Nonlinear Model Reduction for Optimization Based Control of Transient Chemical Processes

Wolfgang Marquardt*
Lehrstuhl für Prozesstechnik
RWTH Aachen
D-52064 Aachen, Germany

Abstract

Optimization based control aims at maximizing the economical performance of a plant in a transient environment employing nonlinear models. Model quality is crucial for achieving good economical performance. The models have to represent plant behavior with sufficient detail, but the computational complexity must be limited to facilitate real-time optimization on a time horizon oriented at the dominating time constant of the process. This contribution reviews nonlinear model reduction techniques from an optimization based control perspective. The use of different variants of process models in a structured control system is particularly emphasized. Challenging research directions are identified.

Keywords

Fundamental modeling, Nonlinear model reduction, Nonlinear model predictive control, Real-time optimization, Transient operation, Process dynamics

Introduction

Optimization based control refers to the class of control techniques which predict the process behavior by means of a dynamic plant model and *optimize the economical performance* of a process while *satisfying operational constraints*. For continuous processes, predominantly operated at steady-state, this concept is implemented in industrial practice by a real-time optimization system (Marlin and Hrymak, 1997), which by means of a steady-state model, computes the setpoints for the model predictive controller (Henson, 1998; Morari and Lee, 1999; Allgöwer et al., 1999; Rawlings, 2000; Mayne et al., 2000) which itself provides setpoints for the base control systems. In industrial applications, model predictive controllers (MPC) are almost exclusively based on a linear plant model determined from experimental identification using plant test data (Qin and Badgwell, 1996). Nonlinear model predictive control has not yet gained significant industrial interest (Qin and Badgwell, 2000) despite the inherently nonlinear behavior of most process plants. This is partly due to the state of technology which yet neither provides mature tools to assist model development nor sufficiently robust algorithms to reliably solve the optimization problem on-line. However, even if these shortcomings could be overcome, nonlinear control technology will only be applied if the significantly increased effort in designing, implementing and maintaining such controllers leads to a significant improvement of plant economics as compared to state of the art linear control technology.

If only stationary operational phases are considered, the need for nonlinear optimizing control technology can hardly be justified, since only few practical situations such as non-minimum phase behavior or steady-state multiplicity may call for a nonlinear controller. However, chemical process systems are often operated in transient phases, where all process variables are *intentionally time-varying*. Transient operation is not only limited to processes which are of an inherent dynamic nature such as batch and periodically forced processes, but it finds increasing attention also in continuous processes to implement feedstock switching or product grade transitions (Helbig et al., 2000a), to realize cross-functional coordination between units in a plant or a site (Lu, 2000) or to exploit low frequency disturbances in the dynamically changing environment of a plant in supply chain conscious plant operation (Backx et al., 1998).

In *transient plant operation*, the operational envelope of the plant naturally covers a large region of the state space. The dynamics can therefore not adequately be represented by a linear model. Hence, *nonlinear models and nonlinear control techniques* are indispensible to achieve satisfactory performance. However, it is not only the nonlinearity which distinguishes the control problem in transient operation from its stationary counterpart. In transient operation, the control task must be considered from a wider perspective. Instead of maintaining a setpoint or tracking a trajectory given by a superior decision layer and rejecting disturbances, the control system has to achieve the *best possible economical performance* within a constrained operational region despite the dynamically changing environment of the plant (Backx et al., 1998). Hence, the targets in stationary phases (Rawlings, 2000) are replaced by an economical objective in transient phases (Helbig et al., 2000b) which consequently results in an integration of optimizing control and on-line economical (dynamic) optimization by means of receding horizon techniques (Backx et al., 2000). We introduce the term *operation support system* to emphasize the wider scope which goes beyond mere setpoint control, trajectory tracking, and disturbance rejection.

These trends in process operations and control necessitate models of sufficient rigor, which are suitable for

*marquardt@lfpt.rwth-aachen.de

implementation in an optimization based operation support system. Since these models must cover the whole operational envelope of a plant, purely *empirical process models* seem to be unfavorable due to the immense effort required for plant testing (Pearson and Ogunnaike, 1997). Consequently, models for optimization and control should capture the major physico-chemical effects in a mechanistic manner at least to the extent accessible. The more a-priori-knowledge can be built into a *fundamental process model* the less experimental effort will be required to fit the model to plant data in order to obtain good extrapolation capabilities in a large region of the state space.

Modeling is considered one of the major bottlenecks of nonlinear model predictive control (Henson, 1998; Lee, 2000) or, more generally, of optimization based process operations and control in the sense of Backx et al. (1998) and Helbig et al. (2000a). Lee (2000) discusses the requirements, the current status and the needs of nonlinear modeling and identification for control and operations with an emphasis on *experimental identification*. This paper aims at complementing and partly detailing Lee's assessment by focusing on *nonlinear model reduction* for the implementation of optimization based operations support systems.

A comprehensive and sensible review of this subject is a formidable task which can hardly be achieved given the limited space available. Hence, the selection of the material presented reflects at least to some extent the interest and the ignorance of the author. The references given should be taken as exemplary rather than as comprehensive. They have been chosen to point the interested reader to relevant approaches and results and to provide a first guide to a more detailed literature study.

Modeling always has to be oriented towards a projected use in an application (Foss et al., 1998). Hence, the next section introduces first a mathematical formulation of the *optimization based operation support problem*, suggests some decomposition strategies and derives general requirements the models have to fulfill. The major phases of a *systematic work process for the development of (fundamental) process models* are given in the following section. The resulting models show a natural structure which should be exploited in model application. This structure can be related to *hybrid modeling* which is discussed in the following section.

Fundamental models are typically of a high computational complexity which is difficult to handle by online optimization algorithms (Henson, 1998). Therefore, nonlinear model reduction techniques are of significant relevance if large-scale applications have to be tackled. Consequently, model reduction techniques are discussed in great detail next. We consider both, *model order reduction* to reduce the number of equations, and *model simplification* to reduce the complexity of the dynamic process model. With the final section, we return to the *optimization based operation support system* and discuss which type of models are good candidates for implementing the different modules in a potentially decomposed system. We conclude with a summary of important *open research problems*.

Optimization Based Operation Support

We introduce a general problem formulation for optimization based operation support and discuss potential decomposition strategies for implementing such a control system in an industrial environment (see Backx et al. (1998), Helbig et al. (2000b) and Backx et al. (2000) for a more detailed discussion). Resulting requirements on the models are summarized to guide fundamental modeling and model order reduction and simplification.

Mathematical Problem Formulation

The goal of optimal process control and operations is the minimization of some economic cost function Φ_i over a certain time horizon $\Delta_i = [t_{0,i}, t_{f,i}]$, both set by a decision maker on a higher level in the automation hierarchy (e.g. a planner or a scheduler), in face of unknown parametric or unstructured model uncertainty and time-varying exogenous inputs represented by the disturbance vector $\boldsymbol{d}_i(t)$. The minimization is subject to the model equations $\boldsymbol{f}(\cdot)$ and to production and process constraints $\boldsymbol{m}_i(\cdot)$ and $\boldsymbol{g}_i(\cdot)$ such as product quality and capacity restrictions or equipment limitations. The constraints $\boldsymbol{m}_i(\cdot), \boldsymbol{g}_i(\cdot)$ can be either path, point or periodicity constraints. The final time $t_{f,i}$ of the operational phase Δ_i is determined by the operational objectives. It could, for example, be the final batch time or the time after which a grade change in a continuous process has been completed. Feedback of available measurements $\boldsymbol{\eta}_i(t)$ of the system outputs $\boldsymbol{y}_i(t)$, inputs $\boldsymbol{u}_i(t)$ and disturbances $\boldsymbol{d}_i(t)$ is introduced to achieve satisfactory performance. Typically, an estimation of the states $\boldsymbol{x}_i(t)$ and the disturbances $\boldsymbol{d}_i(t)$ is required to implement output feedback with high performance. We drop the index i denoting a specific operational phase subsequently to ease notation in the sequel.

The overall output feedback problem can be separated into a dynamic data and model reconciliation problem and into an optimal control problem. At any time instance $t_c \in \Delta$, the reconciliation problem

$$\min_{x_{r,0}, d_r} \Phi_r(\boldsymbol{y}_r, \boldsymbol{\eta}, \boldsymbol{x}_{r,0}, \boldsymbol{d}_r, t_r, t_c) \tag{1}$$

subject to

$$\begin{aligned} 0 &= \boldsymbol{f}_r(\dot{\boldsymbol{x}}_r, \boldsymbol{x}_r, \boldsymbol{u}_r, \boldsymbol{d}_r) \ , \\ \boldsymbol{x}_r(t_r) &= \boldsymbol{x}_{r,0} \ , \\ \boldsymbol{y}_r &= \boldsymbol{h}_r(\boldsymbol{x}_r, \boldsymbol{u}_r, \boldsymbol{d}_r) \ , \\ \boldsymbol{u}_r &= \mathcal{U}[\boldsymbol{u}_c(\cdot)] \ , \\ 0 &\geq \boldsymbol{g}_r(\boldsymbol{x}_r, \boldsymbol{u}_r, \boldsymbol{d}_r) \ , \end{aligned}$$

$$t \in [t_r, t_c]$$

and the control problem

$$\min_{u_c} \Phi_c(\boldsymbol{x}_c, \boldsymbol{u}_c, t_c, t_f) \tag{2}$$

subject to

$$\begin{aligned} 0 &= \boldsymbol{f}_c(\dot{\boldsymbol{x}}_c, \boldsymbol{x}_c, \boldsymbol{u}_c, \boldsymbol{d}_c) \ , \\ \boldsymbol{x}_c(t_c) &= \boldsymbol{x}_r(t_c) \ , \\ \boldsymbol{y}_c &= \boldsymbol{h}_c(\boldsymbol{x}_c, \boldsymbol{u}_c, \boldsymbol{d}_c) \ , \\ \boldsymbol{d}_c &= \mathcal{D}[\boldsymbol{d}_r(\cdot)] \ , \\ 0 &\geq \boldsymbol{g}_c(\boldsymbol{x}_c, \boldsymbol{u}_c, \boldsymbol{d}_c) \ , \\ 0 &\geq \boldsymbol{m}_c(\boldsymbol{x}_c, \boldsymbol{u}_c, \boldsymbol{d}_c) \ , \end{aligned}$$

$$t \in [t_c, t_f]$$

have to be solved in real-time on horizons $[t_r, t_c]$ and $[t_c, t_f]$, respectively. The indices r and c refer to quantities in the reconciliation and the control problem, respectively. Purposely, we have assumed that neither the models nor the production and process constraints are the same in the reconciliation and control problems. For the sake of simplicity, we have not explicitly introduced the discrete nature of measurements $\boldsymbol{\eta}$ and controls $\boldsymbol{u}_c, \boldsymbol{u}_r$. These vectors could, however, be thought of being concatenations of the respective vectors at discrete times t_k in either $[t_r, t_c]$ or $[t_c, t_f]$. Further, though we do not want to exclude hybrid continuous-discrete systems (e.g. Barton et al., 1998; Bemporad and Morari, 1999), we have not accounted for any discrete variables in the problem formulation explicitly for the sake of a simpler notation.

The problems (1) and (2) are coupled and cannot be solved independently. The states $\boldsymbol{x}_r(t_c)$ and the disturbances $\boldsymbol{d}_r(t)$ are estimated in the reconciliation problem (1) and are passed to the control problem (2) to facilitate state and disturbance prediction via the control model and the predictor $\mathcal{D}$. On the other hand, the control variables $\boldsymbol{u}_c(t)$ are passed from the control problem (2) to the reconciliation problem (1) and are processed by $\mathcal{U}$ to update the controls needed for state and disturbance estimation.

From a control perspective, this problem is an output feedback optimal control problem with general (instead of least-squares) objectives reflecting process economics. Since there are no operational targets in the sense of setpoints or reference trajectories (characteristic for a model predictive control problem, see Rawlings (2000)) the problem may also be interpreted from an operational perspective. Hence, the problem can be considered a generalization of state of the art (steady-state) real-time optimization (Marlin and Hrymak, 1997) which aims at establishing economically optimal transient plant operation (Backx et al., 1998, 2000). In any case, the solution of this *operation support problem* would achieve an integration of advanced (predictive constrained) process control and economical optimization in a transient environment.

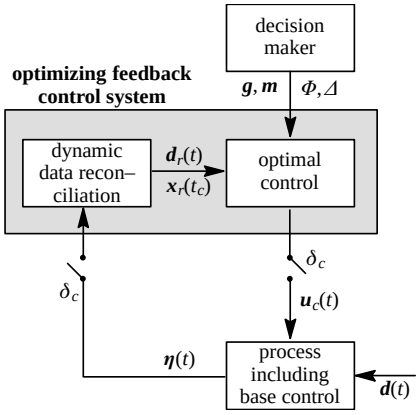

Figure 1: Direct (centralized) optimization approach. δ_c refers to the feedback control sampling time.

Decomposition Strategies

In principle, the problem (1), (2) can be solved simultaneously on the controller sampling frequency δ_c (see Figure 1). This *centralized* or *direct approach* is only computationally tractable if small-scale processes (such as single process units) are considered and/or strongly simplified models are applicable. Obviously, highly efficient solution algorithms and sophisticated model reduction techniques are extremely important to push the frontier of computational complexity (Biegler and Sentoni, 2000). However, even if the problem could be solved easily for large-scale processes, it is questionable whether such an unstructured approach would be accepted by both, the industrial plant operators and the control system vendors. Problems could be associated with a lack of redundancy, reliability and transparency as well as with a high engineering complexity and maintenance effort.

Decomposition of the overall problem seems to be unavoidable in particular if large-scale plant or even site wide control problems with cross-functional integration (Lu, 2000) are considered (Backx et al., 1998). The development of such decomposition strategies can be built on the theory of multi-level hierarchical systems (Mesarovic et al., 1970; Singh, 1977; Findeisen et al., 1980; Morari et al., 1980; Jose and Ungar, 2000) which has been worked out before the mid-eighties. Many of the concepts have not widely been implemented at that time due to a lack of computational power. Though this bottleneck has been largely overcome today, the theory has not yet found adequate attention in the process control community.

Two fundamentally different decomposition strategies can be distinguished. *Horizontal decomposition* refers to

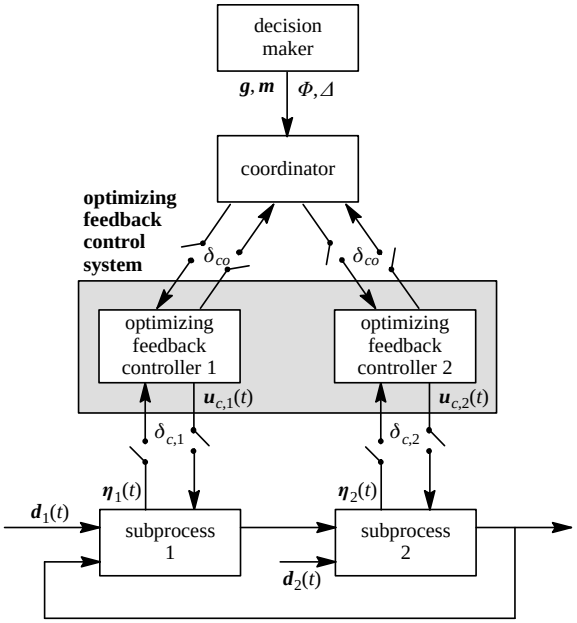

Figure 2: Horizontal decomposition, decentralized optimization approach. δ_c and δ_{co} refer to the feedback control and to the coordination sampling times.

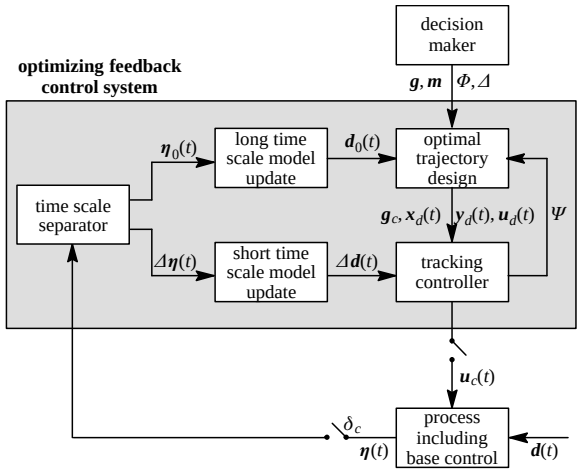

Figure 3: Vertical, two time-scale decomposition of optimization based operations support for transient processes. δ_c refers to the sampling time of the tracking controller, Ψ refers to a process performance indicator.

a decentralization of the control problem typically (but not necessarily, e.g. Lee et al. (2000)) oriented at the functional constituents of a plant (e.g. the process units). Coordination is required to guarantee that the optimal value of the objective reached by a centralized optimizing control system (see Figure 1) can also be achieved by decentralized dynamic optimization. Various coordination strategies for dynamic systems have been described, for example, by Findeisen et al. (1980). Figure 2 shows one possible structure, where the coordinator adjusts the objective functions of the decentralized optimizing feedback controllers to achieve the "true" optimum of the centralized approach.

Vertical decomposition refers to a multi-level separation of the problem (1), (2) with respect to different time-scales. Typically, base control, predictive reference trajectory tracking control, and dynamic economic optimization could be applied with widely differing sampling rates in the range of seconds, minutes, and hours (see Findeisen et al. (1980) for example). According to Helbig et al. (2000b), the feasibility of a multiple time-scale decomposition does not only depend on the dynamic properties of the autonomous system but also on the nature of the exogenous inputs and disturbances. If, for example in a stationary situation, the disturbance can be decomposed into at least two contributions,

$$\boldsymbol{d}(t) = \boldsymbol{d}_0(t) + \Delta\boldsymbol{d}(t) \;, \tag{3}$$

a *slow trend* $\boldsymbol{d}_0(t)$ fully determined by slow frequency contributions and an additional *zero mean contribution* $\Delta\boldsymbol{d}(t)$ containing high frequencies, some sort of decomposition should be feasible. Figure 3 shows a possible structure of the optimization based operations support system in this case. The upper level is responsible for the design of a desired optimal trajectory $\boldsymbol{x}_d(t), \boldsymbol{u}_d(t), \boldsymbol{y}_d(t)$ whereas the lower level is tracking the trajectory set by the upper level. Due to the time varying nature of the disturbances $\boldsymbol{d}(t)$, feedback is not only necessary to adjust the action of the tracking controller but also to adjust the optimal trajectory design to compensate for variations in $\boldsymbol{d}_0(t)$ and $\Delta\boldsymbol{d}(t)$, respectively. The control action $\boldsymbol{u}_c(t)$ is the sum of the desired control trajectory $\boldsymbol{u}_d(t)$ and the tracking controller output $\Delta\boldsymbol{u}(t)$. Reconciliation is based on the slow and fast contributions $\boldsymbol{\eta}_0(t)$ and $\Delta\boldsymbol{\eta}(t)$ separated by a time-scale separation module. Control and trajectory design are typically executed on two distinct sampling intervals δ_c and $k\delta_c$ with integer $k > 1$. The performance of the controller, coded in some indicator Ψ, needs to be monitored and communicated to the trajectory design level to trigger an update of the optimal trajectory in case the controller is not able to achieve acceptable performance. Though this decomposition scheme is largely related to so-called composite control in the singular perturbation literature Kokotovic et al. (1986), the achievable performance will be determined by the way the time-scale separator is implemented.

Model Requirements

Optimization based control requires appropriate models to implement solutions to the reconciliation and control problems.

The model in (2) must predict the cost function, out-

puts and states over the time horizon $[t_c, t_f]$ with sufficient accuracy. Any notable inaccuracy will inevitably result in an economical loss because of a violation of the constraints or a deviation from the true economic optimum. The requirements on the prediction quality of disturbances on $[t_c, t_f]$ are high, since they influence the cost function in (2). This is in contrast to model predictive control, where even a crude disturbance prediction is sufficient to eliminate offset in trajectory tracking. The same (high) accuracy requirements hold in the whole operational envelope covered during nominal plant operation.

The prediction accuracy on the control horizon $[t_c, t_f]$ crucially depends on the quality of the state and disturbance estimates on the reconciliation horizon $[t_r, t_c]$ as determined from a solution of the reconciliation problem (1) employing the accessible measurements. Due to unavoidable model uncertainty, the model needs to be reconciled simultaneously. In the most simple case, a number of carefully chosen model parameters has to be updated periodically. Often, unknown time-varying exogenous functions or plant upsets and operator interaction at discrete time instances complicate the reconciliation problem.

Since all the models are used as part of an optimization algorithm, the gradients with respect to $\boldsymbol{x}_{0,r}$ and the parameterization of $\boldsymbol{d}_r$ in (2) and the parameterization of $\boldsymbol{u}_r$ in (1) must be of high accuracy, too, to avoid unnecessary iterations or even convergence to the wrong optimum (see Biegler et al. (1985) and Ganesh and Biegler (1987) for a discussion in steady-state optimization).

