material balances and a nonparametric/parametric submodel. The general hybrid model structure is described in the first subsection. The integration of the nonparametric model, a nonlinear partial least square model, is explained in the second subsection and a novel parameter identification algorithm is presented in the third subsection. The question of choosing the best model structure is finally addressed.

2.1. The general semi-parametric hybrid model structure

The general hybrid model structure proposed is depicted in Fig. 1. The concept is an evolution of the semi-parametric hybrid model proposed originally by Oliveira (2004). The structure is based on a bioreactor dynamic model, consisting of n material balances represented in vectorial terms as:

$$\frac{dc}{dt} = f = r(L_x, w_A) - D \cdot c + u, \qquad (1)$$

where c is the vector of concentrations, D is the dilution rate, u is a vector of volumetric control inputs and r is the vector of kinetic rates, i.e. the reaction term mimicking the cell system, which is modeled with a nonparametric/parametric submodel, using the vector of parameters w_A.

The nonparametric/parametric submodel reads as:

$$r(c, L_x, w_A) = K \cdot \left\langle \phi_j(c) \times \rho_j(L_x, w_A) \right\rangle_{j=1,\dots,m}, \qquad (2)$$

with K being a $n \times m$ matrix of yield coefficients, ϕbeing m kinetic functions and $\rho(L_x,w_A)$ being unknown kinetic functions which include w_A and the inputs L_x. These unknown kinetic functions are modeled with nonparametric techniques, such as artificial neural networks (Oliveira, 2004; Peres et al., 2001; Preusting & Noordover, 1996; Psichogios & Ungar, 1992; Schubert et al., 1994a; Schubert et al., 1994b; Simutis et al., 1997; Thompson & Kramer, 1994) or, as presented in the following, by a Nonlinear Partial Least Square alike model.

The hybrid model can either be classified as a one-step or a multi-step ahead predictor. This is due to the unknown kinetic functions, $\rho(L_x,w_A)$, in Eq. (2), more precisely the inputs, L_x. When, the inputs cover only measured inputs at discrete time points, equivalent to a finite impulse response (FIR) model, then the hybrid model functions behave as a one-step ahead predictor. When, alternatively, L_x comprises only the estimates of the model at discrete time points, equivalent to an AutoRegression (AR) model, then the hybrid model is a multi-step ahead predictor. It should be pointed out that the combination of measured and estimated

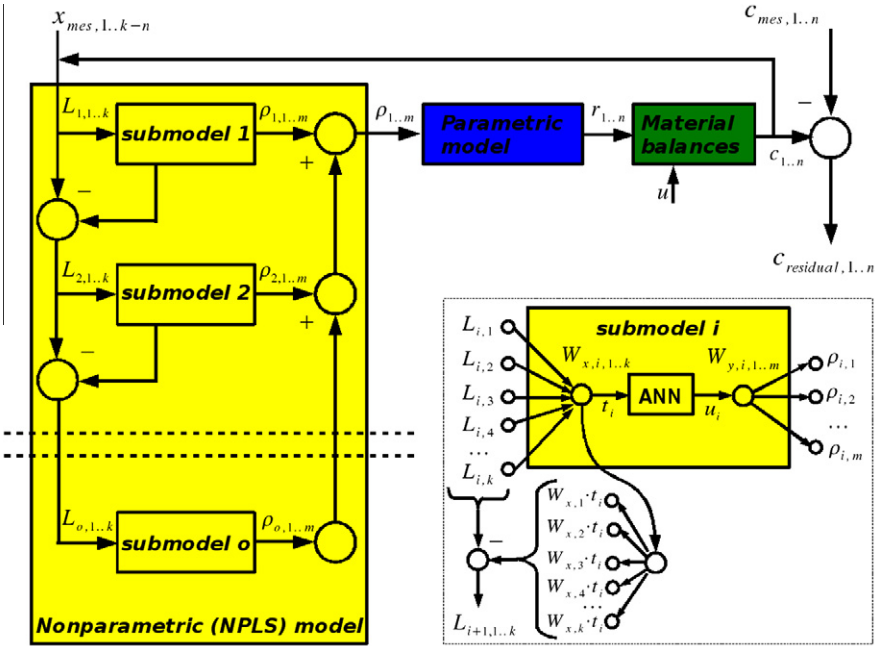

Fig. 1. Diagram of the general semi-parametric hybrid model structure and of the incorporated submodels (mathematical symbols as in the text).

data for L_x, equivalent to an AutoRegression eXogenous (ARX) model, results in a one-step ahead predictor.

2.2. The nonparametric model

The proposed nonparametric submodel, hereafter referred to as nonparametric model, is the key feature of the novel hybrid model. The structure is the one of a NPLS model, which is embedded into the hybrid framework, as reported to have been successfully applied in many areas, (Baffi et al., 2000; Henneke et al., 2005; Lee et al., 2005; Qin & McAvoy, 1996). In fact the structure exhibits all (N)PLS features, such as maximization of the covariance between input and output variables, minimization of redundant information of the inputs and identification of a minimal number of latent variable models. In the method here proposed the estimation of the unknown kinetic rates from noisy and sparse concentration measurement data is circumvented.

2.2.1. The nonparametric model structure

The nonparametric model, for each component, j, of the vector of unknown kinetic functions, $\rho_j(L_x, w_A)$, is composed of o separate latent variable models (referred to as submodels $i = 1,\ldots,o$, see Fig. 1), such that:

$$\rho_j(L_x, w_A) = \sum_{i=1}^{o} \rho_{i,j}(L_x, w_A), j = 1, \ldots, m, \tag{3}$$

where the index i denotes latent variable i. Note that in the following the term “latent variable model” is relaxed to latent variable.

Each submodel can further be divided into two parts, an outer and an inner model (Fig. 1): the outer model firstly linearly compresses the respective high dimensional input, by the use of input loadings, to one inner latent variable; the inner model then correlates, (non) linearly, the input latent variable, t_i, to the output latent variable, u_i; and subsequently the outer model decompresses the outer latent variable, u_i, through the use of the output loadings, into the respective outer vector $\rho_{i,\,j}(L_x, w_A)$ (for details see Baffi et al., 2000; Qin & McAvoy, 1992). In Baffi et al. (2000), ANN and RBF were used as inner models, which proved to be successful. In this approach an ANN model is applied. Mathematically this nonparametric model is expressed as follows:

$$\rho_{1,\ldots,m,i}(L_x, w_A) = W_{y,i} \cdot \left(w_{2,i} \cdot g\left(w_{1,i} \cdot h\left(W_{x,i} \cdot L_{i,1,\ldots,k}\right) + b_{1,i}\right) + b_{2,i}\right), \tag{4}$$

where $W_{x,i}$ and $W_{y,i}$ are the compression factors of the outer model, also called loadings, $w_{2,i}$ and $w_{1,i}$ are parameters of the ANN inner model, $b_{2,i}$ and $b_{1,i}$ are the biases of the ANN inner model, $h(\cdot)$ and $g(\cdot)$ are transfer functions, here linear and tangential, and $L_{i,1\ldots k}$ comprises all inputs 1 to k to the model.

For $i = 1$ the vector of inputs comprises the estimated state variables or/and additional measured data $x_{\text{mes},1\ldots n}$, as illustrated in Fig. 1.

For $i > 1$, the vector of inputs, $L_{i,1\ldots k}$ is the difference between the previous input vector and the information captured by the previous input latent variable, i.e. mathematically:

$$L_{i,1\ldots k} = L_{i-1,1\ldots k} - W_{x,i-1} \cdot L_{i-1,1\ldots k} \cdot W_{x,i-1}^T, \tag{5}$$

The arising advantage when compared to the so far used nonparametric model is that high numbers of redundant experimental data can be considered as inputs to the nonparametric model. In contrast to (N)PLS models the advantage for the identification of the involved parameters is that the kinetic rates do not need to be known explicitly, and that the hybrid structure is inherently dynamic. It should however be stressed that while the structure is a relevant prerequisite, the parameter identification method is essential for the success of the overall procedure.

2.2.2. Identification of the nonparametric model parameters

The identification of the nonparametric model parameters proposed in this paper differs from the NIPALS identification procedure, but the general idea of this algorithm is kept. This idea is somehow identical to a twofold objective optimization, where both the covariance between inputs and outputs and the captured variance of the input are maximized. The maximizations are accomplished by the application of the sensitivity approach (Frank, 1978; Oliveira, 2004; Peres et al., 2001; Simutis et al., 1997), as it was shown to be preferable over the error-prone initial estimation of the kinetic rates with sequent parameter identification for ANNs (e.g. see Oliveira, 2004).

2.2.2.1. Maximization of the covariance between inputs and outputs. The maximization of the covariance between the inputs and outputs is analogous to the minimization of a weighted least-square error function of the state variables, c, which reads as:

$$\min_{w_A}\left\{E_1 = \frac{1}{P \times n}\sum^{P}\sum_{j=1}^{n}\frac{\left(c_{\text{mes},j}(t) - c_j(t, w_A)\right)^2}{c_{\sigma,j}}\right\}, \tag{6}$$

and where w_A are the model parameters, $c_{\text{mes},1\ldots n}$ is the vector of measured-known state variables, and $c_{\sigma,j}$ is the standard variance of the experimentally measured concentration.

