Extended Kalman Filter Based Nonlinear Model Predictive Control

Jay H. Lee*

Department of Chemical Engineering, Auburn University, Auburn, Alabama 36849-5127

N. Lawrence Ricker

Department of Chemical Engineering, University of Washington, Seattle, Washington 98195

This paper formulates a state observer based nonlinear model predictive control (MPC) technique using successive linearization. Based on local linear approximations of state/measurement equations computed at each sample time, a recursive state estimator providing the minimum-variance state estimates (known as the "extended Kalman filter (EKF)") is derived. The same local linear approximation of the state equation is used to develop an optimal prediction equation for the future states. The prediction equation is made linear with respect to the undecided control input moves by making linear approximations dual to those made for the EKF. As a result of these approximations, increase in the computational demand over linear MPC is quite mild. The prediction equation can be computed via noniterative nonlinear integration. Minimization of the weighted 2-norm of the tracking errors with various constraints can be solved via quadratic programming. Connections with previously published successive linearization based approaches of nonlinear quadratic dynamic matrix control are made. Under restrictive assumptions on the external disturbances and measurement noise, the proposed algorithm reduces to these techniques. Potential benefits, hazards and shortcomings of the proposed technique are pointed out using a control problem arising in a paper machine.

1. Introduction

Model predictive control (MPC) has emerged as a powerful tool for dynamic optimization and control. Although different in form, the underlying idea of all the available MPC schemes is the same and can be stated as follows:

1. A dynamic model and on-line measurements are used to build a prediction of future output behavior expressed in terms of current and future mainpulated input moves.
2. On the basis of the prediction, optimization is performed to find a sequence of input moves that minimizes a chosen measure of the output deviation from their respective reference values while satisfying all the given constraints.
3. Since the quality of prediction may improve as more measurements are collected, only the first of the calculated input sequences is implemented and the whole optimization is repeated at the next sampling time. This so called "receding horizon implementation" makes MPC a feedback control algorithm.

A key feature contributing to the success of MPC is that various process constraints can be incorporated directly into the on-line optimization performed at each time step. Because most of the challenging control problems found in the process industry involve multivariable systems with constraints for which no other effective control technique exists, MPC has been steadily gaining acceptance by the industry. Various versions of MPC based on the aforementioned principle have demonstrated their effectiveness in industrial applications (Richalet *et al.*, 1978; Culter and Ramaker, 1980; Rouhani and Mehra, 1982; Garcia and Morari, 1982; Cutler and Hawkins, 1988; also see Garcia *et al.* (1989) for review).

Although the initial versions of MPC (*e.g.*, dynamic matrix control by Culter and Ramaker (1980) or model algorithmic control by Rouhani and Mehra (1982)) were heuristic in nature and limited in generality, many extensions, modifications, and refinements have been proposed during the past decade. The abundance of batch processes and continuous processes with wide operating ranges has motivated the development of nonlinear MPC (NLMPC) techniques, which use nonlinear models for prediction. Proposed NLMPC techniques range from a simple extension of DMC based on successive linearization of the nonlinear model (nonlinear quadratic dynamic matrix control (NLQDMC) by Garcia (1984) and Gattu and Zafiriou (1992)) to more elaborate and computationally intensive techniques involving discretization of the model followed by solution via nonlinear programming (NLP) (see Biegler and Rawlings (1991) and Bequette (1991) for review). Among these techniques, only NLQDMC has seen any significant industrial application, mainly due to its mild computational requirement.

Another major refinement proposed for MPC is in the method for estimating the unmeasured disturbance effect. Most industrial MPC techniques rely on a heuristic *open-loop* prediction method in which state estimates are not corrected by measurements. Measurements are incorporated into the prediction equation in an *ad hoc* fashion, often as a constant bias term (corresponding to the difference between the current measurement and model output) added to the model prediction vector. While this open-loop method is adequate for many cases, it is restricted to open-loop stable processes and its *ad hoc* treatment of feedback measurements limits performance. In addition, such an approach is not directly applicable to inferential control problems in which the controlled variables must be estimated using secondary measurements. Hence, the recent trend has been to use the well-established *closed-loop* state estimation techniques like the Kalman filter for model prediction (Ricker, 1990; Lee *et al.*, 1992, 1994; Lee and Yu, 1994). Thus far, closed-loop state estimation has been addressed mainly in the context of linear MPC. On the other hand, for nonlinear MPC techniques, the advantage of closed-loop state

* To whom all correspondence should be addressed. E-mail: JHL@ENG.AUBURN.EDU. Phone: (205) 844-2060. Fax: (205) 844-2063.

0888-5885/94/2633-1530$04.50/0

estimation over open-loop state estimation stretches even beyond the increased applicability and superior disturbance estimation. For example, in the case of local linearization method, the quality of the linearized model depends critically on how accurately the states are estimated. Because open-loop estimation provides no feedback corrections to the state estimates, they can be significantly biased, leading to linearized models not properly representing the local dynamics.

The main objective of this article is to develop a computationally efficient nonlinear control algorithm by combining the main concepts of NLQDMC with an appropriate state estimation technique. More specifically, we integrate the successive linearization concept used in NLQDMC with a similar state estimation technique (*i.e.*, extended Kalman filter; Kopp and Oxford, 1963) to develop a prediction of future states that is amenable to computational efficiency. The prediction vector can be computed via nonlinear integration and is affine with respect to the undecided manipulated input moves, yielding an optimization solvable via quadratic programming (QP). While a more rigorous treatment of both the filtering and control problems is possible, significant theoretical and practical barriers weighed against often insignificant performance improvement do not justify such an approach at the present time. First, the theory of optimal filtering and prediction for general nonlinear systems with stochastic inputs is still in a developmental stage (see Jazwinski, 1970). Second, a fully nonlinear approach would lead to complex NLPs to be solved at each time step, both for the filtering and control problems, significantly increasing the on-line computational demand (see Jang *et al.* (1987), Sistu and Bequette (1991), and Robertson and Lee (1993) for some approaches in this direction). Finally, because MPC implements the control moves in a receding horizon fashion, the technique based on successive linearization should yield adequate performance for most cases.

We make connections between the proposed technique and previously published NLQDMC techniques by Garcia (1984) and Gattu and Zafiriou (1992). Although our work shares some similarity with NLQDMC of Gattu and Zafirou in that both incorporate state estimation into the original NLQDMC, there are some important differences as well. First, instead of the extended Kalman filter, they suggest using the *steady-state* Kalman filter which is redesigned at each time step on the basis of the locally linearized model. Our approach is more appropriate in the context of a time-varying system (approximating the nonlinear system). Second, the Kalman filter is designed with *white* state noise of a scalar-times-identity covariance matrix; therefore, the load disturbance effects must still be taken into account afterward, similarly as in traditional MPC. This can lead to significant biases in the state estimates, limiting its applicability and performance. In our work, we model load disturbances as (possibly nonstationary) stochastic processes added to the states and develop an *optimal* multistep prediction under local linearization approximations of the state and measurement equations. The advantages of our approach over NLQDMC are several-fold. First, it expands applicability to inferential control problems. Second, the quality of the linearized model should improve as the states are now estimated with the load disturbance effect fully taken into account. Finally, by developing an *optimal* predicton under clearly stated approximations, it is transparent when the proposed technique is expected to perform comparably to more elaborate techniques, and when it will exhibit substantially inferior performance.

We study the benefits and limitations of the proposed technique using a bilinear multivariable control problem arising in a paper machine. We also study the effect of model errors on the performance of the algorithm and practical ways to alleviate these problems. The application study should also provide a tutorial of the technique and some valuable insights needed for successful application.

2. Extended Kalman Filter Based NLMPC: Algorithm Derivation and Implementation

2.1. Model. We assume that our model is expressed through the following nonlinear differential equation:

$$\dot{x} = f(x,u,d) \tag{1}$$

$$y = g(x,d) \tag{2}$$

x is the state vector, u is the manipulated input vector, d is the unmeasured disturbance vector, and y is the measured output vector. It is assumed that dynamics from u to y have relative degree of at least 1. For digital controller design, u and d can be assumed to be constant between the sampling instants. Hence, we can express a discrete version of the model (1) and (2) as follows:

$$x_k = F_{t_s}(x_{k-1},u_{k-1},d_{k-1}) \tag{3}$$

$$y_k = g(x_k,d_k) \tag{4}$$

where $F_{t_s}(x_{k-1},u_{k-1},d_{k-1})$ denotes the terminal state vector obtained by integrating the ordinary differential equation (ODE) (1) for one sample interval (t_s) with the initial condition of x_{k-1} and constant inputs of $u = u_{k-1}$ and $d = d_{k-1}$. In general, F_{t_s} cannot be written in closed form.

For the purpose of state estimation, it is common to express the unmeasured disturbance signal d as a stochastic process. Without loss of generality, we assume that d is generated through the following stochastic difference equation:

$$x_k^w = A^w x_{k-1}^w + B^w w_{k-1} \tag{5}$$

$$d_k = C^w x_k^w \tag{6}$$

w_k is discrete-time white noise with covariance R^w.

It is also possible that the measurements of y_k are corrupted by measurement noise ν_k as follows:

$$\hat{y}_k = g(x_k,d_k) + \nu_k \tag{7}$$

We assume in this study that ν_k is white noise with covariance of R^ν. Extensions of the subsequently introduced estimation technique to the case of nonwhite measurement noise is straightforward and can be found in standard textbooks (*e.g.*, Goodwin and Sin, 1984).

Combining (3) and (4) with (5)–(7), we arrive at the following augmented model:

$$\begin{bmatrix} x_k \\ x_k^w \end{bmatrix} = \begin{bmatrix} F_{t_s}(x_{k-1},u_{k-1},C^w x_{k-1}^w) \\ A^w x_{k-1}^w \end{bmatrix} + \begin{bmatrix} 0 \\ B^w \end{bmatrix} w_{k-1} \tag{8}$$

$$\hat{y}_k = g(x_k,C^w x_k^w) + \nu_k \tag{9}$$

From this point on, our discussion will be based on the augmented form of the model.