Accuracy requirements are much higher here as compared to setpoint or trajectory tracking feedback control. Since any model uncertainty directly influences plant economics, we cannot rely on feedback only to cope with model uncertainty. From a plant economics point of view, a quantification of the model error would be desirable (Tatrai et al., 1994) though hardly achievable in practice. Model validation against plant and cost data is extremely important and needs to be an integral part of the model development activity. We should keep in mind that the predictive quality of the model has to be assessed in a closed-loop rather than an open-loop mode. For one, the gain and frequency characteristics of the model usually differ in open- and closed-loop. Further, plant measurements can be eventually used to update the model as part of the reconciliation problem (1) to compensate for deficiencies in the predictive capabilities of the model.

Due to the high complexity of any large-scale industrial plant, model reduction has always to be considered. Inevitably, any reduction of the model complexity introduces inaccuracies, which—if not significantly smaller in magnitude than the mismatch between plant and original model—will lead to a loss of economical performance and may even give rise to instability. Despite this fact, a compromise between model complexity and predictive quality must always be achieved (Tatrai et al., 1994) since low sampling frequency control action as a consequence of the high complexity of a very accurate model would also reduce the performance of the operation support system.

Though controllability and observability are properties of the plant rather than the model, these structural properties may get lost in case of simple plant models if not properly accounted for. The same holds for the identifiability of model parameters which is not only a matter of the available measurements but also of the detail built into the model. Of course, these requirements are obvious but difficult to assess in the nonlinear case. There is some evidence that fast time-scales should be eliminated from the model to yield better robustness and more favorable stability properties (Christofides and Daoutidis, 1996; Kumar et al., 1998).

To facilitate numerical treatment, the models should be of differential index one (Brenan et al., 1996; Martinson and Barton, 2000), proper boundary conditions have to provided in case of distributed parameter models (Martinson and Barton, 2000), and singular arcs (Bryson and Ho, 1975; Kadam, 2000) should be avoided. Discontinuities are still difficult (and will be at least expensive) to handle numerically (Barton et al., 1998) and should therefore be avoided if possible.

In summary, in order to implement the optimization based control system (1), (2), sufficiently accurate models are required for the nominal intrinsic process dynamics, for the economical objectives (including product quality), for the exogenous disturbances, and for sensor and actuator characteristics if they are relevant for process economics. Major sources of structured or unstructured uncertainty should at least be identified as a prerequisite for appropriate model updating. Different tailored models are necessary for the control and reconciliation subproblems even in a direct (centralized) approach (cf. Figure 1). If some decomposition of the operation support system is employed (cf. Figures 2 and 3), model decomposition is an additional major concern.

Validation of the model quality has to be accomplished under closed-loop conditions. A brute force approach could rely on the formulation and solution of an optimization problem which assesses feasibility, flexibility, and controllability according to the classification given by Abel et al. (1998). The problem formulation leads to a large-scale bilevel dynamic optimization problem the reliable solution of which is out of reach at this point for most industrially relevant processes.

Systematic Model Development

The discussion of the last section clearly reveals the complexity of the model requirements. It is therefore not

surprising that modeling and the proper validation of the resulting models is (and will be for a long time) the major bottleneck in introducing model-based operation support systems into industrial application.

Recently, Foss et al. (1998) undertook an industrial field study to identify current industrial practice in process modeling. They identified the same major steps organized nearly the same way in a work process in all the companies included in the study, if only a coarse task granularity is considered. There seem to be no generally practiced patterns on the subtask level. A first analysis of the modeling work process on a detailed level has been attempted recently (Marquardt, 1995; Lohmann and Marquardt, 1996; Lohmann, 1998) in the context of work process centered computer-aided modeling support systems (Jarke and Marquardt, 1996; Bogusch et al., 2001). The approach pursued in these studies is promising but does not yet address the requirements of optimization based control sufficiently. More emphasis has to be put on model transformations (including model order reduction and simplification), model structure discrimination and parameter identification as well as closed-loop model validation in the future.

We are far from a recommended work process which would lead us to a reasonable set of models for a decomposed optimization based operations support system at minimal cost. This section presents the major modeling steps on a coarse granular task level to guide the development of more elaborate modeling work processes (see Foss et al., 1998, for details) and to put model reduction and model application as discussed in the remainder of the paper into perspective.

(a) *Requirements analysis:* A precise problem formulation is necessary but often omitted in process modeling since most of the requirements are still vague. As in any design activity these requirements have to evolve with the model during the modeling process. Major issues are the purpose of modeling and the intended application of the model, the quantities to be computed from the model, their dependency on time and non-time coordinates, the accuracy to be attained, the available resources for model construction and the available computational resources for model interpretation.

(b) *Abstraction of the process:* The boundaries of the process under consideration are specified by stating all external connections to the process' environment first. Subsequently, the process is decomposed hierarchically into more and more refined interconnected model objects until a desired level of resolution is reached. The properties of the model objects are described in detail. The information collected comprises an *informal descriptive representation* of the model. The extensive quantities to be balanced, the assumptions on the physico-chemical phenomena and the level of detail to be considered are for example part of this description. Canonical model objects and a recommended procedure have been defined to guide this abstraction process (e.g. Marquardt, 1995).

(c) *Formulation of model equations:* For every model object, the descriptive model of step (b) is cast into a set of model equations to precisely define the model object's dynamics. The *informal descriptive model* is converted into a *formal mathematical model*. First, the balance equations are determined accounting for the desired spatial resolution. The process quantities occurring in the balance equations are classified as states, parameters or state functions. Parameters are fixed together with an uncertainty interval. State functions are refined by additional constitutive equations. Appropriate initial and boundary conditions are specified. Simultaneously, a consistency check of physical dimensions and units, an analysis of the remaining degrees of freedom or of the differential index can be carried out (Bogusch et al., 2001).
The resulting process model comprises partial differential equations to cope with spatially distributed model objects, integro-differential-equations to represent particulate model objects by population balances, as well as differential-algebraic equations to describe spatially averaged (well-mixed) model objects. Often, the models are in addition of a hybrid discrete-continuous nature (Barton and Pantelides, 1994) to represent physical state events (such as a phase change) or discrete control action (such as a switching controller).

(d) *Aggregation of model equations:* The equations of the whole process model are deduced by an aggregation of those of every model object. This aggregation process follows the hierarchical structure introduced during the abstraction in step (b). Additional constraints due to the aggregation may be introduced. Again, the model is checked for its index as well as for a proper specification of degrees of freedom and initial conditions. The resulting model may be very large-scale and may comprise of some hundred thousand equations if plant-wide or even site-wide optimization is envisaged. The model structure may be exploited later to accomplish horizontal decomposition (cf. Figure 2).

(e) *Model transformation:* Usually, the model equations are not solved as derived during the modeling process. Instead, the model equations are reformulated with different objectives. For example, reformulation or even model reduction are performed to reduce the computational complexity or the index in case of high index models.

(f) *Implementation and verification:* The model is implemented by means of a modeling and simulation tool. Instead of a formal verification to check whether the model satisfies the intent of the modeler (expressed in the requirements formulated in step (a)) the model is run, the simulation results are checked for plausibility and the computational resources are determined.

(g) *Structure discrimination, parameter estimation, and model validation:* An appropriate model structure has to be chosen from a set of candidates and unknown model parameters have to be determined from experimental data. Optimization based experimental design may be applied to reduce the number of experiments required. Typically, the model fitting process works bottom-up starting with the model objects on the lowest level of the aggregation hierarchy. Another data set is used subsequently to validate the nominal model. Note, that only open-loop experiments are possible at this point.

(h) *Documentation:* A complete documentation of the modeling process, the resulting process model, its implementation, and its validation is provided to facilitate the use of the model or of its parts in a later application.

(i) *Model application:* The model is employed for the intended application, i.e. for the implementation of one of the functional modules in an optimization based control system. An objective function, a disturbance model and a set of constraints have to be defined in addition. The application is validated as a whole which, most importantly boils down to a validation of the model in closed-loop. Not only the modeling, but also the quality of the numerical solution and the measurement data have to be accounted for at this stage.

There is a very close link and a large degree of interdependency between steps (a) to (i). Consequently, a large number of iterations cannot be avoided to meet the complicated and widely varying requirements set out at the beginning. Though, it would be extremely useful to better manage the modeling process in order to come up with a satisfactory solution the first time right, a formalization of the modeling process as a prerequisite for proper work process management seems to be completely out of reach today. This is largely due to a lack of understanding of the modeling process as whole. Still, nonlinear modeling for control is rather an art than an engineering science (cf. also Aris, 1991).

The remainder of the paper will largely deal with steps (c), (e) and (i) with an emphasis on a reduction of model complexity as part of the model transformations in step (e).

Hybrid Modeling

The formulation of model equations in step (c) of the work process above is largely depending on the level of process knowledge available. This knowledge can—at least in part—be organized along the natural structure displayed in model equation sets as derived during fundamental modeling (Marquardt, 1995). On the uppermost level, there are the balances (e.g. a component mass balance) which are composed of generalized fluxes (e.g. a reaction rate), which may be computed from constitutive equations (e.g. the reaction rate expression). Recursively, these constitutive equations contain variables (e.g. a rate constant) which may result from other constitutive equations (e.g. an Arrhenius law). The equation system can be organized as bipartite graph with equations and variables representing the two types of nodes (Bogusch and Marquardt, 1997).

For every process quantity occurring in an equation we can decide whether it is treated as a constant or even time-varying parameter (to be eventually estimated online) or whether it will be refined by another equation. In the latter case, these "constitutive equations" may have a mechanistic basis, or alternatively, they may be of a physically motivated semi-empirical or even of a completely empirical nature. Obviously, the process quantities occurring in these equations can be treated the same way on a next refinement level. In most cases, we are not able or (for complexity reasons) not interested in incorporating truly mechanistic knowledge (on the molecular level) to determine the constitutive equations. Instead, we correlate unknown process quantities by means of a (semi-)empirical equation.

Hence, all process models are by definition *hybrid models*, since they comprise fundamental as well as empirical model constituents. The fundamental model constituents typically represent the balances of mass, energy and momentum and at least part of the constitutive equations required to fix fluxes and thermodynamic state functions as functions of state variables. Empirical model constituents, on the other hand, are incorporated in the overall process model to compensate for a lack of understanding of the underlying physico-chemical mechanisms. These empirical model constituents are typically formed by some regression model such as a linear multivariate or an artificial neural network model. The parameters in the regression model are identified from plant data. Therefore, hybrid modeling is also often referred to as combining a fundamental and a data-driven (or experimental) approach.

Empirical Regression Models

Many different ways of combining empirical (or data-driven) and fundamental modeling have been proposed in recent years (see van Can et al., 1996; Agarwal, 1997), for various alternatives). The most important structures

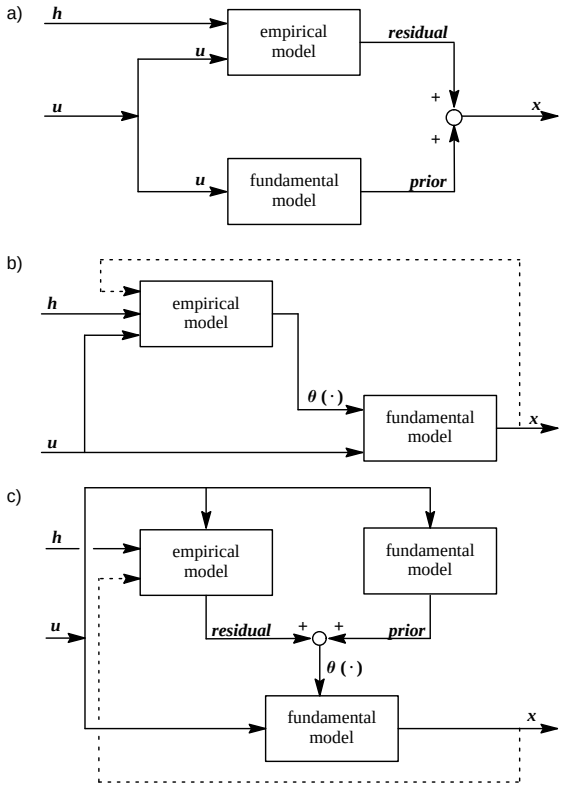

Figure 4: Different structures of hybrid models. (a) parallel structure, (b) serial structure, and (c) an example of a mixed structure.

of combining a fundamental and an empirical model are depicted in Figure 4.

In the *parallel structure*, independently introduced by Su et al. (1992), Kramer and Thompson (1992), Thompson and Kramer (1994), and Johansen and Foss (1992), the model output is a weighted sum of an empirical and a fundamental constituent (cf. Figure 4 (a)). Usually, both of these models are dynamic. The empirical model is often implemented as some type of neural network. It acts as an error model and compensates for any unstructured uncertainty in the fundamental model.

In contrast to this ad-hoc approach to hybrid modeling, the *serial structure* (cf. Figure 4 (b)) is fully consistent with a fundamental model structure. For lumped parameter systems, we find in general on some level of refinement the equation structure

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}(\cdot), \boldsymbol{p}_1, \boldsymbol{u}) \tag{4}$$

with parameters $\boldsymbol{p}_1$ and the unknown function $\boldsymbol{\theta}(\cdot)$. The quantities $\boldsymbol{\theta}$ represent physical quantities which are difficult to model mechanistically. Examples are a flux, a kinetic constant, or a product quality indicator. Typically, $\boldsymbol{f}(\cdot)$ includes the balances of mass and energy, which can always be formulated easily. Obviously, instead of postulating some model structure

$$\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{u}, \boldsymbol{p}_2) \tag{5}$$

which is based on some physical hypotheses as in fundamental modeling, any other purely mathematically motivated model structure can be chosen to implement the constitutive equation. This function and the probably unknown parameters $\boldsymbol{p}_1$ in the fundamental model have to be estimated from the measured outputs $\boldsymbol{\eta}$ and the known inputs $\boldsymbol{u}$ after appropriate parameterization by $\boldsymbol{\theta}(\cdot)$ and $\boldsymbol{p}_2$. This approach to hybrid modeling has been introduced first by Psichogios and Ungar (1992) who suggested the use of a feedforward neural network as the regression model. Identifiability is a serious concern, if $\boldsymbol{\theta}(\cdot)$ not only depends on known inputs $\boldsymbol{u}$ and measured outputs $\boldsymbol{\eta}$ as in most reported studies but also on states $\boldsymbol{x}$.

Any combination of the *parallel and serial structures* in Figure 4 is conceivable. An example, reported by Simutis et al. (1997), is shown in Figure 4 (c).

Hybrid models with serial structure have got a lot of attention. They have found numerous applications in different areas such as in catalysis and multiphase reaction engineering (e.g. Molga and Westerterp, 1997; Zander et al., 1999; Molga and Cherbański, 1999), biotechnology (e.g. Saxén and Saxén, 1996; van Can et al., 1998; Shene et al., 1999; Thibault et al., 2000), polymerization (e.g. Tsen et al., 1996), minerals processing (e.g. Reuter et al., 1993; Gupta et al., 1999), drying (e.g. Zbiciński et al., 1996) or in environmental processes (e.g. de Veaux et al., 1999). In most cases reported so far, the hybrid model has been determined off-line. Satisfactory prediction quality can be obtained if sufficient data are available for training. Typically, the interpolation capabilities are comparable to fully empirical models but the extrapolation capabilities are far superior (e.g. van Can et al., 1998; de Veaux et al., 1999). A tutorial introduction to hybrid modeling with emphasis on the serial structure has been given by te Braake et al. (1998).

On-line updating of the neural network model in the context of model-based control has also been suggested in cases where a high predictive quality cannot be obtained by off-line training due to a lack of sufficient data or to a time-varying nature of the process. An example has been recently reported by Costa et al. (1999) who apply optimal control to a fed-batch fermentation. These authors employ a functional link (neural) network to model the reaction rates in a hybrid model with serial structure and update the parameters on-line to improve control performance.

Empirical Trend Models

Employing nonlinear regression with neural networks for the determination of $\boldsymbol{\theta}(\cdot)$ is not the only one approach to hybrid modeling of uncertain systems. For example, the structure of the model (4), (5) has been explored before

the introduction of hybrid neural network models in the area of reaction calorimetry (see Schuler and Schmidt (1992) for a review). In this context, the unknown function $\boldsymbol{\theta}(\cdot)$ refers to the heat of reaction which is inferred from temperature measurements by some state estimation technique. Instead of the static model (5) a dynamic model

$$\dot{\boldsymbol{\theta}} = \boldsymbol{\vartheta}_1(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{p}_3) \,, \tag{6}$$

$$\dot{\boldsymbol{\pi}} = \boldsymbol{\vartheta}_2(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\eta}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{p}_4) \tag{7}$$

is chosen to complement a fundamental model in the serial structure shown in Figure 4 (b). Here, the quantities $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are interpreted as part of the (extended) state vector rather than as a nonlinear state function. Often, due to a lack of mechanistic knowledge, the dynamic models for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ are chosen in a simple manner. In many cases, constant or linear trends are sufficient to obtain an estimate, which is completely satisfactory for monitoring and control though the predictive capabilities of the model are very limited in theses cases.

This approach is not restricted to reaction calorimetry. For example, Helbig et al. (1998) report on an extension of this idea to real-time optimization of a two-phase polymerisation reactor operated in semi-batch mode to maximize productivity despite lacking knowledge on the detailed physico-chemical phenomena occurring. Empirical models for the overall reaction rate and the interfacial mass transfer rate have been included in a fundamental model comprising mass and energy balances as well as physical property models of the two-phase system. Simple models (6), (7), have been chosen which guarantee observability and controllability even when control variables are constrained.

Wiener/Hammerstein Type Hybrid Models

Hybrid models of a completely different type may be built as follows. In many cases, a fundamental steady-state process model is available either from process design of from steady-state real-time optimization. Often, the effort of converting the existing steady-state model implemented in some process modeling environment to a dynamic model to be employed for the support of transient process operations is quite high. Dynamic real-time optimization may be accomplished at least approximately by means of a hybrid model which combines the nonlinear fundamental steady-state model with an empirical linear dynamic model in a serial manner. The resulting structure corresponds—in the most simple case of a SISO system—to a Hammerstein or a Wiener model depending on the sequence order of the linear dynamic and the nonlinear static submodels.

Three hybrid models of this type—a Wiener, a Hammerstein and a feedback structure, have been treated recently under the assumption of a known nonlinear static map and unknown linear dynamics by Pearson and Pottmann (2000). These authors report tailored identification algorithms which exploit the knowledge of the nonlinear static map given by the fundamental model. In their binary distillation case study, the nonlinear map between reflux and top composition has been approximated by simple spline functions.

In a similar study, reported previously by Norquay et al. (1999), a hybrid model of Wiener type has been developed and employed in a model predictive control strategy for dual composition control of an industrial C_2-splitter in an ethylene plant. First-order plus dead-time linear transfer functions are used to implement an empirical 2×2 linear model followed by two static maps which result from cubic spline approximation of a fundamental steady-state process model. The Wiener model based controller has been successfully implemented on an industrial plant in Australia.

Discussion

Hybrid trend as well as regression models (in particular with serial structure) are not only compensating for lacking mechanistic knowledge. In most cases these models also result in a much lower complexity as compared to representative detailed mechanistic models. Hybrid modeling is therefore closely related to model reduction which is discussed in more detail below. In many cases, the validity of the model can only be guaranteed if online dynamic model reconciliation by either parameter estimation in case of regression models (e.g. Costa et al. (1999)) or by combined state and disturbance estimation in case of trend models (e.g. Helbig et al. (1998)) is employed.

Though there is some evidence that the serial structure has more favorable extrapolation properties than the parallel structure (van Can et al., 1996), methods of systematically designing hybrid model structures—either employing a static regression or a dynamic trend model—and their ranking with respect to prediction and extrapolation quality under open- and closed-loop conditions are still lacking.

The Hammerstein/Wiener type nonlinear hybrid models are very promising in those cases where a fundamental steady-state real-time optimization model is already available. The work of Norquay et al. (1999) and Pearson and Pottmann (2000) are good starting points for further investigation which should aim at an extension of the concept to real-time dynamic optimization and control. The engineering effort can be reduced tremendously if the optimization model implementation can be directly integrated with a dynamic linear model in a modular fashion by a heterogeneous simulation and optimization platform (e.g. von Wedel and Marquardt (2000)).