This objective function requires the determination of the number of latent variables prior to application, which is in contrast to (N)PLS models where consecutive latent variables are added till the desired level of abstraction is reached.

2.2.2.2. Maximization of the captured input variance. The first objective function E_1 serves only to maximize the covariance between the inputs and outputs, while the NIPALS algorithm also provides orthogonality of the latent variables that span the subspace (Baffi et al., 2000). This feature is important, because parameter identification problems arising from redundant input information are prevented and the dimension of the solution space is reduced. As for (N)PLS structures, redundant information is minimized on one hand by the compression of the input dimensions and on the other hand by subtraction of the information covered by the respective latent variable from the input information, Eq. (5), i.e. capturing the variance of the inputs. In analogy to this intrinsic feature of the NIPALS algorithm, the objective defined in the following seeks to account for such.

Capturing the variance of the inputs is analogous to the minimization of the residual of the inputs, i.e. minimizing:

$$L_{res,1\ldots k} = L_{0,1\ldots k} - \sum_{i}^{o} W_{x,i-1} \cdot L_{i-1,1\ldots k} \cdot W_{x,i-1}. \tag{7}$$

A direct application of this equation for optimization is not feasible as uncorrelated inputs hinder the convergence of the optimization. In order to circumvent this problem the following procedure was developed:

(i) The first step therein is to regress the matrix of inputs $L_{i,1\ldots k}$ with the input scores, t_i, in order to obtain the input loadings in a PCA manner, i.e.:

$$W_{x,i,lin,un} = \frac{L_{i,1\ldots k} \cdot t_i}{t_i^T \cdot t_i}. \tag{8}$$

The obtained solution is then normalized to unit length:

$$W_{x,i,lin} = \frac{W_{x,i,lin,un}}{\|W_{x,i,lin,un}\|}. \tag{9}$$

(ii) The second step is the calculation of the residual between the input loading which is incorporated in the system of model equations, $W_{x,i}$, and the one obtained from Eqs. (8) and (9), $W_{x,i,\mathrm{lin}}$. The minimization of this residual is thought to be similar to the minimization of Eq. (7). For the minimization a least square error function is adopted, i.e.:

$$\min_{w_A}\left\{E_2=\frac{1}{o\times k}\sum_{i=1}^{o}\left(W_{x,i,\mathrm{lin}}-W_{x,i}\right)^2\right\}. \tag{10}$$

In such a way the inputs that are not correlated to other inputs or to the outputs are taken into account.

For the maximization of the error functions, E_1 and E_2, the sensitivity equations are employed. This means that the objective functions are differentiated with respect to the parameters w_A.

2.2.2.3. Sensitivity equations for E_1. The sensitivity equations are obtained by differentiating Eq. (6) with respect to w_A, which in general implies the derivation of Eq. (1) with respect to w_A.

For the inner models, i.e. the ANN's, this reduces to the derivation of Eq. (1) with respect to $w_{1,i}$ and $w_{2,i}$ (embodied in the following by w) resulting in:

$$\frac{d}{dt}\cdot\frac{dc}{dw}=\frac{\partial f}{\partial c}\cdot\frac{dc}{dw}+\frac{\partial f}{\partial w}, \tag{11}$$

where the first term on the right hand side of Eq. (11) is due to the optional consideration of estimated state variables as inputs to the nonparametric model, as displayed in Fig. 1.

For the outer models the sensitivity equations are similarly obtained by differentiating Eq. (6) with respect to the input and output loadings $W_{x,i}$ and $W_{y,i}$ (which are in the following embodied by $W_{x|y,i}$). Not yet mentioned, but essential to report in this context is the normalization of the loadings $W_{x|y,i}$. This normalization carried out by analogy with the NIPALS algorithm facilitates mathematical operations since $W^T_{X/y,i}=W^{-1}_{X/y,i}$, where:

$$W_{x/y,i}=\frac{W^{\mathrm{up}}_{x/y,i}}{\left\|W^{\mathrm{up}}_{x/y,i}\right\|}, \tag{12}$$

with $W^{\mathrm{up}}_{x/y,i}$ being the vector of parameters obtained from the optimization procedure.

For the derivation of the sensitivity equation, Eq. (12) is accounted for by the chain rule, i.e. the chain factor resulting from Eq. (12) reads:

$$\frac{\partial W_{x/y,i}}{\partial W^{\mathrm{up}}_{x/y,i}}=\begin{pmatrix}\frac{\left(\left\|W^{\mathrm{up}}_{x/y,i}\right\|-W^{\mathrm{up}}_{x/y,i,1,1}\cdot W^{\mathrm{up}}_{x/y,i,1,1}\right)}{\left\|W^{\mathrm{up}}_{x/y,i}\right\|^2} & \cdots & \frac{\left(-W^{\mathrm{up}}_{x/y,i,1,1}\cdot W^{\mathrm{up}}_{x/y,i,1,p}\right)}{\left\|W^{\mathrm{up}}_{x/y,i}\right\|^2}\\ \vdots & \ddots & \vdots\\ \frac{\left(-W^{\mathrm{up}}_{x/y,i,1,1}\cdot W^{\mathrm{up}}_{x/y,i,q,1}\right)}{\left\|W^{\mathrm{up}}_{x/y,i}\right\|^2} & \cdots & \frac{\left(\left\|W^{\mathrm{up}}_{x/y,i}\right\|-W^{\mathrm{up}}_{x/y,i,q,p}\cdot W^{\mathrm{up}}_{x/y,i,q,p}\right)}{\left\|W^{\mathrm{up}}_{x/y,i}\right\|^2}\end{pmatrix}. \tag{13}$$

The sensitivity equations can then, similarly to Eq. (11), be obtained by differentiating Eq. (1) with respect to $W^{\mathrm{up}}_{x/y,i}$ which for the output loadings results in:

$$\frac{d}{dt}\cdot\frac{dc}{dW^{\mathrm{up}}_{y,i}}=\frac{\partial f}{\partial c}\cdot\frac{dc}{dW^{\mathrm{up}}_{y,i}}+\frac{\partial f}{\partial W_{y,i}}\cdot\frac{dW_{y,i}}{dW^{\mathrm{up}}_{y,i}}, \tag{14}$$

and for the input loadings gives:

$$\frac{d}{dt}\cdot\frac{dc}{dW^{\mathrm{up}}_{x,i}}=\frac{\partial f}{\partial c}\cdot\frac{dc}{dW^{\mathrm{up}}_{x,i}}+\frac{\partial f}{\partial W_{x,i}}\cdot\frac{dW_{x,i}}{dW^{\mathrm{up}}_{x,i}}+\frac{\partial f}{\partial L_{i,1\ldots k}}\cdot\frac{dL_{i,1\ldots k}}{dW_{x,i}}\cdot\frac{dW_{x,i}}{dW^{\mathrm{up}}_{x,i}}. \tag{15}$$

The sensitivity equations of the input loadings, Eq. (15), bare the specialty that the subtraction of the covered information from the input information, namely Eq. (5), must be taken into account, what is accomplished by the third term on the right hand side in Eq. (15) (for details see the Appendix A). The derivation of $\partial f/\partial c$ and $\partial f/\partial w_A$ is straightforward and similar to the nonparametric structure given e.g. in Oliveira (2004) wherefore they are not described in detail.

2.2.2.4. Sensitivity equations for E_2. The sensitivity equations for the second objective function, E_2, are obtained by the differentiation of Eq. (10) with respect to w_A (i.e. the in-output loadings and the ANN parameters). In the case of $W_{x,i}$, the derivative is obtained in a relatively straight forward way, resulting in Eq. (13) for the input loadings, while being zero for the output loadings and ANN parameters. In contrast, deriving the gradients of $W_{x,i,\mathrm{lin}}$ with respect to w_A is operose. The chain rule can be applied using Eq. (13) to account for Eq. (9) and differentiating Eq. (12) with respect to w_A (i.e. the input loadings, output loadings and ANN parameters), as shown in the Appendix A.

The least square problem functions, E_1 and E_2, are optimized simultaneously by using the "lsqnonlin" MATLAB function, which uses a subspace trust region method and is based on the interior-reflective Newton method (MATLAB Optimization toolbox), therefore gradient based, i.e. the sensitivity equations are required. However, when estimates of the state space variables are considered as inputs, then all parameters of the nonparametric model are also used to maximize the captured input variance, which is not desirable.

In order to account for this, first the simultaneous parameter identification is carried out and then, when the best parameter of the respective structure are identified as described below, only w and $W_{y,i}$ are further optimized subject only to the first objective function E_1.

In any case, the sensitivity equations are integrated along with the model equations, namely the system of equations comprised by Eq. (1). In this study an Euler integration scheme is adapted. Initial values of sensitivity equations are zero, because the initial state variables are independent of the parameters.