2.2. State Estimation: Extended Kalman Filtering. A straightforward extension of the optimal linear filter ("Kalman filter") is the extended Kalman filter first proposed by Kopp and Oxford (1963) and Cox (1964). The basic idea of EKF is to perform linearization at each time step to approximate the nonlinear system as a time-varying system affine in the variables to be estimated, and apply the linear filtering theory to it.

Let us consider the model of (8) and (9). We will use the notation $x_{k|l}$ and $x^w_{k|l}$ to denote the optimal estimates (*i.e.*, minimum variance estimates) for x_k and x^w_k based on the measurements up to time l. In probabilistic terms, they represent the conditional expectation of the Gaussian variables x_k and x^w_k with the conditions given by the measurements $\hat{y}_1$, ..., $\hat{y}_l$. We may also express the confidence in these estimates through the conditional covariance matrix $\Sigma_{k|l}$, *i.e.*,

$$\Sigma_{k|l} = E\left\{\begin{bmatrix} x_k - x_{k|l} \\ x^w_k - x^w_{k|l}\end{bmatrix}\begin{bmatrix} x_k - x_{k|l} \\ x^w_k - x^w_{k|l}\end{bmatrix}^{\mathrm{T}}\right\} \qquad (10)$$

The problem of recursive state estimation can be viewed as computing the new estimates $x_{k|k}$ and $x^w_{k|k}$ and covariance matrix $\Sigma_{k|k}$ based on the previous estimates $x_{k-1|k-1}$ and $x^w_{k-1|k-1}$, their covariance matrix $\Sigma_{k-1|k-1}$, and new measurement $\hat{y}_k$. It is instructive to view this as two subproblems: "model prediction" and "measurement correction".

2.2.1. Model Prediction. In the model prediction step of state estimation, the objective is to compute the estimates of the new states without using the new information of $\hat{y}_k$, *i.e.*,

Compute $x_{k|k-1}$, $x^w_{k|k-1}$, and $\Sigma_{k|k-1}$ from $x_{k-1|k-1}$, $x^w_{k-1|k-1}$, and $\Sigma_{k-1|k-1}$.

This is in general a very difficult problem as the assumed normal distributions of the initial state variables are destroyed by propagation through nonlinear dynamics. Extended Kalman filter simplifies the problem significantly by making the following linear approximation of (8) with respect to $x_{k-1} = x_{k-1|k-1}$ and $x^w_{k-1} = x^w_{k-1|k-1}$:

$$\begin{bmatrix} x_k \\ x^w_k\end{bmatrix} \approx \begin{bmatrix} F_{t_s}(x_{k-1|k-1},u_{k-1},C^w x^w_{k-1|k-1}) \\ A^w x^w_{k-1|k-1}\end{bmatrix} + \Phi_{k-1}\begin{bmatrix} x_{k-1} - x_{k-1|k-1} \\ x^w_{k-1} - x^w_{k-1|k-1}\end{bmatrix} + \Gamma^w w_{k-1} \qquad (11)$$

where

$$\Gamma^w = \begin{bmatrix} 0 \\ B^w\end{bmatrix}$$

and Φ_k is calculated using the following formula:

$$\Phi_{k-1} = \begin{bmatrix} A_{k-1} & B^d_{k-1}C^w \\ 0 & A^w\end{bmatrix} \qquad (12)$$

$$A_{k-1} = \exp(\tilde{A}_{k-1}t_s);\quad \tilde{A}_{k-1} = \frac{\partial f(x,u,d)}{\partial x}\Big|(x=x_{k-1|k-1},u=u_{k-1},d=C^w x^w_{k-1|k-1}) \qquad (13)$$

$$B^d_{k-1} = \int_0^{t_s}\exp(\tilde{A}_{k-1}\tau)\,\mathrm{d}\tau\cdot\tilde{B}^d_{k-1};\quad \tilde{B}^d_{k-1} = \frac{\partial f(x,u,d)}{\partial d}\Big|(x=x_{k-1|k-1},u=u_{k-1},d=C^w x^w_{k-1|k-1}) \qquad (14)$$

$\partial f/\partial x$ and $\partial f/\partial d$ represent Jacobian matrices of f with respect to x and d, respectively. For example, $\tilde{A}_{k-1}$ is the Jacobian matrix for $f(x,u,d)$ with respect to x evaluated at $x = x_{k-1|k-1}$, $u = u_{k-1}$, and $d = C^w x^w_{k-1|k-1}$. Note that x_k and x^w_k are Gaussian variables because of the linear approximation of (11).

The expressions for the "sensitivity matrices" A_{k-1} and B^d_{k-1} given in (13) and (14) deserve further comments. We adopt these expressions in this paper because they are simple to compute and work well for most cases. However, they are valid only under the assumption that the Jacobian matrices $\partial f/\partial x$ and $\partial f/\partial d$ remain constant throughout the time period between $t = k - 1$ and $t = k$. In fact, these matrices continuously evolve with x during the time interval according to the ODE (1). Explicit expressions for the exact values of these matrices (without the assumption) do not exist; instead, they must be obtained by solving matrix differential equations (see Li and Biegler (1989) for details). However, such complication is often unnecessary (and almost never found in the extended Kalman filter literature) since the Jacobian matrices change very little for a reasonably chosen sample interval. Hence, (13) and (14) are often very good approximations of these exact values.

Let e_{k-1} and e^w_{k-1} represent $x_{k-1} - x_{k-1|k-1}$ and $x^w_{k-1} - x^w_{k-1|k-1}$, respectively. Then,

$$e_{k-1|k-1} = 0;\quad e^w_{k-1|k-1} = 0;\quad w_{k-1|k-1} = 0 \qquad (15)$$

since $x_{k-1|k-1}$ and $x^w_{k-1|k-1}$ represent the optimal estimates (conditional means of x_{k-1} and x^w_{k-1}) and w_{k-1} is zero-mean white noise, the effect of which does not appear in the measurements up to $\hat{y}_{k-1}$.

Hence, for the approximate system (11),

$$\begin{bmatrix} x_{k|k-1} \\ x^w_{k|k-1}\end{bmatrix} = \begin{bmatrix} F_{t_s}(x_{k-1|k-1},u_{k-1},C^w x^w_{k-1|k-1}) \\ A^w x^w_{k-1|k-1}\end{bmatrix} \qquad (16)$$

In addition, since (11) can be rewritten as

$$\begin{bmatrix} x_k \\ x^w_k\end{bmatrix} \approx \begin{bmatrix} x_{k|k-1} \\ x^w_{k|k-1}\end{bmatrix} + \Phi_k\begin{bmatrix} x_{k-1} - x_{k-1|k-1} \\ x^w_{k-1} - x^w_{k-1|k-1}\end{bmatrix} + \Gamma^w w_{k-1} \qquad (17)$$

$\Sigma_{k|k-1}$, the covariance matrix for the error

$$\begin{bmatrix} x_{k|k-1} - x_k \\ x^w_{k|k-1} - x^w_k\end{bmatrix}$$

can be easily computed from $\Sigma_{k-1|k-1}$ as follows:

$$\Sigma_{k|k-1} = \Phi_{k-1}\Sigma_{k-1|k-1}\Phi^{\mathrm{T}}_{k-1} + \Gamma^w R^w(\Gamma^w)^{\mathrm{T}} \qquad (18)$$

2.2.2. Measurement Correction. In the measurement correction stage of state estimation, the new information $\hat{y}_k$ is used to improve the state estimates. The problem can be formally stated as

Compute $x_{k|k}$, $x^w_{k|k}$, and $\Sigma_{k|k}$ from $x_{k|k-1}$, $x^w_{k|k-1}$, and $\Sigma_{k|k-1}$ and measurement $\hat{y}_k$.

Because $\hat{y}_k$ is a nonlinear function of the states, the above is a nonlinear estimation problem that does not yield an analytical solution in general. In the EKF, the problem is again simplified by linearizing the output equation (9) at $x_{k|k-1}$ and $x^w_{k|k-1}$:

$$\hat{y}_k \approx g(x_{k|k-1},C^w x^w_{k|k-1}) + \Xi_k\begin{bmatrix} x_k - x_{k|k-1} \\ x^w_k - x^w_{k|k-1}\end{bmatrix} + \nu_k \qquad (19)$$

where

$$\Xi_k = [C_k \quad C^d_k\cdot C^w] \qquad (20)$$

and C_k and C^d_k are Jacobian matrices defined as follows:

$$C_k = \frac{\partial g(x,d)}{\partial x}\Big|(x=x_{k|k-1},d=C^w x^w_{k|k-1}) \qquad (21)$$

$$C^d_k = \frac{\partial g(x,d)}{\partial d}\Big|(x=x_{k|k-1},d=C^w x^w_{k|k-1}) \qquad (22)$$

Since we are given the means and covariances of the stochastic variables $x_k = x_{k|k-1}$, $x_k^w - x_{k|k-1}^w$, and ν_k, the linear filtering theory can be applied to find the conditional means $x_{k|k}$ and $x_{k|k}^w$ with the measurement condition (19);

$$\begin{bmatrix} x_{k|k} \\ x_{k|k}^w \end{bmatrix} = \begin{bmatrix} x_{k|k-1} \\ x_{k|k-1}^w \end{bmatrix} + L_k(\hat{y}_k - g(x_{k|k-1}, C^w x_{k|k-1}^w)) \quad (23)$$

where

$$L_k = \Sigma_{k|k-1}\Xi_k^T(\Xi_k \Sigma_{k|k-1} \Xi_k^T + R^\nu)^{-1} \quad (24)$$

In addition, the conditional covariance matrix $\Sigma_{k|k}$ expressing the confidence of the corrected state estimates is

$$\Sigma_{k|k} = (I - L_k \Xi_k)\Sigma_{k|k-1} \quad (25)$$

2.2.3. Implementation. In summary, under the assumption that the nonlinear system (8) and (9) is well approximated by the affine system of (11) and (19) obtained via local linearization, the following Kalman filter provides the optimal estimates:

Model Prediction:

$$\begin{bmatrix} x_{k|k-1} \\ x_{k|k-1}^w \end{bmatrix} = \begin{bmatrix} F_{t_s}(x_{k-1|k-1}, u_{k-1}, C^w x_{k-1|k-1}^w) \\ A^w x_{k-1|k-1}^w \end{bmatrix} \quad (26)$$

Measurement Correction:

$$\begin{bmatrix} x_{k|k} \\ x_{k|k}^w \end{bmatrix} = \begin{bmatrix} x_{k|k-1} \\ x_{k|k-1}^w \end{bmatrix} + L_k(\hat{y}_k - g(x_{k|k-1}, C^w x_{k|k-1}^w)) \quad (27)$$

where

$$L_k = \Sigma_{k|k-1}\Xi_k^T(\Xi_k \Sigma_{k|k-1} \Xi_k^T + R^\nu)^{-1} \quad (28)$$

$$\Sigma_{k|k-1} = \Phi_{k-1}\Sigma_{k-1|k-1}\Phi_{k-1}^T + \Gamma^w R^w (\Gamma^w)^T \quad (29)$$

$$\Sigma_{k|k} = (I - L_k \Xi_k)\Sigma_{k|k-1} \quad (30)$$

Note that the model update equation (26) requires nonlinear integration of ODE (1) with known initial condition and constant inputs.