Model Order Reduction

We have seen that model complexity can be reduced significantly by introducing empirical components in a model. Here, we will review order reduction techniques for a given nonlinear lumped parameter models which may be of whatever origin. We restrict the discussion to models of type

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}) \ , \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \ , \quad \boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) \ , \tag{8}$$

though more general differential-algebraic models should be considered. However, there seems to be only the paper of Löffler and Marquardt (1991), which treats this relevant class. We attempt to unify—at least to the extent possible—the great many variants of reported nonlinear order reduction techniques. Such an exercise will hopefully uncover hidden relationships and foster new developments in the future. Order reduction is always possible for large-scale models because the most significant contribution to the dynamics originates in a subspace of the complete state space. The key questions are what we rate as a significant contribution and how we are going to reveal it by systematic means ultimately under closed loop conditions.

Projection Methods

Projection methods have been suggested in a great variety in the recent literature. A generic procedure can be formulated as follows:

1. Transform the original state space into a state space better revealing the important contributions to process dynamics, i.e.

$$\boldsymbol{x} - \boldsymbol{x}^* = \boldsymbol{T}(\boldsymbol{z}) \ , \tag{9}$$

   with a general diffeomorphism $\boldsymbol{T}$ and the transformed state vector $\boldsymbol{z} \in R^n$. The reference state $\boldsymbol{x}^*$ is often a non-zero nominal operating point. Note that $\boldsymbol{T}$ is a non-singular square matrix in the linear case.

2. To achieve order reduction we decompose the transformed space into two complementary subspaces with state vectors $\boldsymbol{z}_1 \in R^m$ and $\boldsymbol{z}_2 \in R^{n-m}$, respectively. Hence,

$$\boldsymbol{T}(\boldsymbol{z}) = \boldsymbol{T}(\boldsymbol{z}_1, \boldsymbol{z}_2) \ , \tag{10}$$

   or

$$\boldsymbol{T}\boldsymbol{z} = \boldsymbol{T}_1\boldsymbol{z}_1 + \boldsymbol{T}_2\boldsymbol{z}_2 \ , \quad \boldsymbol{T} = [\boldsymbol{T}_1, \boldsymbol{T}_2] \ , \tag{11}$$

   in case of a linear transformation. We call $\boldsymbol{z}_1$ the dominant states and refer to $\boldsymbol{z}_2$ as the non-dominant states. Note, that the new states are linear combinations of the original states in the linear case.

3. Finally, we have to deduce a nonlinear dynamic model for the dominant states

$$\dot{\boldsymbol{z}}_1 = \boldsymbol{f}_1(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{u}) \tag{12}$$

   and a hopefully simple algebraic model for the non-dominant states

$$\boldsymbol{0} = \boldsymbol{f}_2(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{u}) \ . \tag{13}$$

4. Approximate states $\tilde{\boldsymbol{x}}$ and outputs $\tilde{\boldsymbol{y}}$ of the original system can be easily computed from $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ using Equation 9 with Equations 10 or 11. Equation 9 can be viewed as an output equation of the model (12), (13).

The variants of projection methods differ mainly in steps 1–3. The major techniques are introduced and put into perspective next.

Model transformation

Scherpen (1993) suggested nonlinear balancing as a tool for nonlinear model reduction. Her method generalizes balancing of linear systems as introduced by Moore (1981). The idea of balancing is the transformation of a system into an equivalent form which allows the assessment of the importance of the state variables with respect to the energy in its input and output signals. According to Scherpen (1993), we define the controllability and observability functions

$$\begin{gathered} L_c(\boldsymbol{x}_0) = \min_{u, x(0)=x_0} \frac{1}{2} \int_{-\infty}^{0} \boldsymbol{u}(t)^T \boldsymbol{u}(t) dt \ , \\ \boldsymbol{x}(-\infty) = \boldsymbol{0} \ , \quad \boldsymbol{u} \in L_2(-\infty, 0) \ , \end{gathered} \tag{14}$$

$$\begin{gathered} L_o(\boldsymbol{x}_0) = \frac{1}{2} \int_{0}^{\infty} \boldsymbol{y}(t)^T \boldsymbol{y}(t) dt \ , \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \ , \ \boldsymbol{u}(t) \equiv \boldsymbol{0} \ , \ 0 \le t < \infty \ , \end{gathered} \tag{15}$$

for linear and nonlinear systems. These functions denote the amount of input energy required to reach the state $\boldsymbol{x}_0$ and the amount of output energy generated by the state $\boldsymbol{x}_0$, respectively. For a linear system $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$, these functions are quadratic forms

$$L_c = \frac{1}{2} \boldsymbol{x}_0^T \boldsymbol{M}_c^{-1} \boldsymbol{x}_0 \ , \quad L_o = \frac{1}{2} \boldsymbol{x}_0^T \boldsymbol{M}_o^{-1} \boldsymbol{x}_0 \tag{16}$$

of the controllability and observability Gramians

$$\boldsymbol{M}_c = \int_0^\infty e^{\boldsymbol{A}t} \boldsymbol{B}\boldsymbol{B}^T e^{\boldsymbol{A}^T t} \, dt \ , \tag{17}$$

$$\boldsymbol{M}_o = \int_0^\infty e^{\boldsymbol{A}^T t} \boldsymbol{C}^T \boldsymbol{C} e^{\boldsymbol{A}t} \, dt \ , \tag{18}$$

which, for stable systems, can be computed from the Lyapunov equations

$$\boldsymbol{A}\boldsymbol{M}_c + \boldsymbol{M}_c \boldsymbol{A}^T = -\boldsymbol{B}\boldsymbol{B}^T \ , \tag{19}$$

$$\boldsymbol{A}^T \boldsymbol{M}_o + \boldsymbol{M}_o \boldsymbol{A} = -\boldsymbol{C}^T \boldsymbol{C} \ . \tag{20}$$

A system is called internally balanced (Moore, 1981) if

$$\boldsymbol{M}_o = \boldsymbol{M}_c = \boldsymbol{\Sigma} = \text{diag}\,\{\sigma_i\} \tag{21}$$

where $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_n$ are the Hankel singular values. There always exists an orthogonal matrix $\boldsymbol{T}$ (with $\boldsymbol{T}^{-1} = \boldsymbol{T}^T$) which transforms a (stable) linear system $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$[1] into its balanced equivalent $(\boldsymbol{T}^T\boldsymbol{A}\boldsymbol{T}, \boldsymbol{T}^T\boldsymbol{B}, \boldsymbol{C}\boldsymbol{T})$ with $\boldsymbol{x} = \boldsymbol{T}\boldsymbol{z}, \boldsymbol{z} = \boldsymbol{T}^T\boldsymbol{x}$ (Moore, 1981). More precisely, there are two transformation applied subsequently which employ the eigenvectors and eigenvalues of both Gramians to arrive at diagonalized controllability and observability Gramians of the transformed system.

The generalization to stable nonlinear systems (8) requires the determination of the nonlinear controllability and observability functions $L_c(\boldsymbol{x}_0)$ and $L_o(\boldsymbol{x}_0)$, respectively. Scherpen (1993) derives two nonlinear partial differential equations, a Lyapunov and a Hamilton-Jacobi type of equation, to determine these functions. Two nonlinear transformations can be derived from these solutions. If they are applied to $L_c(\boldsymbol{x}_0)$ and $L_o(\boldsymbol{x}_0)$ again, n singular value functions, the nonlinear analogs to the singular values in the linear case, can be identified. Hence, these transformations can be used to balance a nonlinear system. A similar technique based on a different definition of the energy of the input and output signals related to the nonlinear H_∞ control problem as well as extensions to nonlinear systems have been reported more recently (Scherpen and van der Schaft, 1994; Scherpen, 1996). The drawback of these approaches is obvious: the required analytical computations to determine $L_c(\boldsymbol{x}_0)$ and $L_o(\boldsymbol{x}_0)$ are rarely feasible in practice and a fully constructive method for the determination of the nonlinear transformation is not yet available. Hence, an approximation of the analytical balancing method of Scherpen seems to be the only way forward.

Consequently, Newman and Krishnaprasad (1998) explored the possibility of determining $L_c(\boldsymbol{x}_0)$ and $L_o(\boldsymbol{x}_0)$ by a Monte-Carlo technique for a two-dimensional problem.

An alternative approach to the computation of approximate Gramians can be deduced from a remark given by Moore (1981, p.21). He suggested to sample the impulse response matrix at a finite number of times to empirically construct an approximate $\boldsymbol{M}_c$ and $\boldsymbol{M}_o$ if solutions to the Lyapunov equations (19), (20) do not exist or are hard to compute.

Pallaske (1987) seems to be the first who made use of this idea in the context of model order reduction for nonlinear systems (without referring to Moore's original work). He introduced the *covariance matrix*

$$\boldsymbol{M} = \int_{\mathcal{G}} \int_0^\infty (\boldsymbol{x}(t) - \boldsymbol{x}^*)(\boldsymbol{x}(t) - \boldsymbol{x}^*)^T \, dt \, d\mathcal{G} \tag{22}$$

as the basis for the derivation of a linear transformation employing an orthogonal matrix $\boldsymbol{T}$. The symbol $\mathcal{G}$ denotes a set of representative trajectories resulting from a variation of initial conditions and input signals. Hence, the covariance matrix averages the behavior of the nonlinear system over a set of representative trajectories.

In order to compute the covariance matrix, the set $\mathcal{G}$ must be parameterized. Pallaske (1987) did not use impulse responses but suggested to keep the controls at constant values matching to the nominal operating point $\boldsymbol{x}^*$ and to parameterize the initial conditions by

$$\boldsymbol{x}_0 = \boldsymbol{x}^* + \boldsymbol{F}\boldsymbol{r} \tag{23}$$

with $\boldsymbol{F} \in R^{n \times s}$ and $\boldsymbol{r} \in R^s$ with $s \ll n$. Pallaske (1987) did not give any recommendations on suitable choices of $\boldsymbol{F}$. Löffler and Marquardt (1991) suggested to select $\boldsymbol{F}$ as the averaged steady-state gain matrix of the system in some region including $\boldsymbol{x}^*$. This suggestion results from a choice of $\mathcal{G}$ as a set of step responses at $\boldsymbol{x}^*, \boldsymbol{u}^*$. Other choices of trajectories $\mathcal{G}$ (or rather matrices $\boldsymbol{F}$) resulting in different covariance matrices $\boldsymbol{M}$ are obviously possible, if they are representative for the system dynamics in some region of the state space and for the intended use of the reduced model.

For nonlinear systems, the covariance matrix can be determined, in principle, by numerical quadrature of the multi-dimensional integral in Equation 22 after a suitable choice of $\boldsymbol{F}$. Highly accurate adaptive parallel algorithms are becoming available to solve this computationally demanding task even on heterogeneous workstation clusters (Čiegis et al., 1997). However, the number of simulations required scales with 2^n. These algorithms can therefore only be used for small to moderate n. Instead, Monte-Carlo techniques may be used to get a coarser approximation for large-scale problems (Jadach, 2000).

In many cases $\boldsymbol{M}$ has to be computed from a linear approximation to get acceptable results (Löffler and Marquardt, 1991). The covariance matrix can be easily determined from an algebraic Lyapunov equation for any parameterization (23). To show this property, we linearize (8) at the reference state $\boldsymbol{x}^*, \boldsymbol{u}^*$ and evaluate the integral in (22) for the set of trajectories $\mathcal{G}$ resulting from the linearized system after a variation of the initial conditions (23) with $\|\boldsymbol{r}\| < \rho$. The resulting covariance matrix $\boldsymbol{M}_l$ is then given by

$$\boldsymbol{M}_l = \int_0^\infty e^{\boldsymbol{A}t} \, \boldsymbol{S} \, e^{\boldsymbol{A}^T t} \, dt \tag{24}$$

with the Jacobian $\boldsymbol{A} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}|_{\boldsymbol{x}^*, \boldsymbol{u}^*}$ and

$$\boldsymbol{S} = k\boldsymbol{F}\boldsymbol{F}^T \,, \quad k = \frac{2\pi^{\frac{s}{2}} \rho^{\frac{s}{2}}}{s(s+2)\Gamma(\frac{s}{2})} \,. \tag{25}$$

The integration can be replaced by the solution of a Lyapunov equation. Löffler and Marquardt (1991) suggested

[1] See Skogestad and Postlethwaite (1996), pp. 464, for extensions to unstable systems.

to use $\boldsymbol{F} = -\boldsymbol{A}^{-1}\boldsymbol{B}$, the static gain at the reference point $\boldsymbol{x}^*, \boldsymbol{u}^*$ to emphasize the input-state relation of the dynamics. An averaged gain in some neighborhood of the reference point can be used instead, i.e. $\boldsymbol{F} = -\overline{\boldsymbol{A}^{-1}\boldsymbol{B}}$, to reflect some of the nonlinearity in the calculation of $\boldsymbol{M}$.

Motivated by the stiffness occurring in many large-scale systems, Pallaske (1987) suggests to choose the transformation such that the dynamics of the states can be approximately captured in a lower dimensional subspace. This translates to a minimization of the variance of the state in the directions of the coordinate axes of the reduced space. The transformation $\boldsymbol{T} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n]$, with $\boldsymbol{d}_i$ being the normalized eigenvectors of the covariance matrix $\boldsymbol{M}$, results in such a choice (see below).

The close relation of Pallaske's method to model reduction by balancing can be identified as follows. A choice of $\boldsymbol{F} = \boldsymbol{F}_c = \boldsymbol{B} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}|_{\boldsymbol{x}^*,\boldsymbol{u}^*}$ and $\boldsymbol{F} = \boldsymbol{F}_o = \boldsymbol{C}^T = (\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}|_{\boldsymbol{x}^*,\boldsymbol{u}^*})^T$ and ρ determined from (25) with $k = 1$ results in the local controllability or observability Gramian (17), (18) of the system (8) at $\boldsymbol{x}^*$ emphasizing the input-state or the state-output relation of the dynamics. These matrices are used in linear balancing (Moore, 1981) to construct the transformation $\boldsymbol{T}_{c,o}$. This transformation aims at a removal of the weakly controllable and observable subspaces. It is in contrast to Pallaske's objective which is a removal of the fast non-dominant states. Obviously, a transformation determined from a linearization of the nonlinear model does not exactly balance the nonlinear system in the sense of Scherpen (1993) but may qualify as a useful empirical approximation which at least is consistent with the linear theory. In fact, Wisnewski and Doyle III (1996a) successfully demonstrate the applicability of a related approach. They compute the transformation $\boldsymbol{T}$ from the left and right eigenvectors of the Hankel matrix, i.e. the product of the observability and controllability Gramians, of a linearization of the nonlinear model at a stationary reference point.

Empirical balancing of nonlinear systems has been recently introduced by Lall et al. (1999). They suggest empirical controllability and observability Gramians, $\boldsymbol{M}_c$ and $\boldsymbol{M}_o$, which are closely related to Equation 22. Impulse responses of varying magnitude are chosen in the set $\mathcal{G}$ to compute $\boldsymbol{M}_c$, whereas responses to different initial conditions are used in the set $\mathcal{G}$ to compute $\boldsymbol{M}_o$. In this case, Lall et al. (1999) use the covariances of the outputs $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ in Equation 22 instead of the states $\boldsymbol{x}$. These Gramians are used to empirically balance the nonlinear system as in the linear case. This approach has been adopted recently by Hahn and Edgar (Hahn and Edgar, 1999, 2000).

A completely different approach of determining the transformation matrix is reported in structural dynamics (Slaats et al., 1995). As in modal reduction techniques for linear systems (Litz, 1979; Bonvin and Mellichamp, 1982), the transformation matrix $\boldsymbol{T}$ is formed by the dominating modes of the second order model (the eigenvectors associated with complex conjugate eigenvalues). However, these modes are taken as functions of the displacement of a node in a mechanical structure to account for the nonlinearities. Analytical expressions are derived to determine the modes for the model linearized at the initial condition and for first and second order sensitivities of the modes with respect to the nodal positions. The transformation matrix is then computed from these quantities. The method carries over to first order process systems models, but seems only appropriate for less severe nonlinearities in the vicinity of an operating point.

State space decomposition

According to step 2 of the general projection method, the transformed space has to be decomposed next into two subspaces capturing the dominant and the non-dominant states, respectively.

Scherpen (1993) suggests to use the magnitude of the singular value functions occuring in the transformed observability and controllability functions in some domain of the state space as an indication for weakly controllable and observable subspaces. Those states with large values of the singular value functions are grouped into the vector of dominant states $\boldsymbol{z}_1$, and the remaining state variables form $\boldsymbol{z}_2$. The same strategy is employed by Lall et al. (1999) and Hahn and Edgar (1999, 2000). They analyse the singular values of the empirical Gramians and delete those states with small singular values indicating weak observability and controllablility as in the linear case (Moore, 1981).

Pallaske (1987) poses an optimization problem to bound the normalized L_2-error between the approximate and the original state vectors $\tilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ to a user-defined tolerance ε_0 by varying m, the dimension of $\boldsymbol{z}_1$. The solution to this problem is

$$m = \min k \tag{26}$$

subject to

$$\sum_{i=1}^{k} \mu_i \geq (1 - \varepsilon_0^2)\, \mathrm{trace}\,\{\boldsymbol{M}\}$$

with $\mu_1 > \mu_2 \cdots \geq \mu_n$ being the eigenvalues of $\boldsymbol{M}$. Hence, the first m normalized eigenvectors $\boldsymbol{d}_i\,, i = 1, \ldots n$ of $\boldsymbol{M}$ span the subspace for the dominant transformed states. Consequently, $\boldsymbol{T}_1 = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_m]$ and $\boldsymbol{T}_2 = [\boldsymbol{d}_{m+1}, \ldots, \boldsymbol{d}_n]$ in Equation 11. The same approach has been adopted by Löffler and Marquardt (1991).

In some cases, the choice of the dominant states may be based solely on physical insight. This selection is typically done without transformation in the original coordinates.

Formulation of the reduced model

There are different strategies of determining the reduced model equations after the dominant states have been identified. In complete analogy to the linear case, Scherpen (1993, 1996) suggests to simplify the balanced nonlinear model by *truncation* of the balanced state. The non-dominant states $\boldsymbol{z}_2$ are equated to zero, accounting for their negligible influence on the input-output behavior of the system. Since the fully nonlinear balancing method is not applicable in practice, we present this approach in more detail for a linear transformation (11) resulting for example from empirical nonlinear balancing or from Pallaske's method.

After transformation of (8) and subsequent decomposition into two subsystems, we obtain

$$\dot{\boldsymbol{z}}_1 = \boldsymbol{T}_1^T \boldsymbol{f}(\boldsymbol{x}^* + \boldsymbol{T}_1 \boldsymbol{z}_1 + \boldsymbol{T}_2 \boldsymbol{z}_2, \boldsymbol{u}) \;, \tag{27}$$

$$\dot{\boldsymbol{z}}_2 = \boldsymbol{T}_2^T \boldsymbol{f}(\boldsymbol{x}^* + \boldsymbol{T}_1 \boldsymbol{z}_1 + \boldsymbol{T}_2 \boldsymbol{z}_2, \boldsymbol{u}) \;, \tag{28}$$

$$\boldsymbol{z}_1(0) = \boldsymbol{T}_1^T (\boldsymbol{x}_0 - \boldsymbol{x}^*) \;, \tag{29}$$

$$\boldsymbol{z}_2(0) = \boldsymbol{T}_2^T (\boldsymbol{x}_0 - \boldsymbol{x}^*) \tag{30}$$

Truncation of the transformed state yields

$$\dot{\tilde{\boldsymbol{z}}}_1 = \boldsymbol{T}_1^T \boldsymbol{f}(\boldsymbol{x}^* + \boldsymbol{T}_1 \tilde{\boldsymbol{z}}_1, \boldsymbol{u}) \;, \tag{31}$$

$$\tilde{\boldsymbol{z}}_2 = \boldsymbol{0} \;, \tag{32}$$

a reduced model of order $m < n$. Alternatively, we may assume

$$\dot{\tilde{\boldsymbol{z}}}_2 = \boldsymbol{0} = \boldsymbol{T}_2^T \boldsymbol{f}(\boldsymbol{x}^* + \boldsymbol{T}_1 \tilde{\boldsymbol{z}}_1 + \boldsymbol{T}_2 \tilde{\boldsymbol{z}}_2, \boldsymbol{u}) \tag{33}$$

to form the reduced model (27), (33). This concept, often referred to as *residualization*, is closely related to singular perturbation discussed in more detail below. It results in a differential-algebraic system of the same order as the original model which is still difficult to solve in general. However, this way the (small) contribution of $\boldsymbol{z}_2$ to $\tilde{\boldsymbol{x}}$ is captured at least to some extent. Combinations of truncation and residualization have been suggested in the linear case (Liu and Anderson, 1989) and obviously carry over to the nonlinear case. Truncation as well as residualization strategies have been suggested and successfully applied to nonlinear systems by Pallaske (1987), Löffler and Marquardt (1991) as well as by Hahn and Edgar (1999, 2000).