2.2.2.5. Additional challenges for parameter identification. Parameter identification of nonparametric structures, especially when gradient based, exhibit a few additional challenges, namely restoring of the model generalization capabilities and avoiding local minima. The first challenge is usually overcome by (i) splitting the data set into two partitions: the training set that contains about 2/3 of the data; and the validation set, which comprises about 1/3; and (ii) terminating the parameter optimization when a certain level of sophistication is reached (Bishop, 1995; Haykin, 1998; Oliveira, 2004).

The second challenge, namely local minima, arises from the shape of the solution space spanned by the objective functions and the parameters (Bishop, 1995; Haykin, 1998). The consistency of the minima obtained for various random initiations of the parameters (in this study at least four) is on one hand a measure of the quality of the solution obtained, and on the other hand a measure of the problem formulation quality. Notice that the larger the number of random initializations, the larger is the statistical confidence of the solution (Bishop, 1995; Haykin, 1998).

2.3. Model performance criteria

In order to identify the best hybrid model, both a measure of model performance must be defined, i.e. a model performance criteria and a suitable set of model structure variations must be considered. As outlined above, in Section 2.2.2, the identification of the best hybrid model structure goes along with the identification of the number of latent variables. Besides this variation in the number of latent variables, the architecture of the ANN structure usually involves the variation of the number of layers and the number of nodes in these layers.

In this work, a number of decisions were taken, in order to downsize the degrees of freedom, namely: (i) a selection of three layers (input, hidden and output layer) was decided, which is usually sufficient if nonlinear continuous functions are sought to be modeled (Haykin, 1998); (ii) the number of nodes for the hidden layers of the ANN is fixed to be one; (iii) the number of nodes for the input and output layers for each submodel is fixed as one, as it results from the (N)PLS structure. What remains is then the evaluation of the variation of numbers of submodels, i.e. the variation of the number of latent variables, for each hybrid model set-up.

One criterion for model performance is the residual, also addressed as the goodness of fit of the model estimates regarding the data, which can be assessed through the Mean Square Error, MSE, where MSE is defined as:

$$\mathrm{MSE} = \frac{1}{P \times n} \cdot \sum_{P} \sum_{j=1}^{n} \left(c_{\mathrm{mes},j}(t) - c_j(t, w_A)\right). \tag{16}$$

Evaluation and comparison of model concepts and structures cannot however be only built up on the estimation error obtained for the training, validation or test set, in form of the residual (Bishop, 1995; Haykin, 1998). It is known that as model complexity grows, i.e. the number of parameters grows, the quality of fit may apparently improve, but often at the expense of robustness and generalization capabilities, (Bishop, 1995; Haykin, 1998). With respect to these issues the Akaike Information Criteria, AIC, is a suitable and widely applied criteria, but according to Leonard and Hsu (1999), Burnham and Anderson (2004), Peres, Oliveira, and de Azevedo (2008), the Bayesian information criteria (BIC), is more appropriate for the applications which this approach addresses. Therefore the BIC is applied for the model comparison and selection in this study.

The Bayesian information criteria (BIC) is defined as:

$$\begin{aligned}\mathrm{BIC} = {} & \left(-\frac{n \cdot P}{2} \cdot \ln\left(\sum_{P} \sum_{j=1}^{n} \left[c_{\mathrm{mes},j}(t) - c_j(t, w_A)\right]^2\right)\right) \\ & - \left(\frac{n_w}{2} \cdot \ln\left(\frac{n \cdot P}{2\Pi}\right)\right),\end{aligned} \tag{17}$$

where the term in the first bracket is the logarithmic maximum likelihood and n_w is the total number of parameters/weights. In terms of the BIC, the model to be selected is the one that exhibits the larger BIC value for the validation set (Burnham & Anderson, 2004; Leonard & Hsu, 1999; Peres et al., 2008).

3. Application, results and discussion

In this section the application, results and discussion of the proposed hybrid model and of reference dynamic (N)PLS models are reported for two complementary case studies. The first study focuses on the process dynamics and the identification of the number of latent variables. The second study concentrates on the model identification from typical noisy, sparse and infrequent experimental data, a case which hinders the direct application of the reference dynamic (N)PLS models. The results obtained for the hybrid model are rigorously analyzed and benchmarked against reference dynamic (N)PLS models.

3.1. Case studies

3.1.1. A protein synthesis, the Park Ramirez model

3.1.1.1. The protein synthesis process. The method proposed in Section 2 is evaluated in this subsection with simulation data of protein synthesis in a fed-batch reactor, also known as the Park-Ramirez model, as originally proposed by Park and Ramirez (1988). This model found wide application, for similar model structures to the one proposed here, e.g. in Kulkarni, Chaudhary, Nandi, Tambe, and Kulkarni (2004) for the evaluation of their principal component analysis – general regression neural network model, or in Oliveira (2004) for the evaluation of the traditional semi-parametric hybrid model. The reactor model comprises material balances of the secreted and total protein/product, the biomass, the substrate and the volume. The model dynamics, i.e. the offset between formation of secreted and total protein on the one side and biomass growth and substrate uptake on the other, poses some challenge, which is one reason for the application of this model in this study. Also, this model finds application because the number of latent variables therein is expected to be larger than one, but smaller than four as analytically at least two kinetic rates (substrate uptake and biomass growth) are linearly dependent and such accounts for the model capability of identifying the underlying latent variables.

In this paper the model equations, the feeding profile, the variation of the initial concentrations and the corruption of the generated simulation data with a Gaussian error of 5%, were applied for simulation case data generation, as described in Kulkarni et al. (2004). Normal and abnormal (in the sense of initial data outside the usual range, as defined by Kulkarni et al. (2004)) fed-batch data were generated, through variations in the initial values of concentrations, which significantly influence the concentrations dynamics. Three sets were defined, comprising 12 normal plus 4 abnormal fed-batches for the training data set, 2 plus 2 for the validation set and 2 plus 2 for the test set, respectively. After generation, the sets were corrupted with 5% Gaussian noise, except for the feeding and volume data which were corrupted with 1.5% Gaussian noise.

3.1.1.2. The reference models. As reference for comparison with the proposed dynamic hybrid models, (N)PLS models which account for the dynamics by the augmentation of the inputs in the sense of finite impulse response (FIR) or AutoRegression (AR) are used (as in most cases: Baffi et al., 2000; Ljung, 1991; Qin, 1993; Ricker, 1988). The model structure identification of such dynamic (N)PLS models comprises the identification of inputs to the models, namely the number, type and time-points, in the sense of FIR or AR, and the identification of the number of latent variables, i.e. the structure is adapted in order to obtain the smallest mean square prediction error in the validation set. In the following (see Table 1) they will be referred to as FIR-(N)PLS and AR-(N)PLS, respectively. The NPLS models contain the same ANN inner model functions as the hybrid models, which are described in more detail in Section 2.

3.1.1.3. The hybrid models. In this study four different hybrid models are investigated.

In the hybrid structures (A) and (B) no mechanistic knowledge of the process is considered. The model equations for concentrations of secreted protein, total protein, biomass and substrate, read:

$$\frac{d}{dt}\begin{bmatrix} P_{\mathrm{sec}} \\ P_{\mathrm{tot}} \\ X \\ S \end{bmatrix} = \begin{bmatrix} r_{P_{\mathrm{sec}}}(L_x) \\ r_{P_{\mathrm{tot}}}(L_x) \\ \mu(L_x) \\ r_S(L_x) \end{bmatrix} - D \cdot \begin{bmatrix} P_{\mathrm{sec}} \\ P_{\mathrm{tot}} \\ X \\ (S - S_{(F)}) \end{bmatrix}, \tag{18}$$

Table 1
Values of model performance criteria over model types and structural parameters – simulation case study on the protein synthesis, also called the Park Ramirez simulation case.

Model type	BIC train	BIC valid	BIC test	MSE train	MSE valid	MSE test
FIR-PLS (lv[a] = 4; nt[b] = 1)	−1930	−477	−348	0.0334	0.0812	0.0295
AR-PLS (lv[a] = 4; nt[b] = 1)	−2074	−519	−433	0.0408	0.0893	0.0456
FIR-NPLS (lv[a] = 4; nt[b] = 1)	−1945	−458	−383	0.0343	0.0699	0.0388
AR-NPLS (lv[a] = 4; nt[b] = 1)	−1994	−486	−433	0.0349	0.0691	0.0456
Hybrid structure A	−1248	−219	−337	0.0134	0.0228	0.0610
Hybrid structure B	−1189	−140	−222	0.0119	0.0118	0.0235
Hybrid structure C	−1962	−379	−402	0.0595	0.0869	0.1055
Hybrid structure D	−1083	−135	−93	0.0095	0.0114	0.0080

[a] lv: number of latent variables.
[b] nt: number of time series elements.

respectively. This corresponds to the bioreactor dynamic model structure represented by Eqs. (1) and (2), where the matrices K and ϕ are identity matrices.