For control applications with relatively short sample intervals, the required computation time for (27) and control move computation may be comparable to the sampling time. Then, $\hat{y}_k$ may not be used for computing the estimate of x_k under closed-loop conditions. In this case, the EKF can be implemented as an estimator where the measurement correction step precedes the model prediction step.

2.3. Prediction. In order to implement a predictive control algorithm, long-term prediction of the key states is required. Clearly, since the underlying system is nonlinear, the future states (and hence the outputs) are related to the current states and current/future inputs in a nonlinear fashion. This makes the problem of finding the optimal input sequence a complex nonlinear optimization problem. We propose to make the relationship linear via local linearization, which is dual to the local linearization used for deriving the extended Kalman filter.

2.3.1. One-Step Ahead Prediction.

$$\begin{bmatrix} x_{k+1} \\ x_{k+1}^w \end{bmatrix} = \begin{bmatrix} F_{t_s}(x_k, u_k, C^w x_k^w) \\ A^w x_k^w \end{bmatrix} + \begin{bmatrix} 0 \\ B^w \end{bmatrix} w_k \quad (31)$$

Based on the approximation (11) that we used for deriving the EFK,

$$\begin{bmatrix} x_{k+1|k} \\ x_{k+1|k}^w \end{bmatrix} = \begin{bmatrix} F_{t_s}(x_{k|k}, u_k, C^w x_{k|k}^w) \\ A^w x_{k|k}^w \end{bmatrix} \quad (32)$$

Note that since $F_{t_s}(x_{k|k}, u_k, C^w x_{k|k}^w)$ represents the solution obtained by intergating the ODE $\dot{x} = f(x,u,d)$ for one sample interval, the state $x_{k+1|k}$ is related to the *undecided* manipulated input u_k through nonlinear integration. This makes the optimization required for input move computation a nonlinear problem.

In order to prevent the on-line computational requirement of the algorithm from becoming unwieldy, we further approximate the equation by linearizing $F_{t_s}(x_{k|k}, u_k, C^w x_{k|k}^w)$ at the last known input value u_{k-1}:

$$\begin{bmatrix} x_{k+1|k} \\ x_{k+1|k}^w \end{bmatrix} \approx \begin{bmatrix} F_{t_s}(x_{k|k}, u_{k-1}, C^w x_{k|k}^w) \\ A^w x_{k|k}^w \end{bmatrix} + \begin{bmatrix} \mathcal{B}_k^u \\ 0 \end{bmatrix} (u_k - u_{k-1}) \quad (33)$$

where $\mathcal{B}_k^u$ is calculated using the following formula:

$$\mathcal{B}_k^u = \int_0^{t_s} \exp(\tilde{\mathcal{A}}_k \tau)\, d\tau \cdot \tilde{\mathcal{B}}_k^u \quad (34)$$

$$\tilde{\mathcal{B}}_k^u = \frac{\partial f(x,u,d)}{\partial u}\Big|(x = x_{k|k}, u = u_{k-1}, d = C^w x_{k|k}^w)$$

$$\tilde{\mathcal{A}}_k = \frac{\partial f(x,u,d)}{\partial x}\Big|(x = x_{k|k}, u = u_{k-1}, d = C^w x_{k|k}^w)$$

The above linearization can be considered as a dual approximation to (19), which was the linearization of the measurement equation made in order to make the filtering problem linear. For future use, we will also define the corresponding discrete state transition matrix $\mathcal{A}_k$ as follows:

$$\mathcal{A}_k = \exp(\tilde{\mathcal{A}}_k t_s) \quad (35)$$

2.3.2. Generalization to Multistep Prediction. We can generalize this idea to develop multistep prediction. Note from (32) that

$$\begin{bmatrix} x_{k+2|k} \\ x_{k+2|k}^w \end{bmatrix} = \begin{bmatrix} F_{t_s}(x_{k+1|k}, u_{k+1}, C^w x_{k+1|k}^w) \\ A^w x_{k+1|k}^w \end{bmatrix} \quad (36)$$

Note that $x_{k+2|k}$ is related in a nonlinear fashion not only to u_{k+1}, but also to u_k appearing in expression (33) for $x_{k+1|k}$. By appropriate linearization, we would like to derive an approximation that is linear with respect to the *undecided* inputs u_k and u_{k+1}. First, we note that $x_{k+1|k}^w = A^w x_{k|k}$ and that $F_{t_s}(x_{k|k}, u_{k-1}, C^w x_{k|k}^w)$ represents $x_{k+1|k}$ when u_k is kept same as the previous value u_{k-1}. In the absence of any information on how u_k is likely to change from u_{k-1}, the linear relationship that best approximates the local behavior can be obtained by linearizing the expression $F_{t_s}(x_{k+1|k}, u_{k+1}, C^w A^w x_{k|k}^w)$ with respect to $x_{k+1|k} = F_{t_s}(x_{k|k}, u_{k-1}, C^w x_{k|k}^w)$ and $u_{k+1} = u_{k-1}$ as follows:

$$\begin{aligned} F_{t_s}(x_{k+1|k}, u_{k+1}, C^w x_{k+1|k}^w) \approx \\ F_{t_s}\{F_{t_s}(x_{k|k}, u_{k-1}, C^w x_{k|k}^w), u_{k-1}, C^w A^w x_{k|k}^w\} + \\ \mathcal{A}_{k+1|k}\{x_{k+1|k} - F_{t_s}(x_{k|k}, u_{k-1}, C^w x_{k|k}^w)\} + \\ \mathcal{B}_{k+1|k}^u (u_{k+1} - u_{k-1}) \end{aligned} \quad (37)$$

where

$$\mathcal{A}_{k+1|k} = \exp(\tilde{\mathcal{A}}_{k+1|k} t_s) \quad (38)$$

$$\mathcal{B}^u_{k+1|k} = \int_0^{t_s} \exp(\tilde{\mathcal{A}}_{k+1|k}\tau)\, d\tau \cdot \tilde{\mathcal{B}}^u_{k+1|k} \quad (39)$$

$$\tilde{\mathcal{A}}_{k+1|k} = \frac{\partial f(x,u,d)}{\partial x}\Big|(x=F_{t_s}(x_{k|k},u_{k-1},C^w x^w_{k|k}),u=u_{k-1},d=C^w x^w_{k|k})$$

$$\tilde{\mathcal{B}}^u_{k+1|k} = \frac{\partial f(x,u,d)}{\partial u}\Big|(x=F_{t_s}(x_{k|k},u_{k-1},C^w x^w_{k|k}),u=u_{k-1},d=C^w x^w_{k|k})$$

The linearization with respect to $x_{k+1|k} = F_{t_s}(x_{k/k},u_{k-1},C^w x^w_{k|k})$ is to be considered as dual to (11), while that with respect to $u_{k+1} = u_{k-1}$ is to be considered as dual to (19).

Noting from (33) that $x_{k+1|k} - F_{t_s}(x_{k|k},u_{k-1},C^w x^w_{k|k}) \approx \mathcal{B}^u_k(u_k - u_{k-1})$ and $x^w_{k+1|k} = A^w x^w_{k|k}$ and substituting the affine approximation (37) into (36), we obtain

$$\begin{bmatrix} x_{k+2|k} \\ x^w_{k+2|k} \end{bmatrix} \approx \begin{bmatrix} F_{t_s}\{F_{t_s}(x_{k|k},u_{k-1},C^w x^w_{k|k}),u_{k-1},C^w A^w x^w_{k|k}\} \\ (A^w)^2 x^w_{k|k} \end{bmatrix} + [A_{k+1|k}\mathcal{B}^u_k \quad \mathcal{B}^u_{k+1|k}] \begin{bmatrix} u_k - u_{k-1} \\ u_{k+1} - u_{k-1} \end{bmatrix} \quad (40)$$