Approximation of original states

An approximation of the original state $\tilde{\boldsymbol{x}}$ is obtained from

$$\tilde{\boldsymbol{x}} = \boldsymbol{x}^* + \boldsymbol{T}_1 \tilde{\boldsymbol{z}}_1 + \boldsymbol{T}_2 \tilde{\boldsymbol{z}}_2 \tag{34}$$

according to Equations 9 and 11. The second term vanishes identically in case of truncation. Steady-state accuracy cannot be guaranteed in the general nonlinear case.

Proper Orthogonal Decomposition

Proper orthogonal decomposition (POD), often also called Karhunen-Loeve expansion or method of empirical eigenfunctions (Fukunaga, 1990; Holmes et al., 1996), is somehow related to the projection methods discussed above. This method has gained much attention in fluid dynamics in the context of discovering coherent structures in turbulent flow patterns. Later, the method has been worked out for the construction of low-order models for dynamic (usually distributed parameter) systems with emphasis on fluid dynamical problems (Sirovich, 1987; Aubry et al., 1988; Holmes et al., 1996). POD has found many applications in simulation and optimization of reactive and fluid dynamical systems (e.g. Graham and Kevrekidis, 1996; Kunisch and Volkwein, 1999; Afanasiev and Hinze, 1999) or chemical vapor deposition processes (e.g. Banerjee and Arkun, 1998; Baker and Christofides, 1999). The method comes in a number of variants. We will summarize one of them following the presentation of Ravindran (1999) next.

Assume we have a representative trajectory of (8) for a certain initial condition $\boldsymbol{x}_0$ and control $\boldsymbol{u}(t)$ defined on a finite time interval $[t_0, t_1]$. The trajectory is uniformly sampled for simplicity to form the ensemble $\mathcal{S} = \{\boldsymbol{x}(t_k) - \boldsymbol{x}^*\}_{k=1}^p = \{\Delta\boldsymbol{x}(t_k)\}_{k=1}^p$ containing p data sets of length n which are often called *snapshots*. As before, $\boldsymbol{x}^*$ is the reference which can be either a steady-state or the ensemble average of the snapshots. We are interested in a unit vector $\boldsymbol{d}$ which is in some sense close to the snapshots in $\mathcal{S}$. We may request that $\boldsymbol{d}$ is as parallel as possible to all the snapshots. This requirement leads to the optimization problem

$$\max \frac{1}{p} \sum_{k=1}^{p} \frac{(\Delta\boldsymbol{x}(t_k)^T \boldsymbol{d})^2}{\boldsymbol{d}^T \boldsymbol{d}}. \tag{35}$$

We assume $\boldsymbol{d}$ to be a linear combination of the data, i.e.

$$\boldsymbol{d} = \sum_{k=1}^{p} w_k \Delta\boldsymbol{x}(t_k) \;, \tag{36}$$

and determine the weights w_k to solve the optimization problem (35). Solving this optimization problem is the same as finding the eigenvectors of the *correlation matrix* $\boldsymbol{N}$ with elements

$$N_{i,j} = \Delta\boldsymbol{x}(t_i)^T \Delta\boldsymbol{x}(t_j) \;. \tag{37}$$

Since this matrix is nonnegative Hermitian, it has a complete set of orthogonal eigenvectors $\{\boldsymbol{w}_1, \ldots \boldsymbol{w}_p\}$ along with a set of eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$. We can now construct an orthogonal basis $\mathrm{span}\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p\}$ by means of (36) with

$$\boldsymbol{d}_i = \frac{1}{\sqrt{\lambda_i}} \sum_{k=1}^{p} w_{i,k} \, \Delta\boldsymbol{x}(t_k); \,, \quad i = 1, \ldots p \tag{38}$$

where $w_{i,k}$ denote the elements of the eigenvector $\boldsymbol{w}_i$. It can be shown that any approximation of $\boldsymbol{x}(t_k)$ in a subspace spanned by the first $p_1 < p$ basis vectors $\boldsymbol{d}_i$ maximizes the captured energy $\boldsymbol{x}(t_k)^T\boldsymbol{x}(t_k)$ of the data set. Due to this property, we may just use a reduced basis span$\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{p_1}\}$ with $p_1 \ll p$ to obtain sufficient approximation quality. The value of p_1 is determined after some experimentation. The ratio

$$\kappa = \frac{\sum_{k=1}^{p_1} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \tag{39}$$

indicates the percentage of energy contained in the first p_1 basis vectors. Obviously this ratio should be close to unity. Note, that we therefore do not have to match the number of snapshots (or basis vectors) p to the dimension n of the dynamic system. Often, we want to use $p \ll n$ for convenience, if very large-scale systems are considered, which, for example, may arise after discretization of a distributed parameter system.

There are at least two common ways of determining the basis vectors. Banerjee and Arkun (1998) employ a singular value decomposition and construct the basis from the left singular vectors of $\boldsymbol{N}$. An alternative approach does not rely on the correlation matrix $\boldsymbol{N}$ but on the $n \times p$ *snapshot matrix*

$$\boldsymbol{X} = [\Delta\boldsymbol{x}(t_1), \Delta\boldsymbol{x}(t_2), \ldots \Delta\boldsymbol{x}(t_p)] \tag{40}$$

the columns of which are the snapshots $\Delta\boldsymbol{x}(t_k)$ at t_k (Aling et al., 1996; Shvartsman and Kevrekidis, 1998). Again, the basis is formed by those $p_1 \ll p$ left singular vectors of $\boldsymbol{X}$ which are associated with the largest singular values and hence capture most of the energy in the data set.

The basis constructed from the correlation or snapshot matrices gives rise to a representation of an approximate solution $\tilde{\boldsymbol{x}}$ to (8) by a linear combination of the basis vectors. Hence,

$$\tilde{\boldsymbol{x}} - \boldsymbol{x}^* = \sum_{k=1}^{p_1} a_k \, \boldsymbol{d}_k \; . \tag{41}$$

If the expansion coefficients a_k and the basis vectors $\boldsymbol{d}_k$ are collected in a vector $\boldsymbol{a}_1 = [a_1, a_2, \ldots a_{p_1}] \in R^{p_1}$ and a matrix $\boldsymbol{U}_1 = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots \boldsymbol{d}_{p_1}] \in R^{n \times p_1}$ we can rewrite this equation as

$$\tilde{\boldsymbol{x}} = \boldsymbol{x}^* + \boldsymbol{U}_1\boldsymbol{a}_1 \tag{42}$$

which has the same structure as Equation 34 in case of truncating the non-dominant states. The reduced model is

$$\dot{\boldsymbol{a}}_1 = \boldsymbol{U}_1^T \boldsymbol{f}(\boldsymbol{x}^* + \boldsymbol{U}_1\boldsymbol{a}_1, \boldsymbol{u}) \; , \tag{43}$$

$$\boldsymbol{a}_1(0) = \boldsymbol{U}_1^T(\boldsymbol{x}_0 - \boldsymbol{x}^*) \; , \tag{44}$$

which has exactly the same appearance as the truncated model (31) resulting from model reduction by projection.

As in the projection methods, truncation is not the only possibility of developing a reduced model. Rather, the full basis spanned by $p < n$ vectors can be employed by summing to p instead to p_1 in the approximation (41). This approach would result in a model structure completely analogous to the system (27), (33). Residualization (e.g. setting $\dot{\boldsymbol{a}}_2$ to zero) or more sophisticated *slaving methods* can be used to reduce the computational effort (e.g. Aling et al., 1997; Shvartsman and Kevrekidis, 1998; Baker and Christofides, 2000). A more detailed discussion will be provided in the next section.

The major difference between POD and projection is the lack of a convergence proof which would guarantee the reduced model to match the original model in case the number of basis functions p_1 approaches the state dimension n. Further, there is no diffeomorphism between the original state space and the space spanned by the empirical eigenvectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p$ in POD which is available in the projection methods discussed above. Otherwise, both types of methods are very similar as they heavily rely on data taken from a series of simulations of the original model (8). However, the data is organized and used differently depending on the intention of model reduction. Here, the basis is constructed from the correlation or snapshot matrix $\boldsymbol{N}$ or $\boldsymbol{X}$ (cf. (37) and (40)) whereas the covariance matrix $\boldsymbol{M}$ (cf. (22)) and its specializations are used in the projection methods.

Slaving

Residualization in projection results in a set of differential-algebraic equations (cf. Equations 27, 33). A similar system is obtained in POD, if all basis vectors are employed but only the first p_1 are used to build the dynamic subsystem. If the algebraic subsystem (cf. (33) in projection method) cannot be solved explicitly, the computational effort cannot be reduced and the model reduction largely fails. Truncation could be employed instead, but a significant loss in approximation accuracy would result inevitably.

A reduction of the computational effort is possible, if we could find an explicit relation between the algebraic and the dynamic variables ($\boldsymbol{z}_2$ and $\boldsymbol{z}_1$ in projection or $\boldsymbol{a}_2$ and $\boldsymbol{a}_1$ in POD methods). Hence, we are looking for a function

$$\boldsymbol{z}_2 = \boldsymbol{\sigma}(\boldsymbol{z}_1) \tag{45}$$

in case of residualization in projection methods, or for equivalent functions in case of residualization in POD. This approach has also been called *slaving* in the recent literature. Its roots are in nonlinear dynamics, where the computation of *approximate inertial manifolds* has some tradition (e.g. Foias and Témam, 1988; Foias et al., 1988). Aling et al. (1997) as well as Shvartsman and Kevrekidis (1998) use this idea in their case studies

on nonlinear model reduction based on POD. Here, we present the concept in the context of projection methods, where it does not seem to have been applied yet.

The algebraic equation (33) forms the starting point of defining the family of maps

$$\begin{aligned}\tilde{\boldsymbol{z}}_2^{(k+1)} &= \boldsymbol{\sigma}_k(\boldsymbol{z}_1, \tilde{\boldsymbol{z}}_2^{(k)}) \\ &=: \tilde{\boldsymbol{z}}_2^{(k)} - \boldsymbol{T}_2^T \boldsymbol{f}(\boldsymbol{x}^* + \boldsymbol{T}_1\boldsymbol{z}_1 + \boldsymbol{T}_2\tilde{\boldsymbol{z}}_2^{(k)}, \boldsymbol{u}) \; .\end{aligned} \tag{46}$$

If the map is a contraction, Equation 33 can be solved iteratively from the initial guess $\tilde{\boldsymbol{z}}_2^{(0)} = \boldsymbol{0}$ which would be used in truncation. The computational effort can be reduced significantly, if only few iterations are carried out to improve on the initial value. In case of two iterations we find for example

$$\tilde{\boldsymbol{z}}_2 = \boldsymbol{\sigma}_2(\boldsymbol{z}_1, \boldsymbol{\sigma}_1) \; , \quad \boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_1(\boldsymbol{z}_1, \boldsymbol{0}) \; . \tag{47}$$

This approximation can now be used to eliminate $\boldsymbol{z}_2 \approx \tilde{\boldsymbol{z}}_2$ in Equation 27 and to compute $\tilde{\boldsymbol{x}}$ from Equation 34. An accuracy similar to residualization can be achieved by this method but only a model of the same order as in truncation has to be solved.

Equation Residual Minimization Methods

Equation residual minimization methods have been extensively studied in the linear case (e.g. Eitelberg, 1982). A nice generalization to the nonlinear case has been given recently by Lohmann (1994, 1995). He defines a nonlinear reduced model[2] of (8) as

$$\dot{\boldsymbol{z}}_1 = \boldsymbol{V}\boldsymbol{f}(\boldsymbol{x}^* + \boldsymbol{W}\boldsymbol{z}_1, \boldsymbol{u}) \; , \tag{48}$$

to compute the approximate states

$$\tilde{\boldsymbol{x}} = \boldsymbol{x}^* + \boldsymbol{W}\boldsymbol{z}_1 \; . \tag{49}$$

This reduced model has the same structure as that resulting from a truncated projection method (cf. Equation 31) or a truncated POD method (cf. Equation 43). The matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ are, however, determined differently. Their elements are the decision variables in a parameter optimization problem which minimizes the residuals of Equations 48, 49. First, $\boldsymbol{W}$ is determined from minimizing the sum of weighted errors

$$\sum_{k=1}^{p} q_{1,k} \left\| \boldsymbol{x}(t_k) - \boldsymbol{x}^* - \boldsymbol{W}\boldsymbol{z}_1(t_k) \right\|^2 \tag{50}$$

using a number of representative trajectories sampled at discrete times t_k. A set of carefully chosen step responses is chosen for this purpose. Next, given $\boldsymbol{W}$, the sum of weighted equation residuals

$$\sum_{k=1}^{p} q_{2,k} \left\| \dot{\tilde{\boldsymbol{z}}}_1(t_i) - \boldsymbol{V}\boldsymbol{f}(\boldsymbol{x}^* + \boldsymbol{W}\tilde{\boldsymbol{z}}_1(t_i), \boldsymbol{u}(t_i)) \right\|^2 \tag{51}$$

is minimized to fix the elements of $\boldsymbol{V}$. In contrast to other nonlinear model reduction techniques, steady-state accuracy can be guaranteed by incorporating a steady-state condition as an equality constraint in this residual minimization problem.

The choice of the dominant states $\boldsymbol{z}_1$ can be done on physical insight in the most simple case. However, Lohmann (1994, 1995) also suggests a systematic alternative to the selection of dominant states which usually gives better approximation results. He introduces a transformation of the state space which reveals the dominant modes which is similar to both, the balancing transformation of Moore (1981) and to the transformation of Pallaske (1987). The transformation is applied first, the dominant states are identified, and the residual minimization technique is applied to the transformed system.

A modified version of the Lohmann method has been reported recently by Kordt (1999). It applies to systems where only the dominant states of the original model are entering the nonlinear terms of the system equations whereas the non-dominant states are confined to the linear part. An even more special case has been treated earlier by Hasenjäger (1991). He assumes that nonlinearities occur only in the equations of the dominant states and that these nonlinearities only depend on the dominant states. In this case, it is suggested to first neglect the nonlinear terms, reduce the linear part of the model by any of the linear reduction techniques (e.g. Litz, 1979; Moore, 1981; Eitelberg, 1982; Bonvin and Mellichamp, 1982; Glover, 1984; Samar et al., 1995; Muscato, 2000), and then add the nonlinear terms to the reduced linear model equations.

Perturbation Methods

In many cases, chemical processes are characterized by phenomena on separate time scales. Multiple time scales occur due to orders of magnitude differences in the densities of contacting vapor (or gas) and liquid phases in multi-phase processes, in the thermal capacitances of reactors, in the time constant of chemical reactions, or in the transfer rates of material or energy across phase boundaries. The models (8) describing such multiple time scale systems usually incorporate process parameters varying in a wide range. In the simplest case, where there are only two time scales present, we may identify a small parameter $\epsilon \ll 1$ in (8) to result in

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \epsilon) \; . \tag{52}$$

Let us assume that this equation can be reformulated as

$$\dot{\boldsymbol{x}}_s = \boldsymbol{g}_s(\boldsymbol{x}_s, \boldsymbol{x}_f, \boldsymbol{u}, \epsilon) \; , \tag{53}$$

$$\epsilon\dot{\boldsymbol{x}}_f = \boldsymbol{g}_f(\boldsymbol{x}_s, \boldsymbol{x}_f, \boldsymbol{u}, \epsilon) \; . \tag{54}$$

Here, $\boldsymbol{x}_s \in R^{n_s}, \boldsymbol{x}_f \in R^{n-n_s}$ are denoting the so-called slow and fast state variables, respectively. Obviously,

[2] Lohmann introduces a refined parameterization of the reduced model by splitting the right hand sides into linear and nonlinear terms.

this reformulation is identical to the partitioning of the state vector $\boldsymbol{x}$ into $\boldsymbol{x}_s$ and $\boldsymbol{x}_f$ by some strategy. This system representation is called the *standard form of a singularly perturbed system* (Kokotovic et al., 1986). This model comprises two time scales t and $\tau = \frac{t-t_0}{\epsilon}$ of different magnitude and is therefore called a two-time-scale model.

Basic singular perturbation approach

Since ϵ is a small parameter, the solution of (53), (54) can be determined by means of perturbation methods in the limit of $\epsilon \to 0$. It is given as

$$\boldsymbol{x}_s = \tilde{\boldsymbol{x}}_s + O(\epsilon) \ , \quad \boldsymbol{x}_f = \tilde{\boldsymbol{x}}_f + \boldsymbol{\mu} + O(\epsilon) \tag{55}$$

with $\tilde{\boldsymbol{x}}_s, \tilde{\boldsymbol{x}}_f$ computed from (53), (54) for $\epsilon = 0$. The notation $O(\epsilon)$ is used in the usual sense and refers to small terms of order ϵ. The so-called boundary layer correction $\boldsymbol{\mu}$ can be computed from

$$\frac{d\boldsymbol{\mu}}{d\tau} = \boldsymbol{g}_f(\tilde{\boldsymbol{x}}_s, \tilde{\boldsymbol{x}}_f + \boldsymbol{\mu}, \boldsymbol{u}, 0) \ . \tag{56}$$

It quickly dies out in a short time interval at the beginning of a transient. For an asymptotically stable system (56), the boundary layer correction may be neglected together with the $O(\epsilon)$ contributions to the solution. These considerations result in the simplified differential-algebraic model

$$\dot{\tilde{\boldsymbol{x}}}_s = \boldsymbol{g}_s(\tilde{\boldsymbol{x}}_s, \tilde{\boldsymbol{x}}_f, \boldsymbol{u}, 0) \ , \tag{57}$$

$$\boldsymbol{0} = \boldsymbol{g}_f(\tilde{\boldsymbol{x}}_s, \tilde{\boldsymbol{x}}_f, \boldsymbol{u}, 0) \tag{58}$$

to determine approximations $\tilde{\boldsymbol{x}} = [\tilde{\boldsymbol{x}}_s^T, \tilde{\boldsymbol{x}}_f^T]^T$ of the state $\boldsymbol{x}$. This approximation is often called the quasi-steady-state approximation (QSSA) of (8). Further simplification is possible, if the nonlinear algebraic equations can be solved analytically for $\tilde{\boldsymbol{x}}_f$, i.e.

$$\tilde{\boldsymbol{x}}_f = \boldsymbol{g}_f^{-1}(\tilde{\boldsymbol{x}}_s, \boldsymbol{u}, 0) \ . \tag{59}$$

This case can be interpreted as an exact slaving approach (cf. Equation 45).

If this reduction method is applied to a process model in practice, two issues have to be addressed. First, we have to derive the standard singularly perturbed system (53),(54). This boils down to the determination of n_s and to a proper association of the states $x_i, i = 1 \dots n$, to the vectors $\boldsymbol{x}_s$ and $\boldsymbol{x}_f$. Second, since the QSSA does often not lead to models of sufficient accuracy, corrections to the QSSA solutions are of particular interest. A number of suggestions found in the literature are discussed in the following paragraphs.

Standard singularly perturbed systems and QSSA

Often, the QSSA is derived heuristically based on a thorough understanding of the underlying physics. It is well known, however, that such an approach does not always work out well (e.g. Tatrai et al., 1994). More rigorous techniques are therefore required. The number of slow states n_s is usually determined from an analysis of the eigenvalues of the linearized system (8) along some representative trajectories. Robertson and Cameron (1997b) relate the separation ratio denoting the distance of two separated clusters of eigenvalues to the model reduction error. Their analysis gives quantitative recommendations for determining n_s and the eigenvalues in the cluster to be retained in a reduced model. The same authors present a homotopy-continuation technique to identify the eigenvalue-state association to determine the set of states concatenated into $\boldsymbol{x}_s$ and $\boldsymbol{x}_f$, respectively. Both techniques are demonstrated on various non-trivial examples, where the fast or slow states could not be identified merely on the basis of physical insight (Robertson and Cameron, 1997a,b; Tatrai et al., 1994).

Duchene and Rouchon (1996) demonstrate on a simple example from reaction kinetics that the QSSA is not coordinate-free, because the method may not lead to useful results in the original state space. These and other authors suggest a simple linear transformation for special reaction systems (e.g. van Breusegem and Bastin, 1991) and for binary distillation columns (Levine and Rouchon, 1991) which lead to more favorable coordinates for a QSSA. The linear transformation suggested by Pallaske (1987) and introduced above could qualify for the general case. By construction, the dominant states $\boldsymbol{z}_1$ represent the dominant (and often slow) states in the transformed state space, whereas $\boldsymbol{z}_2$ denote the non-dominant (often fast) states $\boldsymbol{z}_2$. Residualization in projection methods or POD discussed above is identical with a QSSA in transformed coordinates.