The hybrid structures (C) and (D) consider some basic knowledge about the process, and the system of equations is generally represented by

$$\frac{d}{dt}\begin{bmatrix} P_{\text{sec}} \\ P_{\text{tot}} \\ X \\ S \end{bmatrix} = \begin{bmatrix} (P_{\text{tot}} - P_{\text{sec}}) & 0 & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & 0 & X \end{bmatrix} \cdot \begin{bmatrix} r_{P_{\text{sec}}}(L_x) \\ r_{P_{\text{tot}}}(L_x) \\ \mu(L_x) \\ r_S(L_x) \end{bmatrix} - D \cdot \begin{bmatrix} P_{\text{sec}} \\ P_{\text{tot}} \\ X \\ (S - S_{(F)}) \end{bmatrix}. \tag{19}$$

While structures (A) and (C) are one-step-ahead predictor models, structures (B) and (D) are multi-step-ahead predictor models, i.e. while the input vector L_x (see Eq. (2)) contains the measured values of substrate, biomass, total and secreted product concentrations for (A) and (C), it contains only estimated values of substrate, biomass, secreted and total product concentration for (B) and (D).

The only remaining undetermined structural feature is thus the number of latent variables. This was identified, in all cases, by an heuristic search of the number of latent variables that produces the best performance in terms of BIC (Eq. (17)) for the validation data.

These hybrid structures can directly be compared to their dynamic (N)PLS counterpart in terms of one-step or multi-step ahead prediction. By doing so, it is possible to evaluate the different structures regarding their statistical confidence, their calibration properties and the model estimation errors.

In advance it should be pointed out that the one-step ahead predictor hybrid models (A) and (C), are expected to perform worse than the multi-step ahead predictor hybrid models (B) and (D), because: (i) the uncertainty, i.e. noise, in the input data is directly passed to the estimates in the case of one-step ahead predictors; (ii) uncertainty in an estimate is passed to all future estimates due to the numerical integration; and (iii) the one-step ahead predictor hybrid models (A) and (C), in contrast to the multi-step ahead predictor hybrid models (B) and (D), have no feedback of the actual state estimates to the nonparametric model, wherefore the nonparametric model can neither identify nor correct for errors in the actual state estimates.

3.1.2. An experimental case study: B. pertussis

3.1.2.1. The B. pertussis process. The experimental study published by Soons et al. (2008, 2008) is the basis for the second case study of the present paper. The challenge here is to examine a dynamic process where only typically infrequent, sparse experimental data is available. Soons, Streefland et al. (2008) reported runs in batch mode and variations to the process conditions, such as in pH, Temperature and dissolved oxygen. Their measurements of the concentrations of lactate, glutamate and biomass over time for eight batches were reported as PAB0003, PAB0004, PAB0005, PAB0006-1, PAB0006-2, PAB0007, PAB0009-1, and PAB0009-2.

In order to identify and avoid bias from possible measurement errors, two sets of studies were carried out in the present paper:

In Set 1 – batches PAB0003, PAB0005, PAB0006-1, PAB0006-2 and PAB0009-2 were employed for training and batches PAB0007 and PAB0009-1 used for validation.

In Set 2 – batches PAB0003, PAB0005, PAB0006-1, PAB0006-2 and PAB0009-1 were employed for training, and PAB0007 and PAB0009-2 for validation.

It should be pointed out that batch PAB0007 is an "abnormal" batch, where a dissolved oxygen limitation and a lowered pH from 0 to 9 h occurred, whereas batches PAB009-1 or PAB009-2 can be taken as "normal" (Soons, Streefland et al., 2008). By doing so, it is guaranteed that in both sets a "normal" and an "abnormal" batch were used in the validation step. The measured biomass concentration of batch PAB0004, was used as final test data, in order to provide a final assessment of the generalization capabilities of the models.

3.1.2.2. The reference models. The reference models in this case study are, as before described for the other case study, (N)PLS models which account for the dynamics by augmentation of the inputs. Beside the augmentation of the inputs in the sense of FIR, the inputs here are also augmented using the AutoRegressive eXogenous (ARX) approach. As before the model structure identification of such dynamic (N)PLS models comprises the identification of the number of latent variables and of inputs to the models, namely the number, type and time-points, in the sense of FIR or ARX, i.e. the structure is adapted in order to obtain the smallest mean square prediction error in the validation set. In both schema a time lag of 1 h and a maximum number of 4 equidistant lags for each input were investigated. In the context of sparse and infrequent measurements the application of these specifications requires that the measurements are pretreated, i.e. in this study the (N)PLS model inputs at the specific time instances were obtained through a cubic smoothing spline (MATLAB routine: csaps). However, this mandatory procedure must be accounted for when analysing the results, since on one hand the smoothing of the data can be expected to enhance the model performance while on the other hand the data interpolation might diminish the same. The NPLS models contain the same ANN inner model functions as the hybrid models, which are described in more detail in Section 2.

3.1.2.3. The hybrid models. The hybrid model in this case contains mechanistic knowledge about the process, which was reported in Soons et al. (2008). This results in improved convergence of the parameter identification and into less random initiations for the parameters in order to obtain consistent results. The system of model equations reads:

$$\frac{d}{dt}\begin{bmatrix}\mathrm{Lac}\\ \mathrm{Glu}\\ X\end{bmatrix}=\begin{bmatrix}\mathrm{Lac}\cdot X & 0 & 0\\ 0 & \mathrm{Glu}\cdot X & 0\\ 0 & 0 & X\end{bmatrix}\cdot\begin{bmatrix}r_{\mathrm{Lac}}\\ r_{\mathrm{Glu}}\\ \mu\end{bmatrix}-D\cdot\begin{bmatrix}\mathrm{Lac}\\ \mathrm{Glu}\\ X\end{bmatrix},\qquad(20)$$

where Lac, Glu and X are the concentrations of Lactate, glutamate and biomass, respectively and r_{Lac}, r_{Glu} and μ are the respective unknown kinetic functions which are obtained by the nonparametric model.

The input vector L_x of the nonparametric model in this study contains the estimates of all concentrations, pH, temperature and the percentage of dissolved oxygen, as reported to be responsible for the process variations (Soons, Streefland et al., 2008). A gain, as reported in the previous case study, the only remaining undetermined structural feature is the number of latent variables. This was as well identified, in all cases, by an heuristic search of the number of latent variables that produces the best performance in terms of BIC (Eq. (17)).

3.2. Issues of hybrid model development and implementation

The proposed semi-parametric hybrid model might be understood as a dynamic NPLS model wherein the dynamics are modeled by material balances. In the following the dynamics and the performance of the hybrid model are rigorously analyzed.

3.2.1. Performance criteria

3.2.1.1. Statistical confidence – the BIC. In comparison to reference dynamic (N)PLS approaches, such as AR(X)- or FIR-(N)PLS models, it was observed that the hybrid methodologies possess way fewer model parameters, i.e. latent variables. This is a qualitative observation which is reflected in both presented simulation cases by the significantly larger BIC values obtained for the hybrid models when compared to the values obtained for the comparative dynamic (N)PLS models (see Tables 1 or 2). It should be pointed out that the dynamic (N)PLS approaches, namely the AR(X)- and FIR-(N)PLS models, are disadvantaged in terms of BIC, due to: (i) the higher number of latent variables; and (ii) the dynamic structure itself which increases the number of parameters on the input side. From the BIC definition, Eq. (17), the model to prefer is the one with the larger BIC value, i.e. the one which for equal residual and number of data, has fewer parameters, in this way penalizing complex models (Bishop, 1995). In general, models with higher numbers of parameters are thought to be less robust and to exhibit worse generalization capabilities than models that offer similar residua, but with smaller number of parameters. Thus the BIC is a measure of the statistical confidence of the model performance and therefore the proposed hybrid models exhibit a higher statistical confidence than the comparative dynamic (N)PLS approaches.

3.2.1.2. Performance under the MSE criterion. The statistical confidence observed for the hybrid models is in agreement with the performance of such models observed and evaluated in terms of the MSE criteria, as shown in Tables 1 and 2.

It was observed that the proposed hybrid method most times exhibits significantly better and only rarely worse performance than the other evaluated dynamic (N)PLS models.

Cases in which the hybrid method exhibited a worse performance in terms of MSE values than the comparative methods were graphically analyzed. As example, for the Park-Ramirez case study, Figs. 2 and 3, it was observed that the highest deviations are to be found in the substrate concentrations for hybrid structures (A) and (C).

When seeking for an explanation it must be kept in mind that: (i) both hybrid models, (A) and (C), are one-step ahead predictor models, in the sense of FIR; and (ii) the estimations by these hybrid models are sensitive to noise in the feeding rate data, as outlined in sub Section 3.1.1.

In the case of the feeding rates, the hybrid model cannot account for the uncertainty therein, because neither the feeding rate data are inputs to the nonparametric model nor the state estimates are feedback to the nonparametric model. That those uncertainties can partially be accounted for when the state estimates are inputs to the nonparametric model, is demonstrated by the excellent performance of hybrid structures (B) and (D). However for the best performance by hybrid model (D), those uncertainties are still observable in form of the slightly bumpy estimations of biomass and substrate and in form of the bumpy estimations of secreted and total protein towards the end of the abnormal fed-batch, shown in Fig. 3.