Carrying on the same derivation for $x_{k+l|k}$, $1 \le l \le p$, we obtain

$$\begin{bmatrix} x_{k+l|k} \\ x^w_{k+l|k} \end{bmatrix} \approx \begin{bmatrix} F_{l\cdot t_s}(x_{k|k},u_{k-1},C^w(A^w)^i x^w_{k|k}|_{0\le i\le l-1}) \\ (A^w)^l x^w_{k|k} \end{bmatrix} + \begin{bmatrix} \prod_{j=1}^{l-1}\mathcal{A}_{k+j|k}\mathcal{B}^u_k & \prod_{j=2}^{l-1}\mathcal{A}_{k+j|k}\mathcal{B}^u_{k+1|k} & \cdots & \mathcal{A}_{k+l-1|k}\mathcal{B}^u_{k+l-2|k} & \mathcal{B}^u_{k+l-1|k} \\ 0 \end{bmatrix} \times \begin{bmatrix} u_k - u_{k-1} \\ u_{k+1} - u_{k-1} \\ \vdots \\ u_{k+l-2} - u_{k-1} \\ u_{k+l-1} - u_{k-1} \end{bmatrix} \quad (41)$$

where

$$\mathcal{A}_{k+j|k} = \exp(\tilde{\mathcal{A}}_{k+j|k} t_s) \quad (42)$$

$$\mathcal{B}^u_{k+j|k} = \int_0^{t_s} \exp(\tilde{\mathcal{A}}_{k+j|k}\tau)\, d\tau \cdot \tilde{\mathcal{B}}^u_{k+1|k} \quad (43)$$

$$\tilde{\mathcal{A}}_{k+j|k} = \frac{\partial f(x,u,d)}{\partial x}\Big|(x=F_{j\cdot t_s}(x_{k|k},u_{k-1},C^w(A^w)^i x^w_{k|k}|_{0\le i\le j-1}), u=u_{k-1},d=C^w(A^w)^j x^w_{k|k})$$

$$\tilde{\mathcal{B}}^u_{k+j|k} = \frac{\partial f(x,u,d)}{\partial u}\Big|(x=F_{j\cdot t_s}(x_{k|k},u_{k-1},C^w(A^w)^i x^w_{k|k}|_{0\le i\le j-1}), u=u_{k-1},d=C^w(A^w)^j x^w_{k|k})$$

$F_{l\cdot t_s}(x_{k|k},u_{k-1},C^w(A^w)^i x^w_{k|k}|_{0\le i\le l-1})$ represents the terminal states obtained by integrating $\dot{x} = f(x,u,d)$ for l sampling intervals with initial condition $x_{k|k}$, constant input $u = u_{k-1}$ and piecewise constant input d taking the value of $C^w(A^w)^i x^w_{k|k}$ during the time interval $[k + i, k + i + 1)$.

Note that the expression (41) requires integration of the ODE (1) for l sample time steps into the future and computation of the sensitivity matrices for each of the l sample times (*i.e.*, $\mathcal{A}_{k+j|k}$ and $\mathcal{B}^u_{k+j|k}$ for $0 \le j \le l-1$). To reduce the computational complexity, the sensitivity matrices $\mathcal{A}_{k+j|k}$ and $\mathcal{B}^u_{k+j|_k}$ are often kept constant at the initial values of $\mathcal{A}_k$ and $\mathcal{B}^u_k$ throughout the prediction horizon (Garcia (1984) and Gattu and Zafiriou (1992)). The idea is that, because of the receding horizon implementation, more errors can be tolerated as we move away from the current time into the future in the prediction horizon. To avoid the notational complexity, we will also adapt this simplification in this paper. Hence, (41) simplifies to

$$\begin{bmatrix} x_{k+l|k} \\ x^w_{k+l|k} \end{bmatrix} \approx \begin{bmatrix} F_{l\cdot t_s}(x_{k|k},u_{k-1},C^w(A^w)^i x^w_{k|k}|_{0\le i\le l-1}) \\ (A^w)^l x^w_{k|k} \end{bmatrix} + \begin{bmatrix} \mathcal{A}_k^{l-1}\mathcal{B}^u_k & \mathcal{A}_k^{l-2}\mathcal{B}^u_k & \cdots & \mathcal{A}_k\mathcal{B}^u_k & \mathcal{B}^u_k \\ 0 \end{bmatrix} \begin{bmatrix} u_k - u_{k-1} \\ u_{k+1} - u_{k-1} \\ \vdots \\ u_{k+l-2} - u_{k-1} \\ u_{k+l-1} - u_{k-1} \end{bmatrix} \quad (44)$$

However, we emphasize that one can always choose to use the more consistent expression (41) if the increased computational load involving sensitivity matrix computation can be tolerated.

2.3.3. Linear Prediction Equation for Controlled Variables. Now let us assume that the controlled variable y^c is expressed as

$$y^c_k = h(x_k,d) \quad (45)$$

In order to develop a prediction equation for y^c that is linear with respect to the undecided input moves, we again linearize the above expression with respect to $x_{k|k}$:

$$y^c_k \approx h(x_{k|k},C^w x^w_{k|k}) + [H_k \quad H^d_k\cdot C^w] \begin{bmatrix} x_k - x_{k|k} \\ x^w_k - x^w_{k|k} \end{bmatrix} \quad (46)$$

where H_k and H^d_k are Jacobian matrices defined as

$$H_k = \frac{\partial h(x,d)}{\partial x}\Big|(x=x_{k|k},d=C^w x^w_{k|k}) \quad (47)$$

$$H^d_k = \frac{\partial h(x,d)}{\partial d}\Big|(x=x_{k|k},d=C^w x^w_{k|k})\cdot C^w \quad (48)$$

Since the last two terms in the right-hand side have an expectation of zero

$$y^c_{k|k} = h(x_{k|k},C^w x^w_{k|k}) \quad (49)$$

Continuing on with the same idea,

$$y^c_{k+l|k} = h(x_{k+l|k},C^w x^w_{k+l|k}) \approx h(x_{k|k},C^w x^w_{k|k}) + [H_k \quad H^d_k\cdot C^w] \begin{bmatrix} x_{k+l|k} - x_{k|k} \\ x^w_{k+l|k} - x^w_{k|k} \end{bmatrix} \quad (50)$$

Combining (50) with the optimal multistep prediction equation (44) we developed for x_{k+l} and substituting $u_{k+l} - u_{k-1}$ with $\sum_{i=0}^{l}\Delta u_{k+i}$ (where $\Delta u_{k+i} = u_{k+i} - u_{k+i-1}$), we obtain the following prediction equation for y^c:

$$\begin{bmatrix} y^c_{k+1|k} \\ y^c_{k+2|k} \\ \vdots \\ y^c_{k+p|k} \end{bmatrix} = \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix} (h(x_{k|k},C^w x^w_{k|k}) - H_k x_{k|k} - H^d_k C^w x^w_{k|k}) + \begin{bmatrix} H_k F_{t_s}(x_{k|k},u_{k-1},C^w x^w_{k|k}) \\ H_k F_{2\cdot t_s}(x_{k|k},u_{k-1},C^w (A^w)^i x^w_{k|k}|_{0\le i\le 1}) \\ \vdots \\ H_k F_{p\cdot t_s}(x_{k|k},u_{k-1},C^w (A^w)^i x^w_{k|k}|_{0\le i\le p-1}) \end{bmatrix} + \begin{bmatrix} H^d_k C^w A^w \\ H^d_k C^w (A^w)^2 \\ \vdots \\ H^d_k C^w (A^w)^p \end{bmatrix} x^w_{k|k} + \begin{bmatrix} H_k \mathcal{B}^u_k & 0 & \cdots & 0 \\ H_k(\mathcal{A}_k \mathcal{B}^u_k + \mathcal{B}^u_k) & H_k \mathcal{B}^u_k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma^{p-1}_{i=0} H_k \mathcal{A}^i_k \mathcal{B}^u_k & \cdots & \cdots & H_k \mathcal{B}^u_k \end{bmatrix} \begin{bmatrix} \Delta u_k \\ \Delta u_{k+1} \\ \vdots \\ \Delta u_{k+p-1} \end{bmatrix} \quad (51)$$

Note that $F_{l\cdot t_s}(x_{k|k},u_{k-1},C^w(A^w)^i x^w_{k|k}|_{0\le i\le l-1})$, $1 \le l \le p$ can be computed recursively since

$$F_{l\cdot t_s}(x_{k|k},u_{k-1},C^w(A^w)^i x^w_{k|k}|_{0\le i\le l-1}) = F_{t_s}(F_{(l-1)\cdot t_s}(x_{k|k},u_{k-1}, C^w(A^w)^i x^w_{k|k}|_{0\le i\le l-2},u_{k-1},C^w(A^w)^{l-1}x^w_{k|k})) \quad (52)$$

Note also that the first term of the right-hand side drops out if the controlled variable vector y^c consists of linear combinations of the state and disturbance vectors (*i.e.*, $y^c_k = H_k x_k + H^d_k d_k$).

In order to keep the notation simple, we will denote the above prediction equation as

$$\mathcal{Y}_{k+1|k} = \mathcal{I}(h(x_{k|k},C^w x^w_{k|k}) - H_k x_{k|k} - H^d_k C^w x^w_{k|k}) + \mathcal{S}^x_k(x_{k|k},u_{k-1},x^w_{k|k}) + \mathcal{S}^{x^w}_k x^w_{k|k} + \mathcal{S}^{\mathcal{U}}_k \Delta\mathcal{U}_k \triangleq \mathcal{Y}^0_{k+1|k} + \mathcal{S}^{\mathcal{U}}_k \Delta\mathcal{U}_k \quad (53)$$

Note that the term $\mathcal{Y}^0_{k+1|k}$ can be computed from the state estimates (provided by EKF) by performing integration of the nonlinear ODE and calculating the Jacobian matrices H_k and H^d_k. $\mathcal{S}^{\mathcal{U}}_k$ is a matrix that must be recomputed at each time step based on the updated Jacobian matrices $\mathcal{A}_k$ and $\mathcal{B}^u_k$. Note that this is usually the most time-consuming step.