In general, however, the change of coordinates is nonlinear. Existence of a nonlinear map transforming any nonlinear system (8) into standard singularly perturbed form (53),(54) has been investigated by Marino and Kokotovic (1988). These authors give conditions, which assure the two-time-scale property of (52). They also provide general criteria, which guarantee the existence of an ϵ-independent diffeormorphism to transform the two-time-scale system (52) into the standard form (53),(54). A procedure for constructing such a diffeomorphism is given. It reveals a set of integrability conditions the transformation has to suffice. A more refined analysis has been developed more recently by Krishnan and McClamroch (1994) which has been adopted and extended by Kumar et al. (1998). These authors study systems (8) which are affine in the control variables $\boldsymbol{u}$ and in integer powers of a large parameter $\frac{1}{\epsilon}$. Krishnan and McClamroch (1994) give properties sufficient for the system to reveal two-time-scale characteristics and characterize the slow and fast dynamics. They also generalize their results to systems with more than one large parameter. The analysis of Kumar et al. (1998) reveals two distinct cases depending on the properties of the model nonlin-

earities. Both, the transformation as well as the region of the state-space in which the system shows two-time-scale behavior, may either depend on ϵ or not. Again, the nonlinear transformation has to satisfy the set of integrability conditions already identified by Marino and Kokotovic (1988). The analytical determination of the transformation may be restricted to special nonlinear systems of low to moderate order. It is often difficult to obtain.

The methods of Marino and Kokotovic (1988), Krishnan and McClamroch (1994) as well as of Kumar et al. (1998) require the identification of a large parameter. The quality of the resulting reduced model will crucially depend on this choice. While this parameter can often be found based on physical insight, it would be advantageous to have a transformation available which does not require such a maybe arbitrary choice. Nonlinear balancing as introduced by Scherpen (1993, 1996) may qualify as such a transformation at least in those cases where the fast states coincide with the weakly observable and controllable states. The computation of the transformation is, however, even more involved than that suggested by Kumar et al. (1998). Empirical nonlinear balancing (Lall et al., 1999; Hahn and Edgar, 2000) could be used instead, sacrificing however the nonlinear transformation in favor of a linear transformation similar to the approach of Pallaske (1987). The relation between singular perturbation and nonlinear balancing can only be conjectured at this point. The generalization of the linear result of Liu and Anderson (1989) to the nonlinear case is yet an open problem.

Coordinate-free perturbation methods

The quasi steady-state approximation is widely employed in original coordinates (e.g. Robertson and Cameron, 1997a,b) or in heuristically introduced transformed coordinates (e.g. Levine and Rouchon, 1991; van Breusegem and Bastin, 1991; Kumar et al., 1998). In both cases the approximation quality may be limited due to coordinates which are not appropriately revealing the time-scale separation of the fast and slow variables. Therefore, coordinate-free perturbation methods are attractive which do not rely on coordinate transformation but still come up systematically with satisfactory reduced models of type (57),(58).

There are various coordinate-free model reduction methods which not only result in reasonable approximate models but which go beyond the accuracy of the QSSA.

A first coordinate free method is for example reported by Genyuan and Rabitz (1996). They improve on the QSSA by expanding the fast variables in the regular perturbation series $\tilde{\boldsymbol{x}}_f = \tilde{\boldsymbol{x}}_f^{(0)} + \epsilon\tilde{\boldsymbol{x}}_f^{(1)} + \epsilon^2\tilde{\boldsymbol{x}}_f^{(2)} + \ldots$ where the functions $\tilde{\boldsymbol{x}}_f^{(k)}$ only depend on $\tilde{\boldsymbol{x}}_s$. The equations determining these functions follow from a regular perturbation method. The method is computationally attractive in those cases, where the fast equations are linear in the fast states.

A series of methods has been based on a geometrical interpretation of the global nonlinear system dynamics. In fact, they approximate the trajectories of the original system by trajectories on an *attractive invariant manifold* $\mathcal{M}$. A manifold $\mathcal{M}$ is invariant with respect to the vector field $\boldsymbol{f}$ in Equation 52 if $\boldsymbol{f}$ is tangent to $\mathcal{M}$. It is (locally) attractive, if any trajectory (starting close to $\mathcal{M}$) tends to $\mathcal{M}$ as $t \to \infty$. Model reduction means restriction of the system dynamics to this manifold. The key problem is to (at least approximately) obtain a set of equations defining $\mathcal{M}$.

Duchene and Rouchon (1996) present a solution to this problem. Their reduced model can be written as

$$\dot{\tilde{\boldsymbol{x}}}_s = \boldsymbol{C}(\tilde{\boldsymbol{x}}_s, \boldsymbol{\nu})\, \boldsymbol{g}_s(\tilde{\boldsymbol{x}}_s, \boldsymbol{\nu}) \;, \tag{60}$$

$$\boldsymbol{0} = \boldsymbol{g}_s(\tilde{\boldsymbol{x}}_s, \boldsymbol{\nu}) \;, \tag{61}$$

employing the auxiliary variables $\boldsymbol{\nu}$ provided the decomposition of the space into a fast and a slow subspace can be determined beforehand (for example from physical considerations or with the method of Robertson and Cameron (1997a,b). Duchene and Rouchon provide an explicit formula for the symbolic computation of the matrix $\boldsymbol{C}$ which comprises derivatives of the vector fields $\boldsymbol{g}_s$ and $\boldsymbol{g}_f$ with respect to the slow and fast variables. An explicit expression is also provided to determine approximations to the fast states as a function of $\tilde{\boldsymbol{x}}_s$ and $\boldsymbol{\nu}$ if they are of interest. As with all the perturbation techniques, model reduction is most effective if the auxiliary variables can be eliminated symbolically in the differential equations. Otherwise, slaving technique may be used additionally.

Duchene and Rouchon (1996) state that their method is completely equivalent to a method reported earlier by Maas and Pope (1992) if the system truly admits two time-scales. The method of Maas and Pope is also based on the idea of computing the manifold $\mathcal{M}$. However, instead of deriving a system of equations for approximation of the dynamics on $\mathcal{M}$ as Duchene and Rouchon (1996), their algorithm only determines a series of points to approximate $\mathcal{M}$ itself. These data points have to satisfy a set of n_f constraints. Rhodes et al. (1999) have suggested just recently to use this data and employ black-box identification to relate the fast states $\boldsymbol{x}_f$ to the slow states $\boldsymbol{x}_s$ by some explicit nonlinear function

$$\boldsymbol{x}_f = \boldsymbol{\sigma}(\boldsymbol{x}_s) \;. \tag{62}$$

This way a reduced order model of dimension n_s can be obtained without the need of deriving a singularly perturbed system in standard form and regardless whether the fast subsystem can be solved explicitly. Note that this technique is completely equivalent to the idea of slaving as employed in the context of POD (e.g. Aling et al., 1997; Shvartsman and Kevrekidis, 1998). Instead of the

Maas and Pope algorithm, the computational method reported by Davis and Skodje (1999) could be used to generate data points of an approximation to $\mathcal{M}$ which is then used to build the correlation (62) as suggested by Rhodes et al. (1999).

Obviously, this technique can be applied to any other singular perturbation or projection method. The algebraic equation (58) of the QSSA or (61) can be sampled for given values of $\boldsymbol{x}_s$. This data set can then be used to determine an expression (62) which can be used to eliminate the fast states $\boldsymbol{x}_f$ in (58) or the auxiliary variables $\boldsymbol{\nu}$ in (60).

It should be noted that all the methods discussed in this section require the partitioning of the original state vector $\boldsymbol{x}$ into fast and slow variables $\boldsymbol{x}_f$ and $\boldsymbol{x}_s$. The method reported by Robertson and Cameron (1997b)—though cumbersome—seems to be most suitable for this purpose.

Remarks on Distributed Parameter Systems

So far, distributed parameter systems have been represented by a model of type (8) employing either averaging over a spatial domain as part of the modeling procedure or by discretizing the spatial coordinates of a partial differential equation (PDE) model. Hence, the infinite-dimensional model has been first reduced to some potentially high order model (8) which then may be subject to model order reduction as reviewed above. Alternatively, order reduction could directly be applied to the infinite-dimensional PDE model to avoid often heuristic finite-dimensional approximate modeling. There is a lot of literature dealing with this problem which would justify a review in its own. Only a few references are given here as a starting point for the interested reader.

The rigorous model reduction approaches for nonlinear PDE models include Galerkin projection involving empirical (e.g. Holmes et al., 1996) or modal eigenfunctions (e.g. Armaou and Christofides, 2000) as well as weighted residuals methods of various kinds (e.g. Villadsen and Michelsen, 1978; Cho and Joseph, 1983; Stewart et al., 1985; Tali-Maamar et al., 1994). Better prediction quality can usually be obtained if the truncated contributions in the series expansion are captured by the approximate inertial manifold (e.g. Christofides and Daoutidis, 1997; Shvartsman and Kevrekidis, 1998; Armaou and Christofides, 2000). These techniques are closely related to projection and proper orthogonal decomposition as discussed above. Singular perturbation techniques have also been applied directly to PDE models. For example, Dochain and Bouaziz (1994) propose a low order model for the exit concentrations of a flow bioreactor.

Extremely compact low order models can be derived for those distributed parameter systems which show wave propagation characteristics such as separation and reaction processes (Marquardt, 1990). The major state variables comprise the spatial position of the wave front and some properties of the shape of the wave. This concept has been applied most notably to fixed-bed reactors (e.g. Gilles and Epple, 1981; Epple, 1986; Doyle III et al., 1996) as well as to binary or multi-component as well as reactive distillation columns (e.g. Gilles et al., 1980; Marquardt and Gilles, 1990; Hwang, 1991; Han and Park, 1993; Balasubramhanya and Doyle III, 2000; Kienle, 2000).

The results available indicate that reduction techniques for PDE models should be seriously considered at least in the sense of a first model reduction step in particular in a plant-wide model reduction problem to derive an approximate lumped parameter model of type (8) for some of the process units.

Discussion

The review in this section shows a large variety of nonlinear model reduction techniques stemming from different scientific areas. This is largely due to the lack of a unifying theory which could be used to guide the model reduction process. Truly nonlinear approaches with a sound theoretical basis are those singular perturbation techniques, which rely on some approximation of the attractive invariant manifold of the dynamical system (e.g. Duchene and Rouchon, 1996; Rhodes et al., 1999; Davis and Skodje, 1999) and nonlinear balancing techniques (Scherpen, 1993, 1996). As always in nonlinear theory, the computations are tedious or even infeasible—in particular if large-scale problems have to be tackled. An interesting alternative are those projection methods which incorporate the nonlinearity of the system in the reduction procedure at least to some extent (e.g. Pallaske, 1987; Hahn and Edgar, 1999, 2000). Some theoretical justification is available from their close relation to a more general nonlinear theory. POD has gained significant interest in recent years in particular for very large-scale processes which often occur as a result of discretizing distributed parameter systems despite their lack of theoretical foundation.

At this point, there is neither evidence whether any of the nonlinear model reduction techniques could qualify as a generic method which gives good results for any process system, nor are there guidelines available which of them to prefer for a particular class of process systems. A selection of results for type (8) models are presented in Table 1. The reductions presented are those suggested by the authors to give satisfactory results. A quantitative comparison is almost impossible and should not be attempted. Obviously, significant order reduction has only been achieved for those processes which have a distributed nature (i.e. the distillation column, fixed bed reactor, pulp digester and rapid thermal processing cases in Table 1). In these cases, model reduction based on nonlinear wave propagation can lead to even higher levels of reduction. However, in all reported studies, the

authors	system	original	reduced
Löffler and Marquardt (1991)	fixed-bed reactor	80 DAE	6 DAE
Tatrai et al. (1994)	FCC unit	20 ODE	15 ODE
Lohmann (1994)	vehicle suspension	10 ODE	7 ODE
Wisnewski and Doyle III (1996a)	continuous pulp digester	210 ODE	17 ODE
Robertson and Cameron (1997b)	evaporator	15 ODE, 30 AE	8 ODE, 34 AE
Robertson and Cameron (1997a)	compressor	51 ODE, 70 AE	20 ODE, 91 AE
Kumar et al. (1998)	CSTR	5 ODE	3 ODE
Aling et al. (1997)	rapid thermal processing	5060 ODE	10 ODE
Hahn and Edgar (1999)	distillation column	32 ODE	3 ODE
Kordt (1999)	aircraft	29 ODE	10 ODE
Hahn and Edgar (2000)	CSTR	6 ODE	4 ODE

Table 1: Selected results of nonlinear model reduction.

complexity of the reduced order model equations is significantly higher than that of the original model. For projection and POD methods, this is due to the linear combination of all the right hand sides of the original model in any equation of the reduced order model (cf. Equations 27, 43). Despite this increase in complexity, significant reductions in computational time have been observed in dynamic simulation in most cases summarized in Table 1. Plant wide models have not yet been considered. Also, the computational complexity as well as the model quality under closed loop conditions and in particular in optimization based controllers has not yet been studied in the context nonlinear model reduction.

More theoretical analysis with an emphasis on closed-loop properties and a reduction of the computational load in dynamic optimization as well as comparative studies on realistic large-scale problems are required to build up more experience which could guide the model order reduction process in a concrete context. Order reduction must be considered to be complemented by model simplification which is introduced in the following section.

Model Simplification

Model simplification is a special type of model reduction where the order of the model is preserved but the complexity of the functional expressions in the model equations is reduced. Since the computational effort is to a large extent determined by the function evaluation of the model equations, such methods are at least as important as model order reduction techniques. We will briefly present promising approaches which are applicable in general and two exemplary areas specific to chemical process systems models.

Linearization

The classical approach to the simplification of nonlinear models for control is a linearization at some nominal operating point. In particular, in a model predictive control framework, the computational complexity can be reduced drastically and the reliability and robustness of the optimization algorithms can be improved significantly, since a (convex) quadratic program has to be solved online instead of a (nonconvex) nonlinear program.

However, since the control system is required to operate in a large operational envelope with satisfactory performance, a linear model resulting from mere Jacobian linearization of the fundamental model at a nominal operating point will not suffice to adequately predict process dynamics. Instead, feedback linearization (Isidori, 1989) of the nonlinear fundamental model can be applied to produce a linear system with a set of state dependent constraints (Nevistić and Morari, 1995; Kurtz and Henson, 1997, 1998). An algorithm close to linear model predictive control can then be applied to handle the constraints. Though conceptually attractive, these techniques are limited to feedback linearizable (small-scale) processes (Morari and Lee, 1999). Therefore, they are not expected to get significant attention for optimization based control of industrial processes.

Instead of feedback linearization, Jacobian linearization at different reference points along a transient trajectory can be envisioned. Many variants of this modeling approach have been reported in the recent literature (e.g. García, 1984; Gattu and Zafiriou, 1992; Lee and Ricker, 1994) to limit the complexity of the optimization in nonlinear model predictive control. Most of the reported studies have been limited to low order models. If large-scale systems are considered, linear model reduction (e.g. Litz, 1979; Moore, 1981; Eitelberg, 1982; Bonvin and Mellichamp, 1982; Glover, 1984; Samar et al., 1995; Muscato, 2000) can be applied to reduce the computational load in a predictive control strategy. Wisnewski and Doyle III (1996b) and Doyle III and Wisnewski (2000) use such a strategy. They keep the reduced nonlinear model constant to avoid the computational burden of on-line linear model reduction along the trajectory.

An alternative to successive linearization along the tra-

jectory are interpolated piecewise linear models which are valid only locally in a certain region of the operational envelope. Different realizations of this idea have been reported for example by Banerjee et al. (1997), Johansen and Foss (1997), Chikkula et al. (1998), Banerjee and Arkun (1998), Lakshmanan and Arkun (1999), Foss et al. (2000), or Dharaskar and Gupta (2000). Though these authors largely aim at experimental identification to develop the local linear models, they could also be constructed by linearization of the fundamental model at a number of reference points followed by subsequent linear model reduction. The local models are then glued together by some interpolation strategy to provide an aggregated model valid in the whole operational envelope. At a first glance, these approaches seem to be more favorable than successive linearization and subsequent on-line model reduction, since model building can be done off-line. This is, however, not completely true, since the parameters of the fundamental model have to be adjusted on-line as part of the optimizing control system to reduce plant-model mismatch. Therefore, a tailored approach is required to adapt the piecewise linear model on-line to plant data or to the updated fundamental model.

Nonlinear Approximation of Functional Expressions

Often, the model equations contain quite complicated nonlinear expressions which result from detailed fundamental modeling and/or from subsequent nonlinear model reduction. In many cases, the right hand sides of the differential-algebraic models are formed by some nonlinear function, which comprises a number of additive, mostly nonlinear terms according to

$$\phi(\boldsymbol{x},\boldsymbol{u}) = \sum_{j=1}^{n_t} \alpha_j \phi_j(\boldsymbol{x},\boldsymbol{u}) \; . \tag{63}$$

Here, ϕ and ϕ_j are scalar functions, α_j are constant weights and $\boldsymbol{x}$ and $\boldsymbol{u}$ are vectors of given states and inputs which vary with time. Functions of this type arise, for example, in reaction kinetic models, or inevitably in reduced order models if derived by projection (cf. Equations 31, 43). In these cases, the right hand sides of the model are always linear combinations of nonlinear functions, which either comprise the right hand sides of the original model in case of a reduced order model (cf. Equation 31) or the reaction rates of elementary reactions in a reaction kinetic model.

We are interested in systematic methods which replace the probably complex functional expression of ϕ by a simpler functional expression $\tilde{\phi}$ which approximates ϕ up to a user specified tolerance for a set of trajectories denoted by $\mathcal{G}$. Obviously, the problem can be generalized by replacing the linear combination in (63) by a general nonlinear expression, which would lead to the problem of approximating a general function $\phi(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\alpha})$ by some simpler function $\tilde{\phi}(\boldsymbol{x},\boldsymbol{u},\tilde{\boldsymbol{\alpha}})$.

At a first glance, this seems to be a classical problem of multi-variate nonlinear approximation, for which many solution techniques should be readily available. However, the problem is quite complicated due to the probably large number of independent variables occurring as arguments of ϕ_j and due to the fact that the approximation should cover a set of trajectories $\mathcal{G}$. Further, we have a combinatorial component in the problem because there is no preferred candidate functional structure for $\tilde{\phi}$ a priori.

Desrochers and Al-Jaar (1985) have studied a closely related problem in a discrete-time setting. Their problem formulation can be met, if the trajectories in $\mathcal{G}$ are combined to one composite trajectory (by putting them in a sequence in time). The composite trajectory is then sampled on some time grid to result in a sequence $\{\boldsymbol{x}_k,\boldsymbol{u}_k,\phi_k\},\ k = 1\ldots K$. Their approximation problem can be reformulated as the mixed-integer nonlinear programming problem (MINLP)

$$\min_{y,\tilde{\alpha}} \sum_{k=1}^{K} e_k^2 + \boldsymbol{w}^T \boldsymbol{y} \tag{64}$$

subject to

$$\begin{aligned} e_k &= \phi_k - \tilde{\phi}_k \\ &:= \phi_k - \sum_{j=1}^{n_t} y_j \tilde{\alpha}_j \phi_j(\boldsymbol{x}_k,\boldsymbol{u}_k) \; , \\ 0 &< \sum_{j=1}^{n_t} y_j < n_t \; , \qquad \boldsymbol{y} \in \{0,1\}^{n_t} \; . \end{aligned}$$

This problem can be solved (after an appropriate reformulation to replace the disjunctions by a more favorable constraint set) by any MINLP method (at high computational expense). However, an elegant tailored solution technique has been reported by Desrochers and Al-Jaar (1985). Their method completely decouples the combinatorial part of the problem from the parameter identification problem. They first identify the most promising combination of functions ϕ_j in the simplified model on the basis of the residual error and then solve a single parameter estimation problem for the most favorable model structure. The penalty term in the objective can be chosen to account for those terms (and variables) which are preferably eliminated to directly influence the sparsity pattern of a model equation. It should be noted that—at least in some cases—some variables (and hence equations) may be eliminated simultaneously, if they only occur in the discarded functions ϕ_j.

A related technique to model simplification, specifically tailored to rapid thermal processing, a microelectronics manufacturing process, has been reported by Aling et al. (1997). Their objective is to further simplify

a nonlinear model stemming from proper orthogonal decomposition.

Obviously, there is no need to rely on expressions ϕ_j which are already present in ϕ to form the approximation $\tilde{\phi}$. Rather, any functional structure could be postulated for $\tilde{\phi}$. For example, Duchene and Rouchon (1996) suggest to consider multivariate interpolation and approximation techniques. However, such an approach seems to be impractical if the number of arguments of ϕ is large.

A more promising approach could be built on methods developed for nonlinear empirical modeling. For example McKay et al. (1997) and Marenbach et al. (1997) present a method for the identification of the structure and the parameters of nonlinear models for steady-state and dynamic processes, respectively. In their approach, process models are postulated to consist of a given set of elementary functional building blocks $\phi_i(\boldsymbol{x}, \boldsymbol{u}, \tilde{\boldsymbol{\alpha}})$. In contrast to Equation 63, these functions depend nonlinearly on unknown parameters $\tilde{\boldsymbol{\alpha}}$. They are combined in a nonlinear fashion to form the approximation $\tilde{\phi}$. Genetic programming is applied to select the best combination and to determine appropriate parameters to get the best fit of measurements. Obviously, this formulation generalizes problem (64) at the expense of a significantly higher computational complexity.