For the experimental case study it is observed in Table 2, that the performance, in terms of MSE, for the hybrid models on data Sets 1 and 2, is non-coherent: using as example the hybrid model with 3 latent variables, the MSE values obtained for the training data of data Set 1, are half as big when compared to the MSE values on the training data of data Set 2. In order to identify the reason for this contradiction an additional analysis, reported below, was carried out on the influences which errors in the initial concentration values have on the whole dynamics. However, observations for the MSE values of the test data for both sets, wherein the performance of the hybrid models are found to be significantly better than the ones of dynamic (N)PLS models, show the excellent generalization capabilities of the hybrid models.

Table 2
Values of model performance criteria over model types and structural parameters – experimental case study on *Bordetella pertussis* cultivation data.

Model type	Structure	Set[c]	BIC train	BIC valid	BIC test	MSE train	MSE valid	MSE test
FIR-PLS	[lv[a] = 5, nt[b] = 3]	1	−486	−168	–	0.1486	0.0736	–
ARX-PLS	[lv[a] = 6, nt[b] = 3]	1	−549	−209	−55	0.1567	0.0703	0.3996
FIR-NPLS	[lv[a] = 5, nt[b] = 3]	1	−473	−173	–	0.1318	0.0831	–
ARX-NPLS	[lv[a] = 6, nt[b] = 3]	1	−516	−224	−41	0.1142	0.1028	0.0488
Hybrid-NPLS	[lv[a] = 1]	1	−430	−103	9	0.2884	0.1397	0.0160
Hybrid-NPLS	[lv[a] = 2]	1	−330	−88	10	0.1020	0.0836	0.0106
Hybrid-NPLS	[lv[a] = 3]	1	−317	−96	9	0.0842	0.0910	0. 0094
FIR-PLS	[lv[a] = 5, nt[b] = 3]	2	−478	−163	–	0.1483	0.0568	–
ARX-PLS	[lv[a] = 6, nt[b] = 3]	2	−542	−204	−52	0.1573	0.0540	0.2361
FIR-NPLS	[lv[a] = 2, nt[b] = 3]	2	−402	−115	–	0.1509	0.0737	–
ARX-NPLS	[lv[a] = 2, nt[b] = 4]	2	−410	−134	−14	0.1328	0.0782	0.0511
Hybrid-NPLS	[lv[a] = 1]	2	−416	−79	−3	0.2694	0.0687	0.0878
Hybrid-NPLS	[lv[a] = 2]	2	−357	−79	13	0.1420	0.0609	0.0075
Hybrid-NPLS	[lv[a] = 3]	2	−393	−82	16	0.1882	0.0582	0.0037

[a] lv: number of latent variables.
[b] nt: number of time series elements.
[c] Set: Set 1 or 2 refer to the grouping of batches the respective model has been trained and validated on.

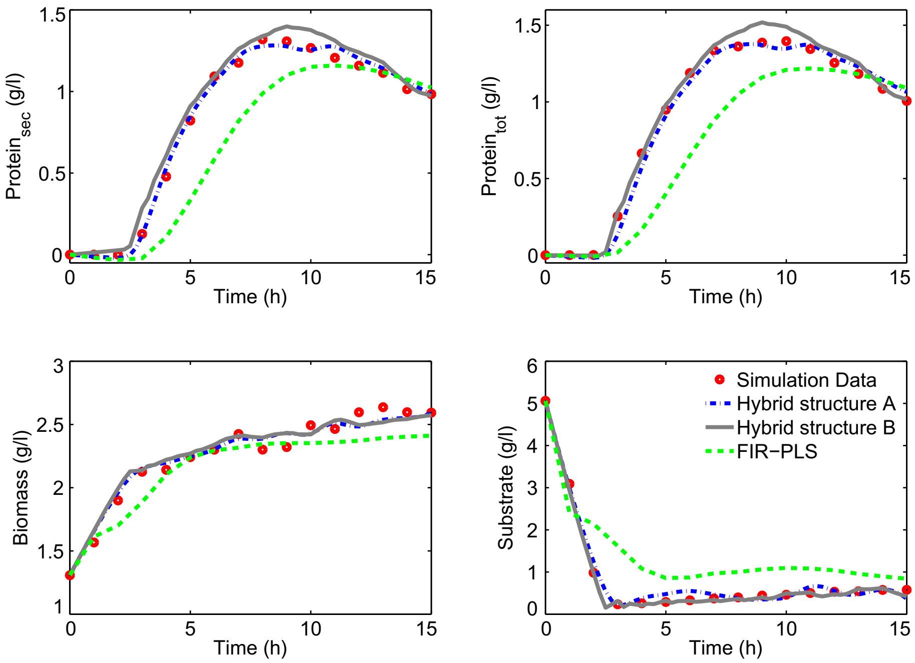

Fig. 2. Park-Ramirez case study – plots of secreted protein, total protein, substrate and biomass concentrations, over time: predictions of hybrid structures A (dashed dotted blue line) and B (grey line), and of the best reference FIR-PLS model (dashed green line, Table 1) vs. the process simulation data (red dots), for a 'normal' validation run. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

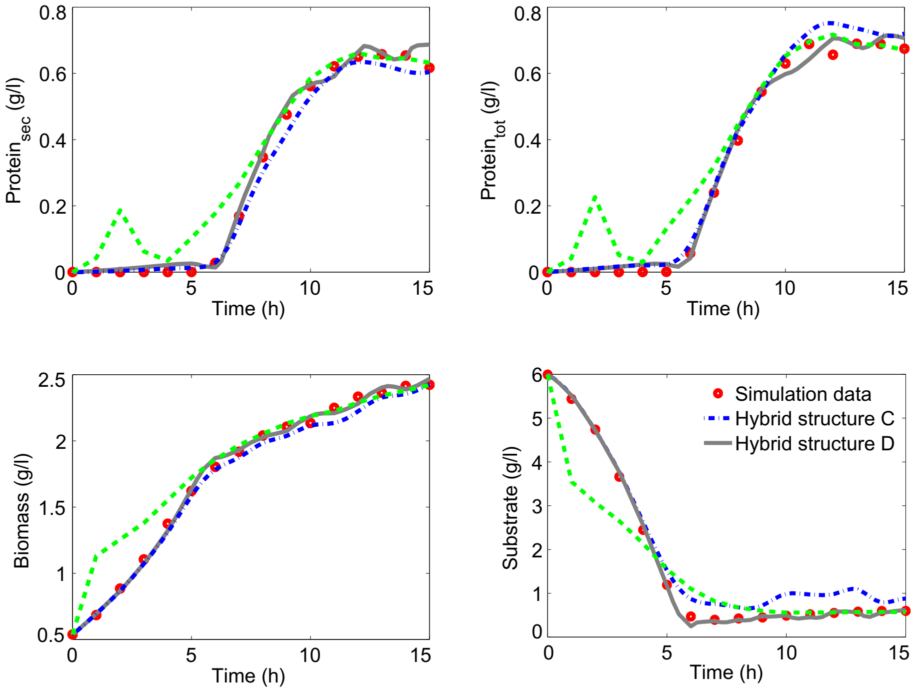

Fig. 3. Park-Ramirez case study – plots of secreted protein, total protein, substrate and biomass concentrations, over time: predictions of hybrid structures C (dashed dotted blue line) and D (grey line), and of the best reference NPLS-AR model (dashed green line, Table 1) vs. the process simulation data (red dots), for an 'abnormal' test run. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2.2. Model structures and error propagation issues

In the case studies presented several sources of errors can be identified, namely: (i) noises in input measurements, (ii) errors inherent to model structures, (iii) errors in estimated inputs and/or in estimated parameters, and (iv) errors associated to 'defective' initial values. These are representative of essentially all experimental applications.

Leaving aside the trivial, though in practice often difficult, issues of error propagation due to the nature of numerical integration methods employed, it is relevant to analyze the issues associated to the nature of model structures chosen.

3.2.2.1. Error propagation due to state feedback to the nonparametric model. One way of propagation of the error in the estimates occurs in all those model structures in which the state estimate is a nonparametric model input, e.g. hybrid structures (B) and (D) in the Park Ramirez case study, or the ARX-(N)PLS model in the experimental case study. However, the form of the time evolution of the sensitivity equations, Eqs. (11), (14), (15), in hybrid structures where state feedback is embedded, tend to have a damping effect on such error propagation. This can be excellently seen by the enhanced performances, in terms of MSE, through hybrid structures (B) and (D) in contrast to the ones for (A) and (C) which are all-together shown in Table 1.

3.2.2.2. Error propagation due to state feedback to the parametric model. Another way in which the error is propagated arises when mechanistic knowledge, namely knowledge about the kinetics, in form of the model estimates, is incorporated, such as in hybrid structures (C) and (D).

The incorporation of the estimates is somewhat identical to the case when the inputs to the nonparametric model comprise the estimates, with the significant difference that an error in the estimation (e.g. from noisy feeding rates as in hybrid structure (C)), depending on the arithmetic operator, (e.g. a multiplication sign for hybrid structures (C)) might amplify the error (e.g. rather large deviations in the substrate concentrations, Fig. 3, and a rather large MSE value, Table 1, are obtained for hybrid structure (C)).