2.4. Control. On the basis of the prediction equation (53), a sequence of optimal control moves can be computed, which minimizes a chosen norm of the expected future error. The following quadratic minimization has been by far the most popular choice (*e.g.*, QDMC by Garcia and Morshedi (1984)):

$$\min_{\Delta\mathcal{U}_k} \|\Lambda^y[\mathcal{Y}_{k+1|k} - R_{k+1|k}]\|_2^{\,2} + \|\Lambda^u \Delta\mathcal{U}_k\|_2^{\,2} \quad (54)$$

where $\mathcal{Y}_{k+1|k}$ is related to $\Delta\mathcal{U}_k$ linearly through the prediction equation (53). $R_{k+1|k} = [r^T_{k+1|k}, ..., r^T_{k+p|k}]^T$ is the future reference vector for y^c available at time k. Λ^y and Λ^u are weighting matrices that are chosen as (block) diagonal matrices in most cases. Note that, because the variance of the estimates for y^c_{k+i} are unaffected by the choice of $\Delta\mathcal{U}_k$, (54) is equivalent to the following minimum variance control criterion:

$$\min_{\Delta\mathcal{U}_k} E\{\|\Lambda^y[\mathcal{Y}_{k+1} - R_{k+1|k}]\|_2^{\,2}\} + \|\Lambda^u \Delta\mathcal{U}_k\|_2^{\,2} \quad (55)$$

Sometimes, the number of input moves computed in the optimization is chosen to be smaller than p in order to reduce the computational load and increase flexibility in tuning. This is called "blocking" and most commonly the last few (say $p - m$) input moves are set to zero *a priori*. In this case, the matrix $\mathcal{S}^{\mathcal{U}}_k$ in the prediction equation (53) is replaced by $\mathcal{S}^{\mathcal{U}_m}_k$ that consists of the first $m \times n_u$ columns of $\mathcal{S}^{\mathcal{U}}_k$. To the above minimization objective, the following input magnitude, rate and output magnitude constraints may be added:

$$u^{\text{low}}_{k+l} \le u_{k+l} \le u^{\text{high}}_{k+l}, \quad 0 \le l \le m-1 \quad (56)$$

$$-\Delta u^{\max}_{k+l} \le \Delta u_{k+l} \le \Delta u^{\max}_{k+l} \quad 0 \le l \le m-1 \quad (57)$$

$$y^{\text{low}}_{k+l} \le y^c_{k+l} \le y^{\text{high}}_{k+l}, \quad 1 \le l \le p \quad (58)$$

In the absence of constraints (56)–(58), the solution to (54) can be derived analytically:

$$\Delta\mathcal{U}(k) = \{(\mathcal{S}^{\mathcal{U}_m}_k)^T(\Lambda^y)^T\Lambda^y\mathcal{S}^{\mathcal{U}_m}_k + (\Lambda^u)^T\Lambda^u\}^{-1}(\mathcal{S}^{\mathcal{U}_m}_k)^T \times (\Lambda^y)^T\Lambda^y\,[R_{k+1|k} - \mathcal{I}(h(x_{k|k},C^w x^w_{k|k}) - H_k x_{k|k} - H^d_k C^w x^w_{k|k}) - \mathcal{S}^x_k(x_{k|k},u_{k-1},x^w_{k|k}) - \mathcal{S}^{x^w}_k x^w_{k|k}] \quad (59)$$

In the presence of constraints, the optimization can be solved via QP. Readers are referred to Ricker (1985) for a detailed solution procedure. The computed control moves are implemented in receding horizon fashion; that is, only the first move Δu_k is implemented and the whole optimization is repeated at the next sampling time.

2.5. Iterative Refinement of Prediction. One way to improve the quality of prediction developed from a linearized model is to iterate between the steps of the control input computation and sensitivity matrix computation. Recall that, in developing the affine expression for $x_{k+l|k}$ given in (41), we linearized the nonlinear expression $F_{t_s}(x_{k+l-1|k},u_{k+l-1},C^w(A^w)^{l-1}x^w_{k|k})$ with respect to $x_{k+l-1|k} = F_{(l-1)\cdot t_s}(x_{k|k},u_{k-1},C^w(A^w)^i x^w_{k|k}|_{0\le i\le l-2})$ (which is $x_{k+l-1|k}$ with the assumption of u remaining constant at u_{k-1}) and $u_{k+l-1} = u_{k-1}$. Clearly, this local linearization makes sense only when the computed inputs u_k, ..., u_{k+l-1} do not deviate from u_{k-1} by much. This assumption of constant u in the sensitivity matrix computation was necessary since these input values are not available for use at the prediction stage (*i.e.*, before performing the optimization at time k). Naturally, the trajectory of (u_k, ..., u_{k+l-1}) computed in the subsequent optimization is likely to be a better approximation of the actual future input sequence than the constant input profile of u_{k-1}.

This consideration suggests to us a way to improve the prediction from the linearized model by iterating between the steps of linearization/prediction and control move optimization. The error introduced from the above-mentioned assumption can be reduced by relinearizing the nonlinear expression at $\bar{u}_{k+l-1}$ and $\bar{x}_{k+l-1|k}$, which denote the input trajectory computed from the previous iteration and the corresponding state trajectory. Then, the *deviation* from the previous input trajectory (*i.e.*, the trajectory computed during the previous iteration) can be optimized through QP and used to update the trajectory. This idea was explored in the multistep Newton-type control algorithm studied by Li and Biegler (1989). Readers are referred to their paper for details and some limited stability results.

This iterative refinement should lead to performance improvement for many cases. However, the computational demand increases and there is no guarantee that the sequence of computed input trajectories will converge to a single trajectory as the number of iteration increases without optimizing the step size for the correction made at each iteration through a laborious line search (as suggested by Li and Biegler). In addition, because of the receding horizon implementation, the first computed move u_k is of highest importance and improvement gained from such iteration is often small. This fact was also pointed out by Li and Biegler.

3. Output Disturbance Models and Connections with NLQDMC

In this section, we make connection between the extended Kalman filter based MPC technique and NLQDMC techniques proposed by Garcia (1984) and Gattu and Zafiriou (1992).

3.1. NLQDMC for Stable Processes and Connection with EKF-Based MPC. 3.1.1. Output Disturbance Model. In the previous section, we rigorously developed an optimal multistep prediction equation using the concept of the extended Kalman filter. The extended Kalman filter and multistep prediction were based on a nonlinear ODE model that includes load disturbance effects on the *states* (recall $\dot{x} = f(x,u,\mathbf{d})$). In the earlier versions of MPC (for example, DMC, QDMC, or IMC), it was customary to express the overall effect of load disturbances on the *output* directly. Such a way of describing the disturbance effect will be referred to as the "output disturbance model" and can be written as

$$\dot{x} = f(x,u) \tag{60}$$

$$y = g_1(x) + d \tag{61}$$

Often elements of d are assumed to be uncorrelated (*i.e.*, the covariance of d is diagonal) since disturbance correlation among different outputs are very difficult and time-consuming to develop. (60) and (61) are certainly a special case of our general model (1) and (2) in which $g(x,d) = g_1(x) + d$ and hence all the techniques developed in the previous section apply.

3.1.2. Calculating the Steady-State Filter Gain. Let us assume for this section that the system is open-loop *stable* (*i.e.*, all eigenvalues of A_k lie strictly inside the unit disk $\forall\ k$). Model the output disturbance d as output of a stochastic process $C^w(qI - A^w)^{-1}B^w$ driven by white noise w_k. The augmented, discrete model derived from (60) and (61) is

$$\begin{bmatrix} x_k \\ x_k^w \end{bmatrix} = \begin{bmatrix} F_{t_s}(x_{k-1},u_{k-1}) \\ A^w x_{k-1}^w \end{bmatrix} + \begin{bmatrix} 0 \\ B^w \end{bmatrix} w_{k-1} \tag{62}$$

$$\hat{y}_k = g_1(x_k) + C^w x_k^w + \nu_k \tag{63}$$

The EKF gain L_k for (62) and (63) reaches a steady state as $k \rightarrow \infty$ since the disturbance dynamics is completely decoupled from the nonlinear process dynamics. Its steady-state value $\bar{L}$ ($\triangleq \lim_{k\rightarrow\infty} L_k$) can be calculated by solving the following algebraic Riccati equation of reduced dimension:

$$\bar{L} = \begin{bmatrix} 0 \\ \bar{\Sigma}^w (C^w)^T (C^w \bar{\Sigma}^w (C^w)^T + R^\nu)^{-1} \end{bmatrix} \tag{64}$$

$$\bar{\Sigma}^w = A^w \bar{\Sigma}^w (A^w)^\mathrm{T} + R^w - A^w \bar{\Sigma}^w (C^w)^\mathrm{T} \{R^\nu + C^w \bar{\Sigma}^w (C^w)^\mathrm{T}\}^{-1} C^w \bar{\Sigma}^w (A^w)^\mathrm{T} \tag{65}$$

Hence, x_k receives no feedback correction at steady state.

3.1.3. Parametrization of Steady-State Filter Gain. While an arbitrary choice for (A^w, B^w, C^w) can be used to shape the disturbance spectrum, in the absence of any fundamental model or identification experiment results, the following choice strikes a good balance between generality and convenience of design/tuning:

$$A^w = \begin{bmatrix} I & \mathrm{diag}\{\alpha_1,...,\alpha_{n_y}\} \\ 0 & \mathrm{diag}\{\alpha_1,...,\alpha_{n_y}\} \end{bmatrix};\quad B^w = \begin{bmatrix} I \\ I \end{bmatrix};\quad C^w = [I \quad 0] \tag{66}$$

$$R^w = \mathrm{diag}\{r_1^w,...,r_{n_y}^w\};\ R^\nu = \mathrm{diag}\{r_1^\nu,...,r_{n_y}^\nu\} \tag{67}$$

This makes d output of the following stochastic process driven by white noise w_k:

$$d_k = \mathrm{diag}\left\{\frac{q}{(q-1)(q-\alpha_1)},...,\frac{q}{(q-1)(q-\alpha_{n_y})}\right\} w_k \tag{68}$$

This integrated white noise filtered through the first-order dynamics (whose time constant is determined by α_i) should adequately describe the "over-damped, persistent" characteristics of most disturbances occurring in stable processes. Note that the limiting case of $\alpha_i = 0$ corresponds to the integrated-white noise (*i.e.*, random walk) disturbance and the other limiting case of $\alpha_i = 1$ corresponds to the double-integrated-white noise (*i.e.*, random ramp) disturbance. It can be shown that the steady-state optimal filter gain $\bar{L}$ can be parametrized as follows:

$$\lim_{k\rightarrow\infty} L_k = \bar{L} \triangleq \begin{bmatrix} 0 \\ \mathrm{diag}\{f_1^a,...,f_{n_y}^a\} \\ \mathrm{diag}\{f_1^b,...,f_{n_y}^b\} \end{bmatrix};\quad f_i^b = \frac{(f_i^a)^2}{1 + \alpha_i - \alpha_i f_i^a} \tag{69}$$

where f_i^a is a parameter determined by the signal-to-noise ratio (r_i^w/r_i^ν) and takes a value between 0 ($r_i^w/r_i^\nu = 0$) and 1 ($r_i^w/r_i^\nu \rightarrow \infty$). Both f_i^a and α_i can be adjusted on-line for desired closed-loop response and robustness. $(f_a)_i$ mainly determines the speed of the closed-loop response and α_i the balance between the overshoot and the settling time (Lee *et al.*, 1994; Lee and Yu, 1994). For open-loop stable, continuous processes, the above steady-state filter gain can be used for all practical purposes. For batch processes, it may be advantageous to employ the time-varying Kalman filter gain to correct for the initialization errors in x more efficiently.