The principle advantage of this kind of methods lies in the ability to use knowledge on favorable functional forms available from fundamental modeling in defining a set of candidate building blocks. Alternatively, one could employ truly black-box nonlinear identification methods such as neural networks. For example, Shvartsman et al. (2000) report on simplification of a reduced order model of a distributed reaction system derived by proper orthogonal decomposition.

Simplification of Chemical Kinetics Models

Large-scale chemical kinetics models arise in many applications. Model complexity stems from the large number of reactions and components. For the simplification of reaction mechanisms we assume a reaction network with n_r reactions and n_s species. The complexity of the reaction kinetics model can be reduced by eliminating both, reactions and species from the reaction network. Elimination of reactions corresponds to model simplification, whereas elimination of species is a special case of order reduction. Here, we focus therefore on the first problem.

Sensitivity analysis is the most classical approach to assess the importance of individual reactions on the evolution of the concentration of all species (e.g. Seigneur et al., 1982; Brown et al., 1997, and the references cited therein). These methods determine the effect of a perturbation in a kinetic rate constant on the concentrations at some point in time or on average during the course of the reaction. Those reactions with rate constants resulting in a large sensitivity of the concentrations are considered important and should be retained in the reaction kinetics model, whereas those reactions leading to small sensitivity can be eliminated without sacrificing prediction accuracy. Sensitivity methods have been successfully applied to a variety of large-scale reaction mechanisms. However, sensitivity analysis may lead to wrong results as illustrated by means of a simple example by Petzold and Zhu (1999). Therefore, optimization based techniques have been suggested more recently by Edwards et al. (1998), Petzold and Zhu (1999), Edwards and Edgar (2000), Edwards et al. (2000), and Androulakis (2000) for a reduction of the number of reactions in a network. The problem formulations presented by these authors are variants of the MINLP in the previous section and aim at facilitating the numerical solution for large-scale problems. However, the parameters α_i in (64) are the stoichiometric coefficients of the reaction model and are (usually) not considered as degrees of freedom in model simplification.

Simplification of Physical Property Models

A classical example of reducing the computational complexity of a process model is related to the simplification of physical property models. The development of local thermodynamic models dates back into the seventies. This research has been initiated by the observation of the large fraction of computational time spent with physical property calculations in steady-state flowsheeting (Grens, 1983). The calculation of K-values in phase equilibrium models

$$y_i = K_i(\boldsymbol{x}, \boldsymbol{y}, p, T)\, x_i \;, \quad i = 1 \ldots n_c \;, \tag{65}$$

for ideal as well as strongly nonideal mixture has got particular attention due the high complexity of the models. Here, $\boldsymbol{x}, \boldsymbol{y}, p$ and T are the liquid and vapor concentrations as well as pressure and temperature under equilibrium conditions. A local model is intended to approximate the K-values as well as their derivatives with a functional expression of strongly reduced complexity. The structure of the local models is derived on physical arguments. For example, Leesley and Heyen (1977) neglect concentration dependencies and suggest a modification of Raoult's law, whereas Chimowitz and coworkers (Chimowitz et al., 1983; Chimowitz and Lee, 1985), Hager (1992) and Ledent and Heyen (1994) consider concentration dependencies by modified Porter or Margules models. Hager's equation, for example, is

$$\ln(K_i p) = A_{i,1} + \frac{A_{i,2}}{T} + (B_{i,1} + \frac{B_{i,2}}{T})(1 - x_i)^2 + B_{i,3}(1 - x_i)^2(1 + 2x_i) \;. \tag{66}$$

The local models are only valid in a limited region of the operating envelope. Hence, at least some of the model parameters ($B_{i,1}, B_{i,2}, B_{i,3}$ in the example given) must be updated along a trajectory in order to retain sufficient approximation accuracy. The parameter update can be

triggered by the simulation or optimization algorithm or by an estimate of the error between the approximate local and the original models. Model parameters are obtained from some least-squares fit of data obtained from the original model. Various variants of updating schemes have been reported by Leesley and Heyen (1977), Macchietto et al. (1986), Hillestad et al. (1989) and by Storen and Hertzberg (1997). Obviously, parameter updates result in model discontinuities. If not properly handled, these discontinuities will make simulation and optimization algorithms fail or converge to wrong solutions (Barton et al., 1998). Hence, explicit discontinuity handling or discontinuity smoothing is a necessity with these models. For the latter approach, interpolation strategies employed in linear multiple models (e.g. Foss et al., 2000; Johansen and Foss, 1997)) could be adopted here.

Significant savings in computational time have been reported for steady-state simulation and optimization (Chimowitz et al., 1984; Perregaard, 1993), dynamic simulation (Macchietto et al., 1986; Hager, 1992; Perregaard, 1993; Ledent and Heyen, 1994) and dynamic optimization (Storen and Hertzberg, 1997) if local thermodynamic models are applied.

Discussion

Nonlinear model simplification has not yet got significant attention in the systems and control literature. It is particularly suited to simplify reduced order models arising from projection methods with the objective to regain at least to some extent sparsity in the reduced model Jacobian. There has been significant activity in the context of chemical kinetics and physical property models. The variety of techniques tailored to these special problems deserve careful analysis in order to assess the potential of applying the specific concept after generalization to other model simplification problems.

For example, sensitivity analysis as worked out in chemical kinetics, is applicable in principle to the simplification of any parametric model (cf. the derivation by Seigneur et al., 1982) but—to the author's knowledge—it has not been explored for general model simplification problems. This is also true for optimization based methods given the close correspondence between reaction model and general model simplification.

On the other hand, the success of local physical property models suggests to consider similar strategies in a more general setting. The simplified model should be based on a fundamental principle rather than on some arbitrary empirical ansatz. In many cases, the simple models are only of sufficient accuracy in a limited region of the operating envelope. Then, adaptive updating of the parameters of a simple model structure using data from a rigorous model can be considered as an interesting alternative to globally valid simplified models. Obviously, a compromise needs to be established between model complexity and range of model validity. For example, a globally valid but complex neural network model (e.g. Kan and Lee (1996) for a liquid-liquid equilibrium model or Molga and Cherbański (1999) for a liquid-liquid reaction model) can be used instead of a simpler (local) model with a limited region of validity which requires parameter updating along the trajectory.

There is an obvious relation between model simplification and hybrid models as discussed above. Hybrid models are motivated by a lack of knowledge on the mechanistic details of some physico-chemical phenomena. A nonlinear regression model such as a neural network is used instead of a fundamental model to predict some process quantity (such as a reaction rate, a mass transfer rate or phase equilibrium concentrations). Model simplification on the other hand aims at reducing the complexity of a given fundamental model. Hence, hybrid modeling in the sense of Psichogios and Ungar (1992) can be readily applied to model simplification. The (typically algebraic) mechanistic model, which—for example—determines a flux (e.g. a reaction rate, see Molga and Cherbański, 1999), or a separation product flow rate, (see Safavi et al., 1999), a kinetic coefficient (e.g. a flotation rate constant, see Gupta et al., 1999), some state function (e.g. a holdup in a two-phase system, see Gupta et al., 1999) is replaced by some nonlinear regression model (such as a neural network). This regression model is typically explicit in the quantity of interest and hence can be evaluated extremely efficiently. Note, that these models can be designed for a large or a small region of validity. In the latter case, parameter updating (using the rigorous model to produce the data required) is required along the trajectory.

Model Application

We assume that a detailed dynamic model is available for example from the process design activities. This model can be simplified or reduced by physical insight, by one of the techniques discussed above, or by a combination thereof to meet the requirement of the various model-based tasks in integrated dynamic optimization and control system following a direct or some decomposition approach. This section summarizes some thoughts about the type of reduced and/or simplified models which might be used most appropriately in a certain context.

Direct Approach

We can make use of any model and apply an optimizing predictive control and a suitable reconciliation scheme to realize dynamic real-time optimization by the direct approach. Instead of following a reference trajectory set by some upper decision layer in the control hierarchy, an economical objective is maximized on-line on the receding control horizon to compute the control moves. The computational complexity of the reconciliation and con-

trol problems must, however, be quite low, since both tasks have to be executed roughly with the sampling frequency of the available measurements. Hence, any of the model order reduction and simplification techniques or combinations thereof should be employed to come up with a model of manageable computational complexity under real-time conditions.

Since there exists a huge number of possibilities for reducing a given detailed model of a realistic industrial process, constructive guidelines for designing such models for both, the reconciliation and the control task, would be extremely helpful. Such guidelines do not seem to be available yet. However, hybrid regression or trend models as well as local linearization along a trajectory or in different areas of the operating envelope seem to be attractive candidates provided the prediction horizon is chosen to reflect the prediction quality of the model. The benefit of model order reduction cannot yet be assessed due to a lack of practical experience with challenging plant-wide optimization based control problems. For large-scale systems, a structured approach to order reduction exploiting the natural spatial decomposition of a plant together with an appropriate model simplification procedure as outlined above seems to be crucial for a successful application. Such a structured approach could also be combined with the horizontal decomposition approach introduced before.

Vertical Decomposition Approach

In contrast to the direct approach, two different models of the same process are required to implement both, the dynamic optimizer (DO) and the model predictive controller (MPC) together with their respective estimators on both levels (cf. Figure 3). Ideally, these models should be derived from a detailed master model to guarantee consistency. The requirements on the models are different in both cases:

(a) *Computational constraints:* DO is executed with a much lower frequency (say in the order of once every one or two hours) whereas MPC is executed with a higher frequency (say in the order of once every couple of minutes). Hence, a higher computational complexity can be tolerated for DO as compared to MPC. Obviously, higher frequencies would facilitate better performance in case a sufficiently valid model would be available. Therefore, the complexity of the model should be minimal in both cases, though larger for DO than for MPC, provided the requirements on prediction accuracy are still satisfied.

(b) *Prediction accuracy:* DO should be able to predict economical performance as well as state and output trajectories with sufficient accuracy over the full operating region. Hence, a fairly detailed model incorporating the major process nonlinearities is required. In contrast, the model implemented in the MPC must predict the setpoint deviation, the outputs and possibly the potentially constrained states in the vicinity of the reference trajectory only. Therefore, a simpler model with a much smaller region of validity can be chosen in this case. Even a linear model, updated along the trajectory, may qualify in this case.

(c) *Frequency range:* Due to the different execution frequencies and the different tasks of both levels, the models have to cover the low and high frequency behaviors of the plant for DO and MPC, respectively. Hence, a model with an appropriate prediction quality on a fast time-scale is required for MPC whereas a model a slow time-scale is needed for DO.

Model order reduction by projection, by equation residual minimization or by proper orthogonal decomposition in conjunction with model simplification can be employed for implementation of DO as well as of MPC. Different degrees of reduction should be employed for DO and MPC, however, to account for the specific requirements on prediction errors and computational complexity. A mildly reduced model can be used for DO, whereas a strongly reduced model must be used for MPC to meet the computational complexity constraints. While requirements (a) and (b) could be met by this choice, requirement (c) is definitely in conflict. Though not explicitly incorporated in the model reduction techniques, a mildly reduced model will cover faster time-scales while a strongly reduced model will cover slow time-scales only. This conclusion is based on an interpretation of the strategy employed during model reduction. For example, the integral average of the projection error in Pallaske's method is reduced by a quantifiable amount, if the dimension of the reduced model is increased (Pallaske, 1987; Löffler and Marquardt, 1991). The larger the dimension of the reduced model, the shorter are the time-scales incorporated.

Provided the model used for DO is updated regularly and thus provides updated reference trajectories to the MPC which reflect process economics and comply with constraints the requirements on the model used in the MPC are quite relaxed. In particular, a linearization of some kind (along the reference trajectory for example) together with linear model reduction could be fully sufficient to achieve adequate overall performance. A relatively simple approach for the implementation of DO are the Wiener/Hammerstein hybrid models which build on an available steady-state fundamental (optimization) model. A key issue is in all cases the integration of the model update during reconciliation on the DO and MPC levels.

Some time-scale separation can be achieved by a proper choice of the models used on the DO and MPC levels. However, there is no theoretical basis for keeping

the models on both levels consistent to each other. This problem is well-known even in state of the art (steady-state) real-time optimization and control where serious performance deterioration has been observed in some cases. If time-scale separation is envisioned, singular perturbation methods might be more favorable. They explicitly address this issue by construction to yield two (or even multiple) dynamic models valid on certain time-scales. Though it may be doubted whether exact nonlinear techniques (Marino and Kokotovic, 1988; Krishnan and McClamroch, 1994; Kumar et al., 1998) are widely applicable to construct a singularly perturbed system in standard form, the approximate projection methods of Pallaske (1987) or Lall et al. (1999) employing linear transformations should be applicable to separate even large-scale model into a fast (non-dominant) and a slow (dominant) submodel. If the fast and slow subsystems are envisioned to be used for implementation of DO and MPC on a slow and a fast time-scale, the coupling between both subsystems may lead to serious interaction which could deteriorate control system performance or even stability.

There are cases (e.g. Stiharu-Alexe and O'Shea, 1995; Kumar and Daoutidis, 2000) where not only the state but also the control and output variables are partitioned in the slow and fast subsystems by singular perturbation. In those cases, a completely partitioned DO and MPC level can be implemented (Stiharu-Alexe and O'Shea, 1995), where the control variables of the fast and the slow subsystems are manipulated by MPC and DO completely independently employing the fast and slow output with high and low sampling rates respectively. A consistent time-scale separation can be achieved in this case.

It is still a largely open question, when and how singular perturbation techniques can be applied to partition a model into a fast and a slow submodel to be used in a consistent manner on the DO and MPC levels in time-scale decomposition.

Alternatively, one may employ multi-resolution methods (Binder et al., 1998) to develop models on different time-scales which could be used in DO and MPC, respectively. The basic idea is briefly discussed next using a scalar system (8) for the sake of a simpler notation. The continuous model is projected onto a sparse multi-scale subspace by a Wavelet-Galerkin method. The state and control vectors x and u are expanded in a series according to

$$x = \boldsymbol{d}^T\ \boldsymbol{\psi}(t) = \boldsymbol{d}_s^T \boldsymbol{\psi}_s(t) + \boldsymbol{d}_f^T \boldsymbol{\psi}_f(t)\ , \tag{67}$$

$$u = \boldsymbol{e}^T\ \boldsymbol{\psi}(t) = \boldsymbol{e}_s^T \boldsymbol{\psi}_s(t) + \boldsymbol{e}_f^T \boldsymbol{\psi}_f(t)\ , \tag{68}$$

where $\boldsymbol{\psi}$ denotes the vector of multi-scale basis functions and the vectors $\boldsymbol{d}$ and $\boldsymbol{e}$ contain the expansion coefficients for the state and control variable, respectively. The expansion can be divided into a leading sum referring to the low frequency content (coefficients $\boldsymbol{d}_s, \boldsymbol{e}_s$ and basis functions $\boldsymbol{\psi}_s(t)$) and into a residual sum which covers the high frequency content (coefficients $\boldsymbol{d}_f, \boldsymbol{e}_f$ and basis functions $\boldsymbol{\psi}_f(t)$).

A discretized model of the slow system is obtained as

$$\boldsymbol{\gamma}_s(\boldsymbol{d}_s^s, \boldsymbol{e}_s) = \boldsymbol{0} \tag{69}$$

after Galerkin projection. The vector function $\boldsymbol{\gamma}_s$ results from the projection of the scalar model (8) with the basis functions $\boldsymbol{\psi}_s(t)$. It fixes the expansion coefficients of the states $\boldsymbol{d}_s^s$ as a function of those of the control variable $\boldsymbol{e}_s$ to approximate the low frequency content of the state $x_s = (\boldsymbol{d}_s^s)^T\, \boldsymbol{\psi}_s(t)$. A discretized model of the fast system is obtained from a Galerkin projection with the basis functions $\boldsymbol{\psi}_s(t)$ and $\boldsymbol{\psi}_f(t)$ as

$$\boldsymbol{\gamma}_f(\boldsymbol{d}_s^f, \boldsymbol{d}_f^f, \boldsymbol{e}_s, \boldsymbol{e}_f) = \boldsymbol{0} \tag{70}$$

which fixes the expansion coefficients $\boldsymbol{d}_s^f, \boldsymbol{d}_f^f$ of the state variable as a function of those of the control variable, $\boldsymbol{e}_s, \boldsymbol{e}_f$, to approximate the low and high frequency content of the state

$$\tilde{x} = \tilde{x}_s + \tilde{x}_f = (\boldsymbol{d}_s^f)^T \boldsymbol{\psi}_s(t) + (\boldsymbol{d}_f^f)^T \boldsymbol{\psi}_f(t)\ . \tag{71}$$

We would get a decoupling of the slow and the fast subsystem, if $\boldsymbol{d}_s^s$ and $\boldsymbol{d}_s^f$, the expansion coefficients for the slow contributions to the state in the slow and the fast models, respectively, would be identical. However, due to the wavelet properties, we only find

$$\boldsymbol{d}_s^f = \boldsymbol{d}_s^s + \boldsymbol{\delta}_f \tag{72}$$

with typically small corrections $\boldsymbol{\delta}_f \neq \boldsymbol{0}$, i.e. $\|\boldsymbol{\delta}_f\| = \varepsilon \|\boldsymbol{d}_s^s\|$ with $\varepsilon < 1$.

This approach to time-scale separation may satisfy requirements (b) and (c) above but it is definitely in conflict with requirement (a) since the size of the fast discretized model is much larger than that of the slow. It is still an open issue, whether and how this problem can be solved. In this case, this approach could be an interesting alternative to implement DO and MPC and its associated estimators on both levels.

Closed-loop Model Validation

Obviously, the validity of the various models has to be assessed in the context of their application in the various modules of the operations support system.

A comparison of the *open loop behavior* is possible by a variety of means. Examples are nonlinearity measures (Helbig et al., 2000b, and references cited therein), to compare the loss of nonlinearity between two candidate models, step responses, or nonlinear describing function analysis (Amrhein et al., 1993) to reveal the frequency content of a nonlinear model. One might argue, that measures of uncertainty could be derived at least in principle by comparison of the reduced and the

original model to be used later during estimation and control design. In fact, Andersson et al. (1999) provide a very interesting result on the comparison and also the simplification of two uncertain models. These authors define a simplification error in terms of the L_2-induced gain. They further show that this error can be computed by convex optimization for *linear uncertain systems* and for a certain (broad) class of *nonlinear uncertain systems* with isolated static nonlinearities. The method can be applied to open-loop as well as closed-loop systems. In the linear case, their result generalizes truncation and singular perturbation. Their result is a good basis for comparing and simplifying large-scale nonlinear process models.

Open-loop tests are not sufficient to test the validity of candidate (reduced) models under closed-loop control conditions. A key question is to relate any simplification of the model to the unavoidable loss of economical performance of the feedback control system and to guarantee closed loop stability despite the simplifications made. Stability loss and performance degradation are well-known phenomena if a controller is designed by means of a reduced order model and applied to the plant in a linear setting if no special design technique is applied (see Zhou et al. (1995), Bendotti and Beck (1999), Wortelboer et al. (1999) for recent examples). All these problems will carry over to the nonlinear case at least in principle. However, there is very little knowledge yet about these issues for integrated dynamic optimization and control as investigated in this work. Obviously, the problem could be addressed from a robust control perspective. If we assume, at least for the moment, the detailed (or nominal) model to perfectly match the plant, any model reduction introduces quantifiable uncertainty. This is in contrast to mainstream research in robust (model predictive) control, where the model error cannot be quantified precisely. The knowledge on the uncertainty introduced by model reduction could be employed to robustly accommodate the mismatch between the reduced models (of the estimator and controller) and the real plant (perfectly matched by the nominal model).

There are some starting points for future research in the recent literature on nonlinear control. For example, Scherpen (1993, 1996) proves the stability of reduced models derived from truncation after nonlinear balancing. It is worth noting, that similar stability results are not available for the more empirical methods suggested by Pallaske (1987), Lohmann (1994), Lall et al. (1999), or Kordt (1999) though they are based on linear transformations only and should therefore be simpler to address. The stability problem of a closed loop system with a controller designed by means of a reduced model obtained from nonlinear H_∞ balancing (Scherpen, 1996) is addressed by Pavel and Fairman (1997). They generalize results on closed loop H_∞ balanced truncation by Mustafa and Glover (1991) from the linear to the nonlinear case. The authors provide first a criterion for maintaining the closed loop stability if the controller is designed by solving the nonlinear normalized $H_\infty(L_2)$ control problem for the reduced model and applied to the plant. Further, the degradation of performance is analyzed in closed loop.