The excellent performance observed with hybrid structure (D), whose mechanistic knowledge is equivalent to (C), is explained by the damping qualities of the nonparametric model.

3.2.2.3. Errors in the Initial values, a special case. A relevant issue in all model analysis is that of the 'condition' of the model structures proposed.

As addressed above, the results of the experimental case study in Table 2, of applying the hybrid models to data Sets 1 and 2, show some inconsistency (see Section 3.1.2). In order to find the reason for such, an additional analysis of the experimental data was carried out, namely a PCA. It was observed that the correlations for the initial values of concentrations in some of the batches vary significantly from the correlations obtained for the whole data set, which is in line with the eye observations made. When (i) correcting the initial values in the validation and test batches of data Set 1 by using PCA and (ii) applying on these sets the hybrid model with two latent variables and the ARX-PLS, which both were prior trained on Set 1, then (iii) the results shown in Table 3 are obtained. Therein it can be seen that the performance in terms of both BIC and MSE values obtained for the hybrid model is significantly better than the performance of the ARX-PLS model. The performance of the hybrid model in Table 3 compared to the very same hybrid model in Table 2 led to more than 50% reduction in the MSE values of the validation and test batches.

This outlines the sensitivity of the proposed hybrid model to a high noise to signal ratio, which is in line to the observations made for the hybrid structures (A) and (C), i.e. the noise in the measurements enters directly the nonparametric model, leading to deviations of the estimations regarding the simulation data. In particular, defective initial values due to noise effects constitutes a special case, as those values are the base for the integration and as such are a significant source of misprediction.

3.3. Challenges of the Park Ramirez case study

The challenges offered by the Park Ramirez simulation case are on the model dynamics and on the identification of the number of latent variables.

3.3.1. The first challenge: the model dynamics

The dynamic delay between formation of secreted and total protein on one hand and biomass growth and substrate uptake on the other hand, varies depending on the initial values of concentrations. This dynamic feature was very well modeled by all applied hybrid structures, apart from the slightly "bumpy" shape of the trajectories, which were ascribed to the error propagation in the discussion above.

Small deviations between estimates and reference values of concentrations can be observed, especially for the one-step ahead predictor hybrid structures (A) and (C), but the general dynamic state behavior is well predicted, as can e.g. be seen in Fig. 3 for the substrate concentration.

Even the dynamics of the abnormal fed-batches are very well predicted by the hybrid structures, in contrast to the observations made for the reference dynamic (N)PLS approaches, as illustrated in Fig. 3. For these special batches it can be concluded that the proposed hybrid models, in comparison to the other dynamic (N)PLS models, even when applied to "regions" where they have been poorly trained on, offer smaller deviations from the simulation data, which confirms the higher statistical confidence of the estimates from such models.

The preceding also means that even if the training set does not contain all possible variations, which can occur during the process, still the performances of the proposed hybrid model for different operating conditions, can be expected to be superior to the one of the comparative dynamic (N)PLS models. These conclusions are according with the findings reported by Thompson and Kramer (1994), Oliveira (2004).

3.3.2. The second challenge: the number of latent variables

The second challenge of the Park Ramirez simulation case is the identification of the number of latent variables for both, the hybrid and the reference (N)PLS models.

Analytically, it is clear that at least two kinetic rates, namely the substrate and biomass rates, are linearly correlated. However, from observations made on the simulation data it might be concluded that also the rates of secreted and total protein are linearly correlated, which in total then sums up to two independent latent variables. This number is observed for the identified hybrid model structures, where it was found that the best hybrid structures always comprised only two latent variables.

In contrast, identification of the best model structure for reference dynamic (N)PLS models always revealed four latent variables. Partially this is due to the fact that linear correlations of the kinetic rates do not necessarily mean that the respective concentrations

Table 3
Values of model performance criteria over model types and structural parameters – corrected initial value data of data Set 1 of the experimental case study on the *Bordetella pertussis* cultivation.

Model type	Structure	BIC train	BIC valid	BIC test	MSE train	MSE valid	MSE test
ARX-PLS	[lv[a] = 6, nt[b] = 3]	−550	−213	−53	0.1582	0.0784	0.3047
Hybrid-NPLS	[lva = 2]	−329	−54	22	0.1019	0.0371	0.0018

[a] lv: number of latent variables.
[b] nt: number of time series elements.

are linearly correlated in the same way, because the initial value of the concentrations poses a bias.

In such context it has to be kept in mind that prior to the application of the chemometric tools the data are, as usual, zero-mean-centred and scaled by the standard variance, which might also contribute to the bias. Hence, three latent variables would be justifiable in the identification of the reference (N)PLS models.

The extra latent variable in these structures might be thought to account for the dynamics, which however is for the cost of a higher number of parameters involved, with the subsequent cost of lower BIC value.

In general, it is worth noting that for the identification of the number of latent variables for the hybrid model the kinetic dimensions with or without mechanistic knowledge incorporation can be reduced to two independent rates, which might suggest that any additional kinetic rate of the simulation model may be redundant.

3.4. Challenges of the experimental case study

The challenge in this case study on *B. pertussis*, arises mainly from the typical infrequent, sparse and noisy experimental concentration data available. The main objective was to show that the developed hybrid model is under these circumstances competitive with the reference dynamic (N)PLS models. The number of latent variables was unknown a priori and such was also subject of the study.

3.4.1. The "best" number of latent variables

The BIC values of the hybrid models were significantly better when compared to the ones of the reference dynamic (N)PLS approaches, this being mainly due to the smaller number of modeling parameters involved in the former.

For both data sets of this study (Section 3.1.2) the BIC values obtained on the application of the hybrid models to the validation batches suggest the selection of two latent variables, which is partially in agreement with the reference (N)PLS structures identified (see Table 2). It has been seen that due to defective initial values, the MSE values obtained for the same validation batches were inconsistent among themselves, but the BIC values obtained for the corrected files nevertheless reinforce the selection of two latent variables.

For the case of the reference dynamic (N)PLS models, it was observed that in general five to six latent variables are necessary to obtain model performances which are, in terms of the MSE, in the same range than those of the hybrid models. Exceptions to this observation exhibit the performances of the dynamic NPLS models of Set 2 (Table 2), which both only comprise two latent variables. It seems that nonlinear inner functions are capable to account better for the general process dynamics than linear ones. This assumption is further supported by the observation that the MSE values of the test data obtained for the nonlinear models are significant smaller than the ones obtained for the linear models.

3.4.2. General performance

The following observations hold concerning general model performance:

(i) The MSE values obtained for the hybrid structures are seen to be significantly better than the ones obtained with the (N)PLS structures for the test set, as presented in Table 2. The correction of those initial values of the substrate concentrations in the test batch (initial data corrected as described in Section 3.2.2), have lead to further improved performance in terms of MSE for the hybrid model, as expressed in Table 3.
Considering that only the initial substrate concentrations were corrected, it can be further concluded that the hybrid model captures well the known fact that the estimation of the biomass concentration is sensitive to the initial substrate concentration.

(ii) The MSE values obtained with the application of the hybrid structure to the "corrected" validation batches (initial data corrected as described in Section 3.2.2) were significantly better than those of the reference ARX-PLS model.
In the context of this analysis, it should be pointed out that

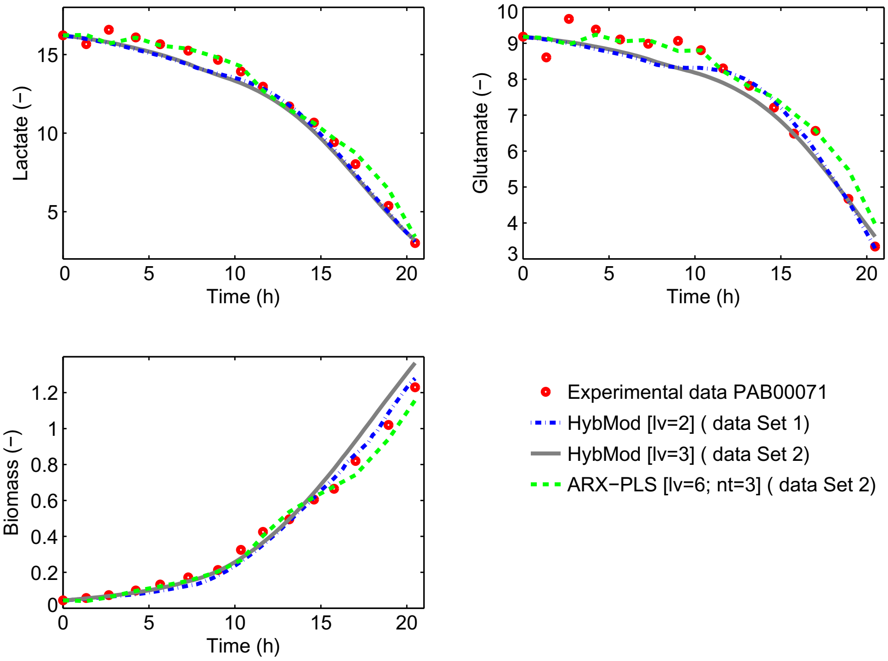

Fig. 4. *Bordetella pertussis* experimental case study – plots of concentrations of lactate, glutamate and biomass concentrations over time for the validation batch PAB00071 (red dots): predictions of the NPLS hybrid model with 2 latent variables (dashed dotted blue line) and 3 latent variables (grey line), vs. estimates of a ARX-PLS, with 3 latent variables (dashed green line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