3.1.4. Equivalence with NLQDMC. When every (α_i) in the disturbance model (66) and (67) is set to zero (*i.e.*, every element of d is integrated white noise) and measurement noise is assumed negligible (*i.e.*, $R^\nu = 0$), the multistep prediction equation (53) developed using the extended Kalman filter with the steady-state optimal gain $\bar{L}$ is identical to the prediction equation used in NLQDMC by Garcia (1984). In NLQDMC, a constant bias of $\hat{y}_k - g(x_{k|0})$ (where $x_{k|0}$ is an open-loop estimate of x_k) is added to the open-loop output prediction terms $y_{k+l|0}$ to form the prediction for y_{k+l}. This is entirely equivalent to assuming that the unmeasured disturbances are independent, integrated white noise directly entering each output and measurement noise is negligible.

3.1.5. Limitations and Potential Hazards for Using the Output Disturbance Model. While the output disturbance model has proven useful in many process control applications, there are some potential hazards and limitations. First, it is not applicable to unstable processes. Note that the steady state gain expression (64) is valid only for open-loop stable processes. For unstable processes, the time-varying EKF gain must be calculated from RDE of (29) and (30) which is initialized with a *positive definite* covariance matrix $\Sigma_{0|0}$. Second, it is not applicable to inferential control problems (where y differs from y^c) since measurements of secondary outputs cannot be used to improve the prediction of unmeasured controlled variables in the absence of a disturbance model correlating these outputs. Finally, although the estimates of the measured *outputs* are unbiased at steady state (because of the integrated white noise disturbance added to the output), the estimates of the states are clearly biased. This may adversely effect the quality of the model obtained via local linearization.

3.2. NLQDMC for Unstable Processes and Connection with EKF-Based MPC. 3.2.1. Output Disturbance Model and Connection with NLQDMC. The output disturbance model (60) and (61) can be used also for unstable processes if the EKF gain is recomputed at each time step based on RDE (29) and (30) initialized with a positive definite covariance matrix. In this case, the states do receive feedback continuously. This enables direct extension of NLQDMC (based on the output disturbance model) to unstable processes.

Gattu and Zafiriou (1992) proposed a slightly different approach to handling unstable processes within the framework of NLQDMC. In their approach, independent white noise is added to each state in addition to the integrated white noise disturbance added to each output:

$$\begin{bmatrix} x_k \\ x_k^w \end{bmatrix} = \begin{bmatrix} F_{t_s}(x_{k-1},u_{k-1}) \\ x_{k-1}^w \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} (w_1)_{k-1} \\ (w_2)_{k-1} \end{bmatrix} \quad (70)$$

$$\hat{y}_k = g_1(x_k) + x_k^w + v_k \quad (71)$$

where w_1 and w_2 are white noise of scalar-times-identity covariance matrices R_1^w and R_2^w, respectively.

Although the technique of Gattu and Zafiriou is best interpreted within our proposed framework as above, it differs from the EKF-based technique in several aspects. First, instead of using the Kalman filter gain calculated explicitly via RDE (29) and (30), they suggest using

$$L_k = \begin{bmatrix} (\bar{L}_1)_k \\ I \end{bmatrix} \quad (72)$$

where $(\bar{L}_1)_k$ is the *steady-state* Kalman filter gain matrix calculated using the formula

$$(\bar{L}_1)_k = \bar{\Sigma}^x C_k^{\mathrm{T}}(C_k\bar{\Sigma}^x C_k^{\mathrm{T}} + R^v)^{-1} \quad (73)$$

$$\bar{\Sigma}^x = A_{k-1}\bar{\Sigma}^x A_{k-1}^{\mathrm{T}} + R_1^w - A_{k-1}\bar{\Sigma}^x C_k^{\mathrm{T}}(R^v + C_k\bar{\Sigma}^x C_k^{\mathrm{T}})^{-1} C_k\bar{\Sigma}^x A_{k-1}^{\mathrm{T}} \quad (74)$$

Not only is the use of the *steady-state* Kalman filter gain for a time-varying linear model improper, but the gain matrix (72) is not even the *optimal* steady-state filter gain. This can lead to biased state estimates and prediction. Second, in the prediction phase, instead of adding the measurement correction term $(\bar{L}_1)k(y_k - y_{k|k-1})$ to the states and integrating the ODE with the corrected states as initial condition, they add the correction term $\Phi_k(\bar{L}_1)_k(y_k - y_{k|k-1})$ uniformly to the predicted states in the prediction horizon after the nonlinear integration. They provide no theoretical justification for this. Finally, our linearization points are more consistent.

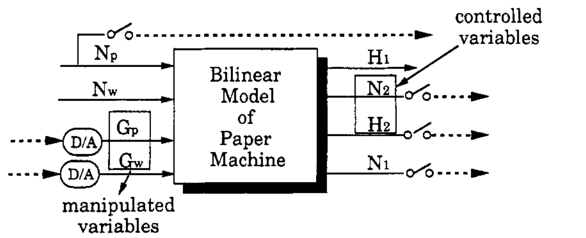

Figure 1. Schematic diagram representing the bilinear model and control problem of a paper machine.

3.2.2. Effect of Model Errors. Gattu and Zafiriou claim that the white noise disturbances to the states x_k, in addition to assuring that the Kalman filter gain is stabilizing, account for the state errors caused by model parameter uncertainty. However, errors caused by model parameter uncertainty are correlated with previous state and input values and therefore nonwhite. It is generally inappropriate to account for errors for such sources with white state excitation noise. Our experience has been that, in order to reduce the bias in the state estimates caused by the model parameter errors, we need to add disturbances of integrating characteristics to the states. The number of integrated white noise disturbances added should coincide with the number of measurements and enter into the state vector such that the system remains detectable. This will give sufficient degrees of freedom to the Kalman filter so that the estimates for the *outputs* are unbiased. Because the outputs are unbiased, the particular combinations of the states corresponding to the outputs will be unbiased. However, the remaining part of the state space (*i.e.*, the null space of the measurement equation) will be biased in general. This fact will be further elaborated in the following example.

4. Application to Paper Machine Headbox Control

4.1. Model. Ying *et al.* (1992) studied the control of composition and liquid level in a paper machine headbox. The system and its control problem are illustrated in Figure 1. The system is modeled as a bilinear, continuous-time system as follows:

$$\dot{x} = Ax + B_0u + \sum_{i=1}^{n_u} u_iB_ix + B_\upsilon u + B_dd \quad (75)$$

$$y = Cx \quad (76)$$

The state vector consists of the liquid level in the feed tank (H_1), the liquid level in the headbox (H_2), the consistency (percentage of pulp fibers in suspension) in the feed tank (N_1), and the consistency in the headbox (N_2). All states except H_1 are measured. The primary control objective is to regulate N_2 and H_2. Manipulated variables are the flow rate of stock entering the feed tank (G_p) and the flow rate of the recycled white water (G_w). Disturbances are the consistency of the stcok entering the feed tank (N_p), which is measured, and the consistency of the white water (N_w), which is unmeasured. Hence, the state vector x, manipulated input vector u, measured disturbance υ, unmeasured disturbance d and output vector y are defined as

$$x = [H_1 \;\; H_2 \;\; N_1 \;\; N_2]^{\mathrm{T}}$$

$$u = [G_p \;\; G_w]^{\mathrm{T}}$$
$$v = N_p$$
$$d = N_w$$
$$y = [N_2 \;\; H_2 \;\; N_1]^{\mathrm{T}} \tag{77}$$

Ying *et al.* derived the following constant parameter matrices:

$$A = \begin{bmatrix} -1.93 & 0 & 0 & 0 \\ 0.394 & -0.426 & 0 & 0 \\ 0 & 0 & -0.63 & 0 \\ 0.82 & -0.784 & 0.413 & -0.426 \end{bmatrix}$$

$$B_0 = \begin{bmatrix} 1.274 & 1.274 \\ 0 & 0 \\ 1.34 & -0.65 \\ 0 & 0 \end{bmatrix}; \quad B_1 = B_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -0.327 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}; \quad B_v = \begin{bmatrix} 0 \\ 0 \\ 0.203 \\ 0 \end{bmatrix}; \quad B_d = \begin{bmatrix} 0 \\ 0 \\ 0.406 \\ 0 \end{bmatrix} \tag{78}$$

All variables are zero at the nominal steady state. The process is open-loop stable.

We design a control system based on the known disturbance characteristics, assuming that N_w can be modeled as stepwise variations of random magnitude and duration (integrated white noise). Then, in (5) and (6), $A_w = 1, B_w = 1,$ and $C_w = 1$. The Jacobian matrices required in (13), (14), and (34) can be computed analytically:

$$\tilde{A}_{k-1} = A + ((u_1)_{k-1} + (u_2)_{k-1})B_1 \tag{79}$$

$$\tilde{B}^d_{k-1} = \tilde{\mathcal{B}}^d_k = B_d \tag{80}$$

$$\tilde{\mathcal{B}}^u_k = B_0 + [B_1 x_{k|k} \;\; B_2 x_{k|k}] \tag{81}$$

The output is a linear, time-invariant function of the states, so $C_k = H_k = C$ and $C^d_k = H^d_k = 0$. The desired closed-loop response time is in the order of 1 min, so we choose a sampling period of $T_s = 0.25$ min.