For reduced models obtained from some singular perturbation analysis, there are not only strong results available for a number of nonlinear control system design techniques but also for open-loop optimal control. For example, Artstein and Gaitsgory (2000) proved just recently convergence of the value function of the perturbed system to that of the slow system for $\epsilon \to 0$ under mild assumptions (such as controllability of the fast subsystem) for general systems in standard singularly perturbed form. Such results could be a starting point for the analysis under closed-loop conditions in future research.

Conclusions

Optimization-based control of transient processes requires nonlinear models of sufficient predictive quality which can be employed for the various tasks in the feedback control system in a real-time environment. On the basis of a suitable formulation of the control problem and some thoughts on its implementation, we focussed on fundamental modeling and in particular on nonlinear model reduction which comprises both, model order reduction and model simplification. A large variety of methods with differing theoretical justification has been reviewed and put into perspective. Though, there has been significant progress in the last 10 years, a thorough understanding which technique could and should advantageously be used in optimization-based control and how it should be tailored to a specific problem is largely lacking. The situation is even worse, if a (vertical) decomposition of the optimizing control system is envisioned in order to extend the state of the art in (steady-state) real-time optimization where a multi-level architecture is typically implemented. Some of the major open research problems are:

(a) Lumped process systems models are usually of differential-algebraic type. With the exception of the work of Löffler and Marquardt, there are no general model order reduction techniques for this class of systems available.

(b) All of the nonlinear reduction techniques rely on a representative set of trajectories. The selection of this set is crucial for the success of the reduction. To the author's knowledge there are no systematic techniques yet to guide this selection. Obviously, the set should be as close as possible to the trajectories occurring in closed-loop.

(c) In contrast to linear model order reduction, only

truncation reduces the model order significantly. If residualization or some sort of steady-state assumption is introduced, differential-algebraic equations result which are often as difficult to solve as the original problem. Systematic approaches based on slaving as introduced in POD or even nonlinear regression of the set of nonlinear equations with a simple nonlinear map are required.

(d) Most of the model reduction techniques lead to a reduced number or equations which are however of a significantly higher functional complexity. The structure and the sparsity of the original model is lost. In particular, in dynamic optimization, exploitation of model structure is a key to a high performant numerical solution. Any means of preserving structure and sparsity at least to some extent during model order reduction is highly beneficial. In addition, sparsity can be reintroduced by systematic model simplification for example by eliminating most of the nonlinear terms in the right hand sides of the reduced model.

(e) There has been almost no work on the system theoretical properties of resulting nonlinear reduced models. Most of the reduction techniques cannot be expected to preserve stability, observability, or controllability properties. In principle, equation residual minimization (Lohmann, 1994, 1995) could be extended by additional constraints not only preserving steady-state accuracy but also stability (using the approach of Mönnigmann and Marquardt (2000), for formulating appropriate constraints).

(f) All process models are hybrid by nature since they comprise fundamental and empirical parts. The appropriate combination of fundamental and empirical knowledge is still an open issue even in case of open-loop (simulation) applications. In closed-loop, an appropriate parameterization of the uncertainty in both model constituents as a basis for an efficient reconciliation is an even more challenging problem.

(g) Excitation frequencies or the magnitude of the controls and disturbances driving the process in closed-loop are largely unknown in most cases. Ideally, on-line adaptation of the structure and not only the parameters of the reduced model on the basis of actual or historical process data would be most preferable. Computational singular perturbation (Massias et al., 1999) or adaptive Galerkin methods (von Watzdorf and Marquardt, 1997; Briesen and Marquardt, 2000) as developed for the treatment of multi-component separation and reaction processes could be a first starting point for the development of a more general technique.

(h) The validity of the model under closed-loop conditions is critical for the success of integrated dynamic optimization and control. Since there are most likely many models interacting in the various functional blocks of the control and optimization system systematic means of constructing consistent models are required to reach high performance. It is a largely open issue how to efficiently assess the validity of the individual models with respect to the prediction of states and outputs as well as gradient and sensitivity information, the consistency between these models, as well as the stability and performance of the integrated system a priori (i.e. during the early phases of the design phase of the control system).

Hopefully this review and this list of major research challenges rises interest in the systems and control community to work in this fruitful and rewarding area. Obviously, most of the questions are not hard core process control problems, but they are rather at the interfaces to related fields which renders them more interesting and more challenging at the same time.

Acknowledgments

This research is partly supported by the European Commission under grant G1RD–CT–1999–00146 in the INCOOP project. INCOOP strives for the integration of control and optimization of transiently operated chemical processes. Fruitful discussions with all the INCOOP team members, in particular with Ton Backx, Okko Bosgra, Jogchem v.d. Berg and Rob Tousain, are gratefully acknowledged. The support of Elke Krumrück, Jitendra Kadam, Adel Mhamdi, Martin Schlegel, and Thomas Binder during the preparation of the manuscript is also acknowledged.

References

Abel, O., A. Helbig, and W. Marquardt, Optimization approaches to control-integrated design of industrial batch reactors, In Berber, R. and C. Kravaris, editors, *Nonlinear Model Based Process Control*, NATO-ASI Series, pages 513–551. Kluwer, Dordrecht (1998).

Afanasiev, K. and M. Hinze, Adaptive control of a wake flow using proper orthogonal decomposition, Technical Report 648, Fachbereich Mathematik, TU Berlin (1999).

Agarwal, M., "Combining neural and conventional paradigms for modelling, prediction and control," *Int. J. Sys. Sci.*, **28**(1), 65–81 (1997).

Aling, H., R. L. Kosut, A. Emami-Naeini, and J. L. Ebert, Nonlinear model reduction with application to rapid thermal processing, In *Proceedings of the 35th Conf. of Decision and Control*, pages 4305–4310, Kobe (1996).

Aling, H., S. Banerjee, A. K. Bangia, V. Cole, J. Ebert, A. Emami-Naeini, K. F. Jensen, I. G. Kevrekidis, and S. Shvartsman, Nonlinear Model Reduction for Simulation and Control of Rapid Thermal Processing, In *Proceedings of ACC 1997*, volume 4, pages 2233–2238, Albuquerque (1997).

Allgöwer, F., T. A. Badgwell, J. S. Qin, J. B. Rawlings, and S. J. Wright, Nonlinear Predictive Control and Moving Horizon Estimation, In Frank, P. M., editor, *Advances in Control*, pages 391–449. Springer Verlag (1999).

Amrhein, M., F. Allgöwer, and W. Marquardt, Validation and analysis of linear distillation models for controller design, In *Proceedings of the Proc. Europ. Control Conf.*, pages 655–660 (1993).

Andersson, L., A. Rantzner, and C. Beck, "Model comparison and simplification," *Int. J. Robust and Nonlinear Control*, **9**, 157–181 (1999).

Androulakis, I. P., "Kinetic mechanism reduction based on an integer programming approach," *AIChE J.*, **46**(2), 361–371 (2000).

Aris, R., "Manners makyth modellers," *Trans. Inst. Chem. Eng.*, **69**(Part A), 165–174 (1991).

Armaou, A. and P. D. Christofides, "Wave suppression by nonlinear finite-dimensional control," *Chem. Eng. Sci.*, **55**, 2627–2640 (2000).

Artstein, Z. and V. Gaitsgory, "The value function of singularly perturbed control systems," *Applied Mathematics and Optimization*, **41**, 425–445 (2000).

Aubry, N., P. Holmes, J. L. Lumley, and E. Stone, "The dynamics of coherent structures in the wall region of a turbulent boundary layer," *J. Fluid Mech.*, **192**, 115 (1988).

Backx, T., O. Bosgra, and W. Marquardt, Towards Intentional Dynamics in Supply Chain Conscious Process Operations, FOCAPO 98, Snowmass, www.lfpt.rwth-aachen.de/Publication/Techreport/ (1998). Revised version submitted to *J. Proc. Control*, July 2000.

Backx, T., O. Bosgra, and W. Marquardt, Integration of model predictive control and optimization of processes, In *Proceedings ADCHEM 2000*, volume 1, pages 249–260 (2000).

Baker, J. and P. D. Christofides, "Output feedback control of parabolic PDE Systems with nonlinear spatial differential operators," *Ind. Eng. Chem. Res.*, **38**, 4372–4380 (1999).

Baker, J. and P. D. Christofides, "Finite-dimensional approximation and control of non-linear parabolic PDE systems," *Int. J. Control*, **73**(5), 439–456 (2000).

Balasubramhanya, L. S. and F. J. Doyle III, "Nonlinear model-based control of a batch reactive distillation column," *J. Proc. Cont.*, **10**, 209–218 (2000).

Banerjee, A. and Y. Arkun, "Model predictive control of plant transitions using a new identification technique for interpolating nonlinear models," *J. Proc. Cont.*, **8**(5/6), 441–457 (1998).

Banerjee, A., Y. Arkun, B. Ogunnaike, and R. Pearson, "Estimation of nonlinear systems using linear multiple models," *AIChE J.*, **43**(5), 1204–1226 (1997).

Barton, P. I. and C. C. Pantelides, "Modeling of combined discrete/continuous processes," *AIChE J.*, **40**(6), 966–979 (1994).

Barton, P. I., R. J. Allgor, W. F. Feehery, and S. Galán, "Dynamic optimization in a discontinuous world," *Ind. Eng. Chem. Res.*, **37**(3), 966–981 (1998).

Bemporad, A. and M. Morari, "Control of systems integrating logic, dynamics, and constraints," *Automatica*, **35**(3), 407–427 (1999).

Bendotti, P. and C. L. Beck, "On the role of LFT model reduction methods in robust controller synthesis for a pressurized water reactor," *IEEE Trans. Cont. Sys. Tech.*, **7**, 248–257 (1999).

Biegler, L. T. and G. B. Sentoni, "Efficient Formulation and Solution of Nonlinear Model Predictive Control Problems," *Latin American Applied Research*, **30**(4), 315–324 (2000).

Biegler, L. T., I. E. Grossmann, and A. W. Westerberg, "A note on approximation techniques used for process optimization," *Comput. Chem. Eng.*, **9**(2), 201–206 (1985).

Binder, T., L. Blank, W. Dahmen, and W. Marquardt, Towards multiscale dynamic data reconciliation, In Berber, R. and C. Kravaris, editors, *Nonlinear Model Based Process Control*, NATO-ASI Series, pages 623–665, Dordrecht. Kluwer (1998).

Bogusch, R. and W. Marquardt, "A Formal Representation of Process Model Equations," *Comput. Chem. Eng.*, **21**(10), 1105–1115 (1997).

Bogusch, R., B. Lohmann, and W. Marquardt, "Computer-aided process modeling with MODKIT," *Comput. Chem. Eng.* (2001). In press.

Bonvin, D. and D. A. Mellichamp, "A unified derivation and critical review of modal approaches to model reduction," *Int. J. Control*, **35**(5), 829–848 (1982).

Brenan, K. E., S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics (1996).

Briesen, H. and W. Marquardt, "Adaptive model reduction and simulation of the thermal cracking of multicomponent hydrocarbon mixtures," *Comput. Chem. Eng.*, **24**, 1287–1292 (2000).

Brown, N. J., G. Li, and M. L. Koszykowski, "Mechanism reduction via principal component analysis," *Int. J. Chem. Kinet.*, **29**(6), 393–414 (1997).

Bryson, A. E. and Y.-C. Ho, *Applied Optimal Control*. Taylor and Francis, Bristol (1975).

Chikkula, Y., J. H. Lee, and B. A. Ogunnaike, "Dynamically scheduled MPC of nonlinear processes using hinging hyperplane models," *AIChE J.*, **44**(12), 2658–2674 (1998).

Chimowitz, E. H. and C. S. Lee, "Local thermodynamic models for high pressure process calculations," *Comput. Chem. Eng.*, **9**(2), 195–200 (1985).

Chimowitz, E. H., T. F. Anderson, S. Macchietto, and L. F. Stutzman, "Local models for representing phase equilibria in multicomponent, nonideal vapor-liquid and liquid-liquid systems. 1. Thermodynamic approximation functions," *Ind. Eng. Chem. Proc. Des. Dev.*, **22**, 217–225 (1983).

Chimowitz, E. H., S. Macchietto, T. F. Anderson, and L. F. Stutzman, "Local models for representing phase equilibria in multicomponent, nonideal vapor-liquid and liquid-liquid systems. 2. Application to process design," *Ind. Eng. Chem. Proc. Des. Dev.*, **23**, 609–618 (1984).

Cho, Y. S. and B. Joseph, "Reduced-order steady-state and dynamic models for separation processes. I. Development of the model reduction procedure. II. Application to nonlinear multicomponent systems.," *AIChE J.*, **29**, 261–276 (1983).

Christofides, P. D. and P. Daoutidis, "Feedback control of two-time-scale nonlinear systems," *Int. J. Control*, **63**, 965–994 (1996).

Christofides, P. D. and P. Daoutidis, "Finite-dimensional control of parabolic PDE systems using approximate intertial manifolds," *J. Math. Anal. Appl.*, **216**(2), 398–420 (1997).

Čiegis, R., R. Šablinskas, and J. Waśniewski, Numerical integration on distributed-memory parallel systems, In *Recent Advances in Parallel Virtual Machine and Message Passing Interface. Lecture Notes in Computer Science*, pages 329–336. Springer (1997).

Costa, A. C., A. S. W. Henriques, T. L. M. Alves, R. Maciel Filho, and E. L. Lima, "A hybrid neural model for the optimization of fed-batch fermentation," *Braz. J. Chem. Eng.*, **16**(1) (1999).

Davis, M. J. and R. T. Skodje, "Geometric investigation of low-dimensional in systems approaching equilibrium," *J. Chem. Phys.*, **111**(3), 859–874 (1999).

de Veaux, R. D., R. Bain, and L. H. Ungar, "Hybrid neural network models for environmental process control," *Environmetrics*, **10**, 225–236 (1999).

Desrochers, A. A. and R. Y. Al-Jaar, "A method for high order linear system reduction and nonlinear system simplification," *Automatica*, **21**(1), 93–100 (1985).

Dharaskar, K. P. and Y. P. Gupta, "Predictive control of nonlinear processes using interpolated models," *Chem. Eng. Res. Des.*, **78**(A4), 573–580 (2000).

Dochain, D. and B. Bouaziz, "Approximation of the model of fixed bed reactors via a singular perturbation approach," *Math. Comp. Simul.*, **37**, 165–172 (1994).

Doyle III, F. J. and P. A. Wisnewski, Nonlinear multi-rate MPC with large scale fundamental models: application to a continuous Kamyr Digester, In Allgöwer, F. and A. Zheng, editors, *Nonlinear Model Predictive Control*, pages 419–432. Birkhäuser, Basel (2000).

Doyle III, F. J., H. M. Budman, and M. Morari, ""Linearizing" controller design for a packed-bed reactor using a low-order wave propagation model," *Ind. Eng. Chem. Res.*, **35**, 3567–3580 (1996).

Duchene, P. and P. Rouchon, "Kinetic scheme reduction via geometric singular perturbartion techniques," *Chem. Eng. Sci.*, **51**(20), 4661–4672 (1996).

Edwards, K. and T. F. Edgar, "Reaction set simplification using variable selection techniques," *Chem. Eng. Sci.*, **55**, 551–572 (2000).

Edwards, K., T. F. Edgar, and V. I. Manousiouthakis, "Kinetic model reduction using genetic algorithms," *Comput. Chem. Eng.*, **22**(1-2), 239–246 (1998).

Edwards, K., T. F. Edgar, and V. I. Manousiouthakis, "Reaction mechanism simplification using mixed-integer nonlinear programming," *Comput. Chem. Eng.*, **24**(1), 67–79 (2000).

Eitelberg, E., "Model reduction and perturbation structures," *Int. J. Control*, **35**, 1029–1050 (1982).

Epple, U., Model reduction for nonlinear systems with distributed parameters, In *Proc. IFAC Symposium DYCOPS'86 Dynamics and Control of Chemical Reactors and Distillation Columns, Bournemouth, UK*, pages 279–284 (1986).

Findeisen, W., M. Brdys, K. Malinowski, P. Tatjewski, and A. Wozniak, *Control and Coordination in Hierarchical Systems*. Wiley (1980).

Foias, C. and R. Témam, "Algebraic approximation of attractors: The finite dimensional case," *Physica D*, **32**, 163 (1988).

Foias, C., G. R. Sell, and R. Témam, "Inertial manifolds for nonlinear evolutionary equations," *J. Differential Equations*, **73**, 309 (1988).

Foss, B. A., B. Lohmann, and W. Marquardt, "A field study of the industrial modeling process," *J. Proc. Cont.*, **8**, 325–337 (1998).

Foss, B. A., T. A. Johansen, and A. V. Sørensen, "Nonlinear predictive control using local models: Applied to a batch fermentation process," *Control Eng. Practice*, **3**(3), 389–396 (2000).

Fukunaga, K., *Introduction to Statistical Pattern Recognition*. Academic Press (1990).

Ganesh, N. and L. T. Biegler, "A robust technique for process flowsheet optimization using simplified model approximations," *Comput. Chem. Eng.*, **11**(6), 553–565 (1987).

García, C. E., "Quadratic dynamic matrix control of non-linear processes. An application to a batch reactor process," *AIChe Annual Meeting, San Francisco* (1984).

Gattu, G. and E. Zafiriou, "Nonlinear Quadratic Dynamic Matrix Control with State Estimation," *Ind. Eng. Chem. Res.*, **31**(4), 1096–1104 (1992).

Genyuan, L. and H. Rabitz, "Combined symbolic and numerical approach to constrained nonlinear lumping: with application to an H2/O2 oxidation model," *Chem. Eng. Sci.*, **51**(21), 4801–4816 (1996).

Gilles, E. D. and U. Epple, Nonlinear wave propagation and model reduction of the fixed-bed reactor, In *Modelling of Chemical Reaction Systems*, pages 312–336. Springer-Verlag, Berlin (1981).

Gilles, E. D., B. Retzbach, and F. Silberberger, Mopdeling, simulation and control of an extractive distillation column, In *ACS Symposium Series*, volume 124, pages 481–492 (1980).

Glover, K., "All optimum Hankel-norm approximations of linear multivariable systems and their L_2-error bounds," *Int. J. Control*, **39**(6), 1115–1193 (1984).

Graham, M. D. and I. G. Kevrekidis, "Alternative approaches to the Karhunen-Loeve decomposition for model reduction and data analysis," *Comput. Chem. Eng.*, **20**(5), 495–506 (1996).

Grens, E. A., "Efficient use of thermodynamic models in process calculations," *Proc. Int. Conf. Found. Comp. Aid. Proc. Des., Snowmass*, pages 249–302 (1983).

Gupta, S., P.-H. Liu, S. A. Svoronos, R. Sharma, and N. A. Abdel-Khalek, "Hybrid first-principles/neural networks model for column flotation," *AIChE J.*, **45**(3), 557–566 (1999).

Hager, J. M., *Lokale Modelle zur Berechnung von Gas-Flüssigkeits-Phasengleichgewichten*, PhD thesis, Universität Stuttgart (1992).

Hahn, J. and T. F. Edgar, "An improved method for nonlinear model reduction using balancing of empirical gramians," *Comp. Chem. Eng.* (1999). submitted.

Hahn, J. and T. F. Edgar, Reduction of nonlinear models using balancing of balancing of empirical gramians Galerkin projections, In *Proceedings Amer. Control. Conf. 2000*, pages 2864–2868, Chicago (2000).

Han, H. and S. Park, "Control of high-purity distillation columns using nonlinear wave theory," *AIChE J.*, **39**, 787–796 (1993).

Hasenjäger, E., *Digitale Zustandsregelung von Parabolantennen unter Berücksichtigung von Nichtlinearitäten*. VDI Verlag (1991).

Helbig, A., O. Abel, and W. Marquardt, Model predictive control for on-line optimization of semi-batch reactors, In *Proc. Amer. Control Conf.*, volume 9 (1998).

Helbig, A., W. Marquardt, and F. Allgöwer, "Nonlinearity measurers: definition, computation and applications," *J. Proc. Cont.*, **10**, 113–123 (2000b).

Helbig, A., O. Abel, and W. Marquardt, Structural Concepts for Optimization Based Control of Transient Processes, In Allgöwer, F. and A. Zheng, editors, *Nonlinear Model Predictive Control*, pages 295–311. Birkhäuser, Basel (2000a).

Henson, M. A., "Nonlinear model predictive control: Current status and future directions," *Comput. Chem. Eng.*, **23**, 187–202 (1998).

Hillestad, M., C. Sorlie, T. F. Anderson, I. Olsen, and T. Hertzberg, "On estimating the error of local thermodynamic models—a general approach," *Comput. Chem. Eng.*, **13**(7), 789–796 (1989).