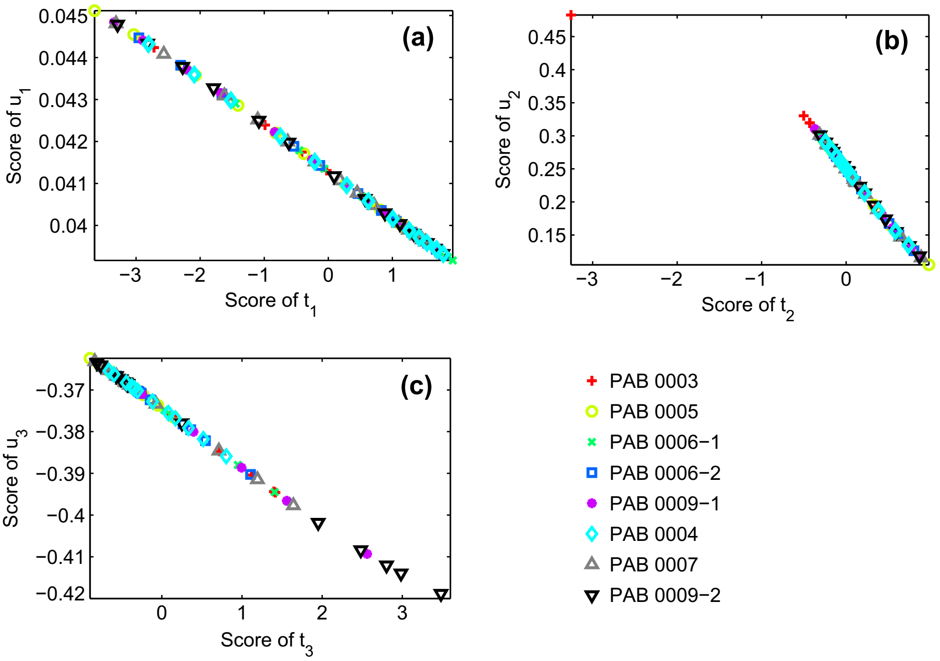

Fig. 5. Bordetella pertussis experimental case study – phase plane plots of input scores, t_i, vs. output scores, u_i, (i = 1, 2, 3 in *a*, *b*, *c*, respectively) for the three latent variables of the inner model ANN functions of the hybrid structure comprising 3 latent variables – application to all batches (PAB0003 red crosses; PAB0005 light green circles; PAB0006-1 green x-es; PAB0006-2 blue boxes; PAB0009-1 purple filled squares; PAB0004 turquoise diamonds; PAB0007 gray upward-pointing triangles and PAB0009-2 black downward-pointing triangle). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the ARX-PLS reference model exhibit a rather low sensitivity to initial values. This can be observed in the difference between the MSE values of the corrected and uncorrected test batch, which is of the order 0.0081, about eleven percent of the respective MSE value.

(iii) The good performance of the hybrid models for the estimation of the biomass concentration is also observed in Fig. 4. The superior performance of the hybrid models is strengthened by comparing, in the same figure, the shape of the trajectories, which are rather bumpy for the (N)PLS model against rather smooth trajectories for the hybrid models (especially for the one with three latent variables). In the case of the dynamic (N)PLS models, the mandatory pretreatment of the data, i.e. the application of a cubic smoothing spline (Section 3.1.2), therefore does not seem to act smoothing on the estimates, but instead the error introduced through the data interpolation seems to board the predictions.

3.5. Complementary features of the hybrid model

In the Park Ramirez case study, it was observed that the identification of the nonparametric model parameters exhibited a faster convergence, a higher consistency of the results and an improved performance, e.g. in form of the MSE criteria in Table 1, when comparing between hybrid models with and without mechanistic knowledge, in favor of the former. Thus, the incorporation of mechanistic knowledge into the hybrid structure leads to a better model performance, which is in line with observations in Psichogios and Ungar (1992), Oliveira (2004).

It is known that for (N)PLS models the analysis of the input and output scores represents a relevant source of information concerning characteristics and features of the processes and of model performance. This important feature of (N)PLS structures is present in the hybrid model developed in this study. For instance, and as illustration with the experimental case study, the plot of scores u_2 over t_2 (Fig. 5) shows a singularity of behavior for batch PAB0003 (see red crosses). This batch, employed in the training stage, distinguishes from the others by (1) having the smallest initial values for all concentrations, namely lactate, glutamate and biomass; (2) exhibiting the highest concentration of biomass in the end of the batch; and (3) comprising a defect in the DO signal towards the end of the batch.

Finally, and still for the experimental case study, the fairly linear inner model functions, which can be seen in Fig. 5, might explain for the fast convergence and the consistency of the hybrid model performance, as in general the optimal parameters of linear models are unique.

4. Conclusion

A novel methodology consisting of a hybrid dynamic (N)PLS model together with an algorithm for parameter identification is proposed for bioprocess modeling. The model consists of a set of macroscopic material balance equations in which the kinetic rates (the reaction terms) are mimicked by a nonlinear partial least square (NPLS) submodel and wherefore the global approach belongs to the class of hybrid models.

This methodology was benchmarked against reference dynamic (N)PLS models (in which the dynamics are modeled by the augmentation of the inputs by lagged variables, such as FIR or AR(X)) through the application to two complementary case studies; (i) a simulation case study, also called the Park Ramirez simulation case after (Park & Ramirez, 1988); and (ii) an experimental case study of a *B. pertussis* cultivation, as published by Soons et al. (2008, 2008).

The following has been observed and can be stated:

(i) The novel approach, due to its inherent dynamics, exhibits, in general, fewer model parameters which results in a higher statistical confidence, observed in form of higher BIC values, when compared to the reference dynamic (N)PLS models.

(ii) In the application to validation data, the model performance, observed in terms of the MSE criterion, was generally significantly better.

(iii) Better calibration properties can be observed, expressed in terms of extrapolation capabilities to broader process conditions (e.g. predictions concerning the abnormal fed-batch data).

(iv) The application of the proposed model to typical infrequent, sparse and noisy experimental data leads to realistic, smooth trajectory estimations of the process states and does not require data interpolation as necessary in the reference dynamic (N)PLS methods.

(v) The integration of mechanistic knowledge into the proposed framework was identified to have a significant impact on the good results obtained, which is in line with the findings of Psichogios and Ungar (1992), Oliveira (2004).

(vi) The novel proposed nonparametric structure and the related parameter identification algorithm exhibit PLS features such as dimension reduction and the opportunity to interpret the plot of scores:
   (a) The Park Ramirez case study involves four kinetic rates, where two of which are linearly correlated. The hybrid model revealed that only two independent latent variables are already sufficient to model the process, in contrast to mostly four obtained by the reference (N)PLS models.
   In general fewer latent variables were required regarding the same process than by the reference dynamic models.
   (b) For the *B. pertussis* case study, from the analysis of the score plots, it was shown that unusual variations in the process conditions could be identified.

(vii) Several sources of errors were identified: (a) noise in the input measurements to the nonparametric model; (b) noise in the measurements of the feeding rates (in the Park Ramirez case study); (c) errors inherent to the feedback nature of the models (where applicable); or (d) defective initial values.

(viii) For all sources of errors, except for the case of defective initial values, it was observed that state feedback to the nonparametric model had a damping effect on error propagation.

(ix) For cases of defective initial values, it was shown that corrective action on such errors has led to improved performance of the hybrid approach in comparison to the reference dynamic (N)PLS models (e.g. a more than twofold improvement of the MSE value in the experimental case study on a *B. pertussis* cultivation).

In all, it can be stated that the application of a suitable hybrid (N)PLS model structure leads to significantly enhanced process estimations when compared to the reference dynamic (N)PLS models.

Acknowledgment

Sincere thanks for the provided data go to the Netherlands Vaccine Institute and for financial support to the Fundacao para a Ciencia e a Tecnologia, where the reference number of the provided scholarship to Moritz von Stosch is: SFRH/BD/36990/2007.

Appendix A

A.1. The calculation of the input and output scores

The input and output scores are an inherent component of the proposed nonparametric structure, Eq. (5). The scores, in analogy to (N)PLS models, give an insight into the information captured by the respective submodel and are further suitable to identify "abnormal" process behavior.

The input scores, also called input latent variable, are directly obtained from multiplication of the input vector $L_{i,1\ldots k}$ (see Eq. (5)) with the input loadings, $W_{x,i}$, i.e.:

$$t_i = W_{x,i} \cdot L_{i,1\ldots k}. \tag{A.1}$$

The output scores are obtained by processing the input scores, t_i, with the ANN inner model, such that:

$$u_i = \left(w_{2,i} \cdot g\left(w_{1,i} \cdot h(t_i) + b_{1,i}\right) + b_{2,i}\right), \tag{A.2}$$

using the weights, biases and functions defined in Section 2.2.1.