We note that, because the only nonlinearity of the model comes from the bilinear terms multiplying the state vector with the inputs, the system can be considered as a linear time-varying system and linearization at each time step provides a perfect model for the particular time step. Hence, we expect the EKF to provide good state estimates assuming the model parameters are accurate and covariance matrices are chosen properly.

4.2. Controller Design. The design parameters are chosen as follows:

$$\Sigma_{0|0} = I, \;\; R^w = \gamma I, \;\; R^v = I$$
$$\Lambda^y = \mathrm{diag}\{1,1,0\}, \;\; \Lambda^u = \lambda I, \;\; p = 5, \;\; m = 3 \tag{82}$$

γ and λ are used as tuning parameters to find a good balance between the closed-loop speed and robustness. The measured disturbance N_p and the setpoints for N_2 and H_2 are projected to remain constant at their current values throughout the prediction horizon.

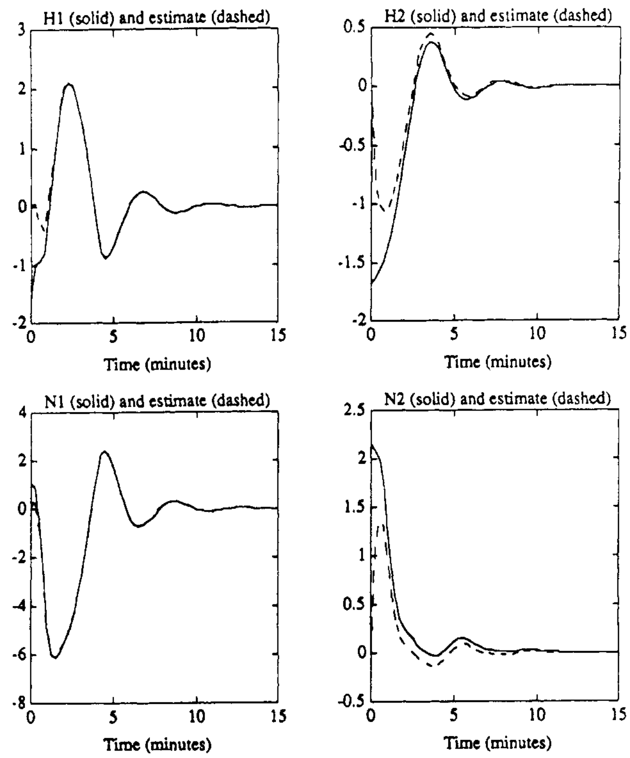

Figure 2. Effect of bias in the initial state estimates, $x_{0|0}$, on the performance of EKF-based NLMPC. Setpoints for H_2 and N_2 are both zero, and no disturbances or model errors are present.

4.3. Simulation Results for Ideal Case. We first test the proposed technique for the idea case where the model matches the plant exactly.

4.3.1. Elimination of Initial State Error. Figure 2 shows the regulatory response when there is a bias in $x_{0|0}$. The tuning parameters used were $\gamma = 3$ and $\lambda = 0.2$. The initial state of the plant x_0 was $[-1.5794\ -1.6811\ 1.0311\ 2.1436]^{\mathrm{T}}$ (chosen randomly), and that of the estimator $x_{0|0}$ was $[0\,0\,0\,0]$. The disturbances were held constant at $N_p = N_w = 0$. Within 15 min, the controller drives H_2 and N_2 to their setpoints of zero, and the state estimates converge to the true values. The closed-loop response time is about 3 min, and the outputs are well-damped.

4.3.2. Servo Response and Comparison with Linear MPC. Figure 3 shows the servo response for a step change of -1 in the setpoint of H_2 from an initial steady state at $x_0 = 0$. The setpoint of N_2 is zero at all times, and there are no disturbances. The initial state estimates are unbiased, *i.e.*, $x_{0|0} = x_0$. Since the "plant" and "model" are the same set of ODEs, EKF provides perfect state estimates for all k. The response time is about 2 min with an overshoot of about 10%. The output decoupling is good—the setpoint change of 1.0 unit in H_2 causes a maximum deviation of about 0.1 unit in N_2. Oscillations are well-damped.

The shortcoming of linear MPC for the problem is illustrated in Figure 4. This MPC was designed using a linear model derived at the nominal steady state. The state estimator is the usual DMC type, which assumes random-step disturbances at the outputs. The values of p, m, and λ used in the EKF-based NLMPC gave an unstable closed-loop response. The simulations shown in Figure 4 used $p = 20$, blocking of $m = [3\ 5\ 12]$ (see Ricker, (1985, 1990) for detail), and $\lambda = 0.4$. The servo response is significantly worse than that of NLMPC. The setpoint change in H_2 causes large, poorly-damped oscillations in N_2. The usual tuning tricks cannot overcome this problem. For example, use of $m = 3$, $\lambda = 2$ reduces the magnitude of the oscillations in N_2, but the response is more sluggish

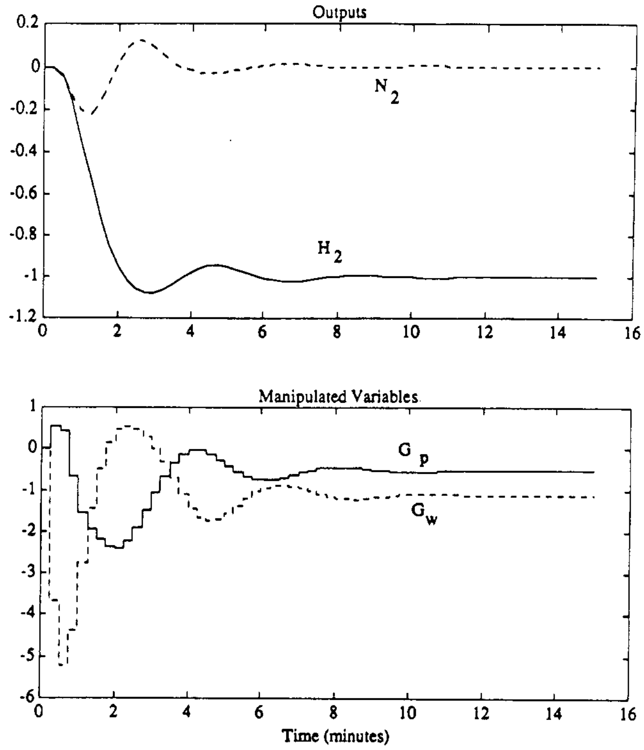

Figure 3. Servo response of EKF-based NLMPC for a step change in the setpoint for H_2 from 0 to −1. At all times, setpoint for N_2 is zero and no disturbances or model errors are present.

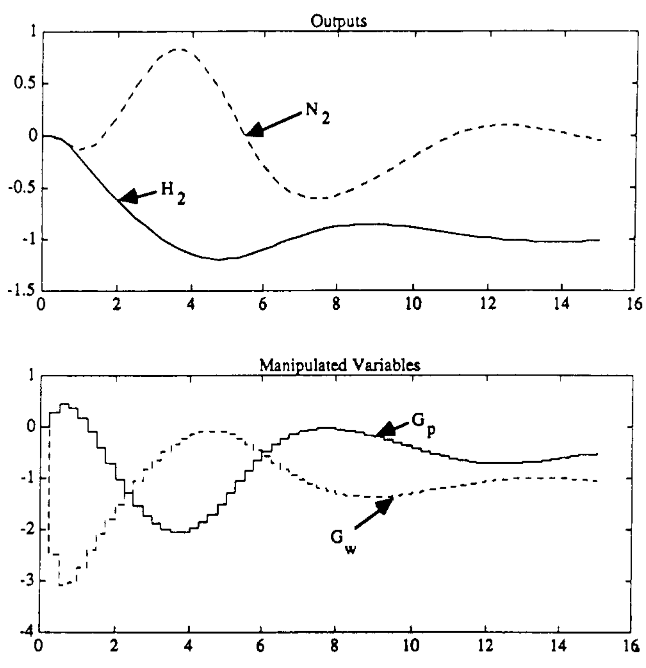

Figure 4. Servo response of Linear MPC for a step change in the setpoint for H_2 from 0 to −1. At all times, setpoint for N_2 is zero and no disturbances or model errors are present.

and is still poorly damped (not shown). For setpoint changes of small magnitude, on the other hand, the MPC servo response approaches that of NLMPC.

4.3.3. Regulatory Response and Comparison with NLQDMC. Figure 5 shows the response to a step change of 10 in the unmeasured disturbance N_w. The initial state estimates are unbiased, *i.e.*, $x_{0|0} = x_0$ and setpoints for both N_2 and H_2 are maintained at 0. Again, because the EKF provides accurate state estimates, EKF-based NLMPC shows excellent disturbance rejection property. The response time is again about 2 min; there is some, but well-damped, oscillation.