Holmes, P., J. L. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge Univ. Press (1996).

Hwang, Y.-L., "Nonlinear wave theory for dynamics of binary distillation columns," *AIChE J.*, **37**, 705–723 (1991).

Isidori, A., *Nonlinear Control Systems*. Springer (1989).

Jadach, S., "Multi-dimensional general purpose Monte Carlo generator with self-adapting simplical grid," *Computer Physics Communications*, **130**(3), 244–259 (2000).

Jarke, M. and W. Marquardt, "Design and evaluation of computer-aided process modeling tool," *AIChE Symp. Ser.*, **92**(312), 97–109 (1996).

Johansen, T. A. and B. A. Foss, "Representing and learning unmodeled dynamics with neural network memories," *Proc. Amer. Control Conf.*, **3**, 3037 (1992).

Johansen, T. A. and B. A. Foss, "Operating regime based process modeling and identification," *Comput. Chem. Eng.*, **21**(2), 159–176 (1997).

Jose, R. A. and L. H. Ungar, "Pricing interprocess streams using slack auctions," *AIChE J.*, **46**(3), 575–587 (2000).

Kadam, J. V., Index analysis and adaptive refinement in multiscale dynamic optimization, Master's thesis, Dept. of Chem. Engng., Indian Institute of Technology Bombay and Lehrstuhl für Prozesstechnik, RWTH Aachen (2000).

Kan, P. and C. J. Lee, "A neural network model for prediction of phase equilibria in aqueous two-phase extraction," *Ind. Eng. Chem. Res.*, **35**(6), 2015–2023 (1996).

Kienle, A., "Low-order dynamic models for ideal multicomponent distillation processes using nonlinear wave theory," *Chem. Eng. Sci.*, **55**(10), 1817–1828 (2000).

Kokotovic, P. V., H. K. Khalil, and J. O'Reilly, *Singular Perturbations in Control: Analysis and Design*. Academic Press (1986).

Kordt, M., "Nichtlineare Ordnungsreduktion für ein Transportflugzeug," *Automatisierungstechnik*, **47**(11), 532–539 (1999).

Kramer, M. A. and M. L. Thompson, "Embedding theoretical models in neural networks," *Proc. Amer. Control Conf.*, **1**, 475 (1992).

Krishnan, H. and N. H. McClamroch, "On the connection between nonlinear differential- algebraic equations and singularly perturbed control systems in nonstandard form," *IEEE Trans. Auto. Cont.*, **39**(5), 1079–1084 (1994).

Kumar, A. and P. Daoutidis, "Nonlinear dynamics and control of process systems with recycle," *ADCHEM Preprints*, **1**, 13–22 (2000).

Kumar, A., P. D. Christofides, and P. Daoutidis, "Singular perturbation modeling of nonlinear processes with nonexplicit time-scale multiplicity," *Chem. Eng. Sci.*, **53**(8), 1491–1504 (1998).

Kunisch, K. and S. Volkwein, "Control of the Burgers equation by a reduced-order approach using proper orthogonal decomposition," *J. Optim. Theo. Appl.*, **102**(2), 345–371 (1999).

Kurtz, M. J. and M. A. Henson, "Input-output linearizing control of constrained nonlinear processes," *J. Proc. Cont.*, **7**(1), 3–17 (1997).

Kurtz, M. J. and M. A. Henson, "Feedback linearizing control of discrete-time nonlinear systems with input constraints," *Int. J. Control*, **70**(4), 603–616 (1998).

Lakshmanan, N. M. and Y. Arkun, "Estimation and model predictive control of non-linear batch processes using linear parameter varying models," *Int. J. Control*, **72**(7/8), 659–675 (1999).

Lall, S., J. E. Marsden, and S. Glavaski, "Empirical model reduction of controlled nonlinear systems," *14th IFAC World Congress, Beijing* (1999).

Ledent, T. and G. Heyen, "Dynamic approximation of thermodynamic properties by means of local models," *Comput. Chem. Eng.*, **18**(Suppl.), S87–S91 (1994).

Lee, J. H. and N. L. Ricker, "Extended Kalman filter based nonlinear model predictive control," *Ind. Eng. Chem. Res.*, **33**(6), 1530–1541 (1994).

Lee, P. L., L. Huaizhong, and I. T. Cameron, "Decentralized control design for nonlinear multi-plants: A gap metric approach," *Chem. Eng. Sci.*, **55**, 3743–3758 (2000).

Lee, J. H., Modeling and identification for nonlinear model predictive control: Requirements, current status and future research needs, In Allgöwer, F. and A. Zheng, editors, *Nonlinear Model Predictive Control*, pages 269–293. Birkhäuser, Basel (2000).

Leesley, M. E. and G. Heyen, "The dynamic approximation method of handling vapor-liquid equilibrium data in computer calculations for chemical processes," *Comput. Chem. Eng.*, **1**, 109–112 (1977).

Levine, J. and P. Rouchon, "Quality control of binary distillation columns via nonlinear aggregated models," *Automatica*, **27**, 463–480 (1991).

Litz, L., "Ordnungsreduktion linearer Zustandsraummodelle durch Beibehaltung der dominanten Eigenbewegungen," *Regelungstechnik*, **28**(3), 80–86 (1979).

Liu, Y. and B. Anderson, "Singular perturbation approximation of balanced systems," *Int. J. Control*, **50**, 1379–1405 (1989).

Löffler, H. P. and W. Marquardt, "Order reduction for nonlinear differential-algebraic process models," *J. Proc. Cont.*, **1**, 32–40 (1991).

Lohmann, B. and W. Marquardt, "On the Systematization of the process of model development," *Comput. Chem. Eng.*, **21**(Suppl.), S213–S218 (1996).

Lohmann, B., "Ordnungsreduktion und Dominanzanalyse nichtlinearer Systeme," *Automatisierungstechnik*, **42**(10), 466–474 (1994).

Lohmann, B., "Application of model order reduction to a hydropneumatic vehicle suspension," *IEEE Trans. Cont. Sys. Tech.*, **3**(1), 102–109 (1995).

Lohmann, B., *Ansätze zur Unterstützung des Arbeitsablaufes bei der rechnerbasierten Modellierung verfahrenstechnischer Prozesse*. VDI Verlag (1998).

Lu, J. Z., Multi-zone control under enterprise optimization: Needs, challenges and requirements, In Allgöwer, F. and A. Zheng, editors, *Nonlinear Model Predictive Control*, pages 393–402. Birkhäuser, Basel (2000).

Maas, U. and S. Pope, "Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space," *Comb. & Flame*, **88**, 252–264 (1992).

Macchietto, S., E. H. Chimowitz, T. F. Anderson, and L. F. Stutzman, "Local models for representing phase equilibria in multicomponent nonideal vapor-liquid and liquid-liquid systems. 3. Parameter estimation and update," *Ind. Eng. Chem. Proc. Des. Dev.*, **25**, 674–682 (1986).

Marenbach, P., K. D. Bettenhausen, S. Freyer, U. Nieken, and H. Rettenmaier, Data-driven structured modelling of a biotechnological fed-batch fermentation by means of genetic programming, In *Proceedings of the Institution of Mechanical Engineers. Part I—J. Sys. Control Engng.*, volume 211, pages 325–332 (1997).

Marino, R. and P. V. Kokotovic, "A geometric approach to nonlinear singularly perturbed control systems," *Automatica*, **24**(1), 31–41 (1988).

Marlin, T. E. and A. N. Hrymak, "Real-time operations optimization of continuous processes," *AIChE Symp. Ser.*, **93**(316), 156–164 (1997).

Marquardt, W. and E. D. Gilles, Nonlinear wave phenoma as fundamentals for model-based control system design in distillation, AIChE Annual Meeting, Chicago, IL (1990).

Marquardt, W., "Traveling waves in chemical processes," *Int. Chem. Engng.*, **30**, 585–606 (1990).

Marquardt, W., Numerical methods for the simulation of differential-algebraic process models, In Berber, R., editor, *Methods of model based control*, volume 293 of *NATO-ASI Ser. E, Applied Sciences*, pages 42–79. Kluwer, Dordrecht (1995).

Martinson, W. S. and A. I. Barton, "A differentiation index for partial differential-algebraic equations.," *SIAM J. Sci. Comp.*, **21**(6), 2295–2315 (2000).

Massias, A., D. Diamantis, E. Mastrorakos, and D. A. Goussis, "An algorithm for the construction of global reduced mechanisms with CSP data," *Combustion and Flame*, **117**, 685–708 (1999).

Mayne, D. Q., J. B. Rawlings, C. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, **36**, 789–814 (2000).

McKay, B., M. Willis, and G. Barton, "Steady-state modelling of chemical process systems using genetic programming," *Comput. Chem. Eng.*, **21**(9), 981–996 (1997).

Mesarovic, M. D., D. Macko, and Y. Takahara, "Two coordination principles and their application in large scale systems control," *Automatica*, **6**, 261–270 (1970).

Molga, E. and R. Cherbański, "Hybrid first-principle—neural-network approach to modelling of the liquid-liquid reacting system," *Chem. Eng. Sci.*, **54**, 2467–2473 (1999).

Molga, E. and K. Westerterp, "Neural network based model of the kinetics of catalytic hydrogenation reactions," *Stud. Surf. Sci. Catal.*, **109**, 379–388 (1997).

Mönnigmann, M. and W. Marquardt, "Normal vectors on manifolds of critical points for parametric robustness of equilibrium solutions of ODE systems," *J. Nonlinear Sci.* (2000). submitted.

Moore, B. C., "Principal component analysis in linear systems: Controllabity, observability, and model reduction," *IEEE Trans. Auto. Cont.*, **26**(1), 17–32 (1981).

Morari, M. and J. H. Lee, "Model predictive control: past, present and future," *Comput. Chem. Eng.*, **23**, 667–682 (1999).

Morari, M., Y. Arkun, and G. Stephanopoulos, "Studies in the synthesis of control structures for chemical processes. Part I: Formulation of the control problem. Process decomposition and the classification of the control tasks. Analysis of the optimizing control structures," *AIChE J.*, **26**(2), 220–232 (1980).

Muscato, G., "Parametric generalized singular perturbation approximation for model order reduction," *IEEE Trans. Auto. Cont.*, **45**(2), 339–343 (2000).

Mustafa, D. and K. Glover, "Controller reduction by H_∞-balanced truncation," *IEEE Trans. Auto. Cont.*, **36**, 668–682 (1991).

Nevistić, V. and M. Morari, Constrained control of feedback-linearizable systems, In *Proceedings of Europ. Control Conf.*, pages 1726–1731, Rome (1995).

Newman, A. and P. S. Krishnaprasad, "Nonlinear model reduction for RTCVD," *IEEE Proc. 32nd Conf. Information Sciences and Systems, Princeton* (1998).

Norquay, S. J., A. Palazoglu, and J. A. Romagnoli, "Application of Wiener model predictive control (WMPC) to an industrial C2-splitter," *J. Proc. Cont.*, **9**, 461–473 (1999).

Pallaske, U., "Ein Verfahren zur Ordnungsreduktion mathematischer Prozeßmodelle," *Chem.-Ing.-Tech.*, **59**(7), 604–605 (1987).

Pavel, L. and F. W. Fairman, "Controller reduction for nonlinear plants: An L_2 approach," *Int. J. Rob. Nonl. Contr.*, **7**, 475–505 (1997).

Pearson, R. K. and B. A. Ogunnaike, Nonlinear process identification, In *Nonlinear process control*, pages 11–110. Prentice-Hall (1997).

Pearson, R. K. and M. Pottmann, "Gray-box identification of block-oriented nonlinear models," *J. Proc. Cont.*, **10**, 301–314 (2000).

Perregaard, J., "Model simplification and reduction for simulation and optimization of chemical processes," *Comput. Chem. Eng.*, **17**(5/6), 465–483 (1993).

Petzold, L. and W. Zhu, "Model reduction for chemical kinetics: an optimization approach," *AIChE J.*, **45**(4), 869–886 (1999).

Psichogios, D. C. and L. H. Ungar, "A hybrid neural-network first principles approach to process modeling," *AIChE J.*, **38**, 1499 (1992).

Qin, S. J. and T. J. Badgwell, An overview of industrial model predictive control technology, In *Chemical Process Control V*, volume 316,93 of *AIChE Symposium Series*, pages 232–256 (1996).

Qin, S. J. and T. J. Badgwell, An overview of nonlinear model predictive control applications, In Allgöwer, F. and A. Zheng, editors, *Nonlinear Model Predictive Control*. Birkhäuser, Basel, Switzerland (2000).

Ravindran, S. S., Proper orthogonal decomposition in optimal control of fluids, Technical Report NASA/TM-1999-209113, National Aeronautics and Space Administration NASA, Hampton, Virginia (USA) (1999).

Rawlings, J. B., "Tutorial overview of model predictive control," *IEEE Cont. Sys. Mag.*, **20**(3), 38–52 (2000).

Reuter, M., J. van Deventer, and P. van der Walt, "A generalized neural-net rate equation," *Chem. Eng. Sci.*, **48**, 1499 (1993).

Rhodes, C., M. Morari, and S. Wiggins, "Identification of low order manifolds: validating the algorithm of Maas and Pope," *CHAOS*, **9**(1), 108–123 (1999).

Robertson, G. and I. Cameron, "Analysis of dynamic models for structural insight and model reduction: Part 2. A multi-stage compressor shutdown case study," *Comput. Chem. Eng.*, **21**(5), 475–488 (1997a).

Robertson, G. and I. Cameron, "Analysis of dynamic process models for structural insight and model reduction: Part 1. Structural identification measures," *Comput. Chem. Eng.*, **21**(5), 455–473 (1997b).

Safavi, A. A., A. Nooraii, and J. A. Romangnoli, "A hybrid model formulation for a distillation column and the on-line optimisation study," *J. Proc. Cont.*, **9**, 125–134 (1999).

Samar, R., I. Postlethwaite, and D.-W. Gu, "Model reduction with balanced realizations," *Int. J. Control*, **62**(1), 33–64 (1995).

Saxén, B. and H. Saxén, "A neural-network based model of bioreaction kinetics," *Can. J. Chem. Eng.*, **74**(1), 124–131 (1996).

Scherpen, J. M. A. and A. J. van der Schaft, "Normalized coprime factorizations and balancing for unstable nonlinear systems," *Int. J. Control*, **60**(6), 1193–1222 (1994).

Scherpen, J. M. A., "Balancing for nonlinear systems," *Sys. Cont. Let.*, **21**, 143–153 (1993).

Scherpen, J. M. A., "H_∞-balancing for nonlinear systems," *Int. J. Rob. Nonl. Contr.*, **6**, 645–668 (1996).

Schuler, H. and C.-U. Schmidt, "Calorimetric state estimators for chemical reactor diagnosis and control: review of methods and applications," *Chem. Eng. Sci.*, **47**, 899–915 (1992).

Seigneur, C., G. Stephanopoulos, and R. W. Carr Jr., "Dynamic sensitivity analysis of chemical reaction systems: a variational method," *Chem. Eng. Sci.*, **17**, 845–853 (1982).

Shene, C., C. Diez, and S. Bravo, "Neural networks for the prediction of the state of Zymomonas mobilis CP4 batch fermentations," *Comput. Chem. Eng.*, **23**, 1097–1108 (1999).

Shvartsman, S. Y. and I. G. Kevrekidis, "Nonlinear model reduction for control of distributed systems: A computer-assisted study," *AIChE J.*, **44**(7), 1579–1595 (1998).

Shvartsman, S. Y., C. Theorodoropuolos, R. Rico-Martinez, I. G. Kevrekidis, and E. S. Titi, "Order reduction for nonlinear dynamic models of distributed reacting systems," *J. Proc. Cont.*, **10**, 177–184 (2000).

Simutis, R., R. Oliviera, M. Manikowski, S. Feyo de Azevedo, and A. Lübbert, "How to increase the performance of models for process optimization and control," *J. Biotech.*, **59**, 73–89 (1997).

Singh, M. D., *Dynamical Hierarchical Control*. North-Holland (1977).

Sirovich, L., "Turbulence and the dynamics of coherent structures: I, II and III," *Quart. Appl. Math.*, **XLV**(3), 561 (1987).

Skogestad, S. and I. Postlethwaite, *Multivariable Feedback Control*. John Wiley (1996).

Slaats, P. M. A., J. Jongh de, and A. A. H. J. Sauren, "Model reduction tools for nonlinear structural dynamics," *Computers & Structures*, **54**(6), 1155–1171 (1995).

Stewart, W. E., K. L. Levien, and M. Morari, "Simulation of fractionation by orthogonal collocation," *Chem. Eng. Sci.*, **40**, 409–421 (1985).

Stiharu-Alexe, I. and J. O'Shea, "Four-dimensional guidance of atmospheric vehicles," *J. Guidance, Control, and Dynamics*, **19**(1), 113–122 (1995).

Storen, S. and T. Hertzberg, "Local thermodynamic models used in sensitivity estimation of dynamic systems," *Comput. Chem. Eng.*, **21**(Suppl.), S709–S714 (1997).

Su, H. T., P. A. Bhat, P. A. Minderman, and T. J. McAvoy, "Integrating neural networks with first principles models for dynamic modeling," *IFAC Symp. Dynamics and Control of Chem. Reactors, Distillation Columns, and Batch Processes* (1992).

Tali-Maamar, N., M. T. Nihtila, and J. P. Babary, "On control of nonlinear distributed parameter bioreactors," *Math. Comp. Simul.*, **37**, 173–181 (1994).

Tatrai, F., P. Lant, P. Lee, I. Cameron, and R. Newell, "Model reduction for regulatory control: an FCCU case study," *Trans. Inst. Chem. Eng.*, **72**(5), 402–407 (1994).

te Braake, H. A. B., H. J. L. van Can, and H. B. Verbruggen, "Semi-mechanistic modeling of chemical processes with neural networks," *Artificial Intelligence*, **11**, 507–515 (1998).

Thibault, J., G. Acuna, R. Pérez-Correa, H. Jorquera, and P. Molin, "A hybrid representation approach for modelling complex dynamic bioprocesses," *Bioprocess Engineering*, **22**, 547–556 (2000).

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

Tsen, A. Y.-D., S. S. Jang, D. S. H. Wong, and J. Babu, "Predictive control of quality in batch polymerization using hybrid ANN models," *AIChE J.*, **42**(2), 455–465 (1996).

van Breusegem, V. and G. Bastin, "Reduced order dynamical modelling of reaction systems: a singular perturbation approach," *Proc. 30th Conf. Decision and Control, Brighton*, pages 1049–1054 (1991).

van Can, H. J. L., C. Hellinga, K. C. A. M. Luyben, J. J. Heijnen, and H. A. B. te Braake, "Strategy for dynamic process modeling based on neural networks in macroscopic balances," *AIChE J.*, **42**(12), 3403–3418 (1996).

van Can, H. J. L., H. A. B. te Braake, S. Dubbelmann, C. Hellinga, and K. C. A. M. Luyben, "Understanding and applying the extrapolation properties of serial gray-box models," *AIChE J.*, **44**(5), 1071–1089 (1998).

Villadsen, J. V. and M. L. Michelsen, *Solution of Differential Equation Models by Polynomial Approximation*. Prentice-Hall (1978).

von Watzdorf, R. and W. Marquardt, "Fully adaptive model size reduction for multicomponent separation problems," *Comput. Chem. Eng.*, **21**(Suppl.), S811–S816 (1997).

von Wedel, L. and W. Marquardt, "Cheops: A case study in component-based process simulation," *AIChE Symp. Ser.*, **96**(323), 494–497 (2000).

Wisnewski, P. A. and F. J. Doyle III, Nonlinear model reduction using Hankel-norm approximation, Technical report, University of Delaware (1996a).

Wisnewski, P. A. and F. J. Doyle III, "A reduced model approach to estimation and control of a Kamyr digester," *Comput. Chem. Eng.*, **20**(Suppl.), S1053–S1058 (1996b).

Wortelboer, P. M. R., M. Steinbuch, and O. H. Bosgra, "Iterative model and controller reduction using closed-loop balancing, with application to a compact disc mechanism," *Int. J. Robust and Nonlinear Control*, **9**(3), 123–142 (1999).

Zander, H.-J., R. Dittmeyer, and J. Wagenhuber, "Dynamic modeling of chemical reaction systems with neural networks and hybrid models," *Chem. Eng. Technol.*, **21**(7), 571–574 (1999).

Zbiciński, I., P. Strumiłło, and W. Kamiński, "Hybrid neural model of thermal drying in a fluidized bed," *Comput. Chem. Eng.*, **20**(Suppl.), S695–S700 (1996).

Zhou, K. M., C. D'Souza, and C. J. R., "Structurally balanced controller order reduction with guaranteed closed-loop performance," *Sys. Cont. Let.*, **24**(4), 235–242 (1995).