A.2. The sensitivity equations

The derivative of Eq. (10) can be split into the derivative of $dW_{x,i,\text{lin}}/dw_A$ and in $dW_{x,i}/dw_A$, where w_A, the vector of parameters, comprises $W_{x,i}$, $W_{y,i}$ and w. The latter derivative is straight forward as described above, Section 2.2.2. Considering Eq. (9), the derivative $dW_{x,i,\text{lin}}/dw_A$ can be extended to:

$$\frac{dW_{x,i,\text{lin}}}{dw_A} = \frac{dW_{x,i,\text{lin}}}{dW_{x,i,\text{lin},un}} \cdot \frac{dW_{x,i,\text{lin},un}}{dw_A}, \tag{A.3}$$

making use of the chain rule. The first term on the right hand side is equivalent to Eq. (13). The second term on the right hand side is the derivative of Eq. (8) with respect to w_A. This term can be reformulated using the quotient rule to:

$$\frac{dW_{x,i,\text{lin},un}}{dw_A} = \frac{\left(t_i^T \cdot t_i\right) \cdot \frac{d\left(L_{i,1\ldots k} \cdot t_i\right)}{dw_A} - \left(L_{i,1\ldots k} \cdot t_i\right) \cdot \frac{d\left(t_i^T \cdot t_i\right)}{dw_A}}{\left(t_i^T \cdot t_i\right)^2}. \tag{A.4}$$

The first derivative in the numerator can be split up, applying the chain rule again, to:

$$\frac{d\left(L_{i,1\ldots k} \cdot t_i\right)}{dw_A} = t_i \cdot \frac{dL_{i,1\ldots k}}{dw_A} + L_{i,1\ldots k} \cdot \frac{dt_i}{dw_A}. \tag{A.5}$$

The second derivative can equivalently be treated, giving:

$$\frac{d\left(t_i^T \cdot t_i\right)}{dw_A} = 2 \cdot t_i \cdot \frac{dt_i}{dw_A}. \tag{A.6}$$

The derivative dt_i/dw_A emerges in (A.5) and (A.6), which, considering Eq. (A.1) and applying the chain rule, can be formulated to:

$$\frac{dt_i}{dw_A} = W_{x,i} \cdot \frac{dL_{i,1\ldots k}}{dw_A} + L_{i,1\ldots k} \cdot \frac{dW_{x,i}}{dw_A}. \tag{A.7}$$

Noting that $w_A^T = [W_{x,i}, W_{y,i}, w]$, then the derivative corresponding to the second term on the right hand side is a matrix comprising the identity submatrix for the derivative of $W_{x,i}$ with respect to $W_{x,i}$ and zero elsewhere.

The derivative in the first term on the right side, namely $dL_{i,1\ldots k}/dw_A$, also appears in Eq. (A.5) and is reformulated using Eq. (5) to:

$$\frac{dL_{i,1\ldots k}}{dw_A} = \frac{dL_{i-1,1\ldots k}}{dw_A} - \frac{d\left(W_{x,i-1} \cdot L_{i-1,1\ldots k} \cdot W_{x,i-1}\right)}{dw_A}, \tag{A.8}$$

where the second term on the right side can be simplified by using the chain rule, a straightforward solution and therefore not carried out here.

The only remaining derivative is $dL_{i-1,1\ldots k}/dw_A$, which is calculated sequentially, starting with $dL_{1,1\ldots k}/dw_A$. It should be noted that only the partition of entries of $L_{1,1\ldots k}$, corresponding to the feedback of model estimates into the nonparametric model (Fig. 1) depend on w_A. As such those derivatives reduce to dc/dw_A which are nothing else than the derivatives given by Eqs. (11), (14), (15).

References

Baffi, G., Martin, E. B., & Morris, A. J. (2000). Non-linear dynamic projection to latent structures modelling. *Chemometrics and Intelligent Laboratory Systems, 52*(1), 5–22.

Bishop, C. (1995). *Neural networks for pattern recognition*. New York: Oxford University Press, Inc..

Burnham, K., & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304.

Clementschitsch, F., & Bayer, K. (2006). Improvement of bioprocess monitoring: Development of novel concepts. *Microbial Cell Factories, 5*(1), 19.

Eykhoff, Pieter (1974). *System identification: Parameter and state estimation*. London, New York: Wiley-Interscience.

Frank, P. M. (1978). *Introduction to system sensitivity theory*. New York: Academic Press.

Haykin, S. (1998). *Neural networks – A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Henneke, D., Hagedorn, A., Budman, H., & Legge, R. (2005). Application of spectrofluorometry to the prediction of phb concentrations in a fed-batch process. *Bioprocess and Biosystems Engineering, 27*(6), 359–364.

Kulkarni, S. G., Chaudhary, A. K., Nandi, S., Tambe, S. S., & Kulkarni, B. D. (2004). Modeling and monitoring of batch processes using principal component analysis (PCA) assisted generalized regression neural networks (GRNN). *Biochemical Engineering Journal, 18*(3), 193–210.

Lakshminarayanan, S., Shah, S. L., & Nandakumar, K. (1997). Modeling and control of multivariable processes: Dynamic PLS approach. *AIChE Journal, 43*(9), 2307–2322.

Lee, D. S., Vanrolleghem, P. A., & Park, J. M. (2005). Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. *Journal of Biotechnology, 115*(3), 317–328.

Leonard, T., & Hsu, J. (1999). *Bayesian methods*. New York: Cambridge University Press.

Ljung, L. (1991). Issues in system identification. *Control Systems Magazine, 11*(1), 25–29.

Oliveira, R. (2004). Combining first principles modelling and artificial neural networks: A general framework. *Computers & Chemical Engineering, 28*(5), 755–766.

Park, S., & Ramirez, W. F. (1988). Optimal production of secreted protein in fed-batch reactors. *AIChE Journal, 34*(9), 1550–1558.

Peres, J., Oliveira, R., & de Azevedo, S. F. (2008). Bioprocess hybrid parametric/nonparametric modelling based on the concept of mixture of experts. *Biochemical Engineering Journal, 39*(1), 190–206.

Peres, J., Oliveira, R., & Feyo de Azevedo, S. (2001). Knowledge based modular networks for process modelling and control. *Computers & Chemical Engineering, 25*(4-6), 783–791.

Preusting, H., Noordover, H., Simutis, R., & Lübbert, A. (1996). The use of hybrid modelling for the optimization of the penicillin fermentation process. *CHIMIA, 50*(9), 416–417.

Psichogios, D., & Ungar, L. (1992). A hybrid neural network-first principles approach to process modeling. *AIChE Journal, 38*, 1499.

Qin, S. (1993). Partial least squares regression for recursive system identification (Vol. 3, pp. 2617–2622).

Qin, S., & McAvoy, T. (1992). Nonlinear PLS modeling using neural networks. *Computers & Chemical Engineering, 16*(4), 379–391.

Qin, S. J., & McAvoy, T. J. (1996). Nonlinear fir modeling via a neural net PLS approach. *Computers & Chemical Engineering, 20*(2), 147–159.

Ricker, N. (1988). The use of biased least-squares estimators for parameters in discrete-time pulse-response models. *Industrial & Engineering Chemistry Research, 27*(2), 343–350.

Schubert, J., Simutis, R., Dors, M., Havlik, I., & Luebbert, A. (1994a). Bioprocess optimization and control: Application of hybrid modelling. *Journal of Biotechnology, 35*(1), 51–68.

Schubert, J., Simutis, R., Dors, M., Havlfk, I., & Luebbert, A. (1994b). Hybrid modelling of yeast production processes – combination of a priori knowledge on different levels of sophistication. *Chemical Engineering & Technology, 17*(1), 10–20.

Simutis, R., Oliveira, R., Manikowski, M., de Azevedo, S. F., & Luebbert, A. (1997). How to increase the performance of models for process optimization and control. *Journal of Biotechnology, 59*(1–2), 73–89.

Soons, Z. I. T. A., Shi, J., Stigter, J. D., van der Pol, L. A., van Straten, G., & van Boxtel, A. J. B. (2008). Observer design and tuning for biomass growth and k(l)a using online and offline measurements. *Journal of Process Control, 18*(7–8), 621–631.

Soons, Z. I. T. A., Streefland, M., van Straten, G., & van Boxtel, A. J. B. (2008). Assessment of near infrared and "software sensor" for biomass monitoring and control. *Chemometrics and Intelligent Laboratory Systems, 94*(2), 166–174.

Thompson, M. L., & Kramer, M. A. (1994). Modeling chemical processes using prior knowledge and neural networks. *AIChE Journal, 40*(8), 1328–1340.

Wang, H., & Yu, J. (2004). Application study on nonlinear dynamic fir modeling using hybrid svm-pls method (Vol. 4, pp. 3479–3482).

Werbos, P. (1988). Backpropagation: Past and future (Vol. 1, pp. 343–353).

Wold, S., Kettaneh-Wold, N., & Skagerberg, B. (1989). Nonlinear PLS modeling. *Chemometrics and Intelligent Laboratory Systems, 7*(1–2), 53–65.