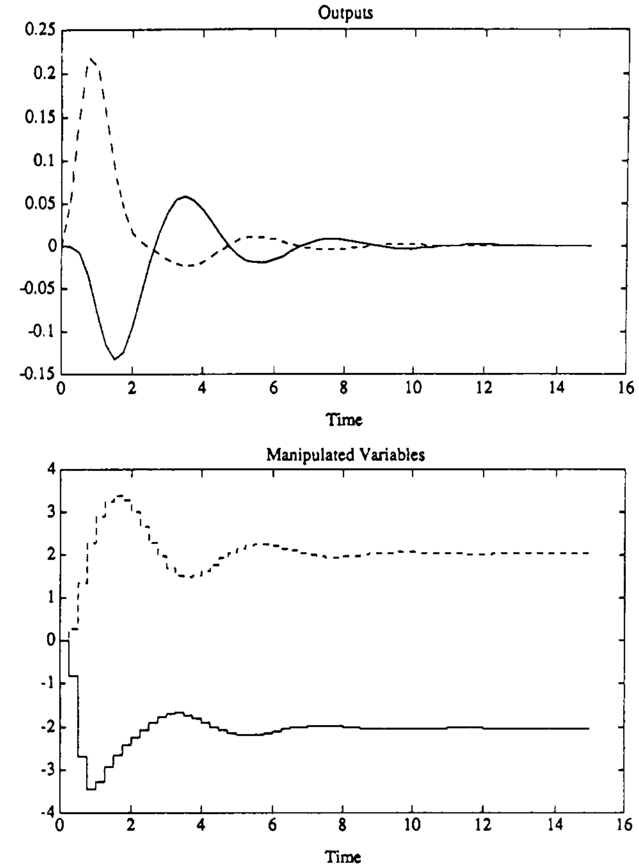

Figure 5. Regulatory response of EKF-based NLMPC for a step change in the unmeasured disturbance N_w from 0 to 10. At all times, setpoints for both H_2 and N_2 are zero. No model errors are present.

Figure 6 shows the response of NLQDMC to the same step change in N_w. The resulting closed-loop responses are markedly worse than those obtained with EKF-based NLMPC. The overshoot is large and the oscillation is poorly damped. This clearly shows the aforementioned problem of using the output disturbance model for nonlinear control. Because the state estimates are never corrected with measurements, they contain large bias. This in turn causes the linearized model (obtained by linearizing the nonlinear model at the biased estimates) to be a poor representation of the local dynamics.

4.4. Simulation Results for Model Error Case. We next test the robustness of the design by introducing parametric plant/model mismatch. This is not a rigorous test of robustness, but gives one an idea of the sensitivity of each method to modeling errors. Each element of the matrices used to represent the plant was multiplied by 1 + ϵ, where ϵ was from a normal distribution, $N(0,0.1)$. Thus, the perturbations are 10% of the nominal value on the average. The C matrix was not modified for an obvious reason, and the controller and estimator calculations used the same model parameters as before.

4.4.1. Servo Response for the Design with One Degree of Freedom. The responses of EKF-based NLMPC degrades considerably (Figure 7), exhibiting an offset of 0.3 unit in H_2 and 0.6 unit in N_2. The problem here is that N_w affects states 3 and 4 only. As the EKF converges to a steady-state condition, the first two rows of the gain matrix, L_k, go to zero. In other words, the original disturbance model stipulates that there should be no disturbances in the first two states, so they receive no feedback correction at steady state. They are biased by model error, however. This causes the estimates of the remaining states to be biased too, as the EKF tries to eliminate the apparent error observed at these outputs.

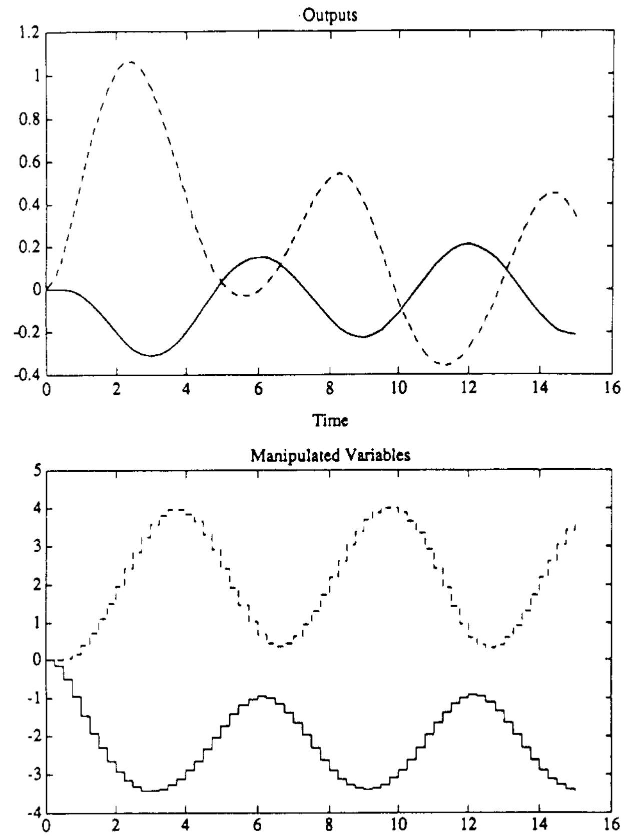

Figure 6. Regulatory response of NLQDMC for a step change in the unmeasured disturbance N_w from 0 to 10. At all times, setpoints for both H_2 and N_2 are zero. No model errors are present.

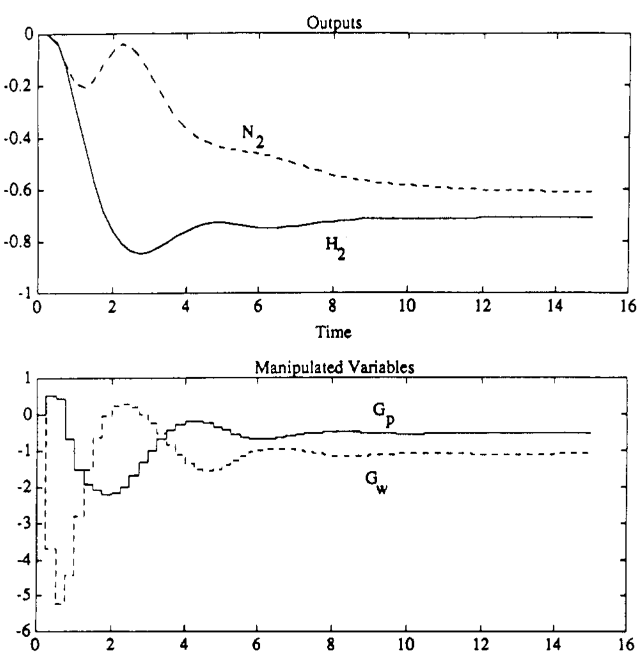

Figure 7. Servo response of EKF-based NLMPC for a step change in the setpoint for H_2 from 0 to -1. The EKF is designed with *one* disturbance degree of freedom. Random errors of 10% are present in the state space parameters. At all times, setpoint for N_2 is zero and no disturbances are present.

In doing so, it only has one degree of freedom: the disturbance N_w. Thus, it is impossible for the EKF (formulated as discussed) to construct the states to track their measured values without bias.

4.4.2. Servo Response for the Design with Three Degrees of Freedom. Understanding the cause for the observed phenomena also suggests a way to improve the design so that the problem may be alleviated. Since we are measuring three of the four states, we should clearly be able to construct unbiased estimates for these states. We can accomplish this by adding two more integrated white noise disturbances into the measured states and making the number of state disturbances (with integrating characteristics) equal to the number of measurements. We put our new disturbances into the second and fourth states. Putting a disturbance into the third state would make it a linearly dependent disturbance (redundant with N_w). Putting another integrated white noise into the first state caused indetectability since the number of disturbances then exceeds the number of measurements. The covariance matrix R^w was chosen as a diagonal matrix and tuned for the best results. All other controller parameters were as in the simulations for the previous cases. Results (with model error) appear in Figure 8. There is less interaction and the process settles more quickly to the setpoint than for the previous simulation. Furthermore, at the final time we have x_k = [-1.3597 -1.0000 1.4679 0.0000]^T and $x_{k|k}$ = [-1.5380 -1.0000 1.4679 0.0000]^T; *i.e.*, bias has been eliminated from the three measured states. This is possible because the EKF now has three degrees of freedom (the three state disturbances). The unmeasured state is slightly biased, but this has no impact on the controller performance.

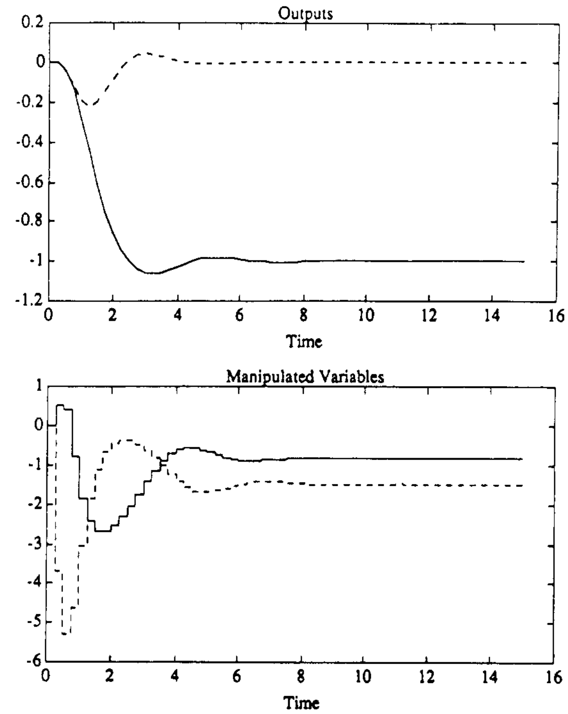

Figure 8. Servo response of EKF-based NLMPC for a step change in the setpoint for H_2 from 0 to -1. The EKF is designed with *three* disturbance degrees of freedom. Random errors of 10% are present in the state space parameters. At all times, setpoint for N_2 is zero and no disturbances or model errors are present.

For the low-order paper machine process, the "artificial" state disturbances included to eliminate offsets caused by model errors are easily tuned, but this might not be the case for poorly scaled, high-order processes, or those for which the outputs are a nonlinear function of the states. It is recommended in general to design these disturbances such that they enter the measured combinations of the states. This ensures that the estimates for the measured states are unbiased. The effect of model errors on the unmeasured states cannot be corrected since these errors cannot be detected from the measurements.

5. Conclusions

In this paper, we proposed a nonlinear MPC technique that is applicable to models of general stochastic state disturbances and noise. The prediction equation was developed based on the concept of successive linearization and was shown to be optimal under certain affine approximations of the state and measurement equations obtained via linearization at the current state estimate and previous input values. The technique requires integration of the nonlinear ODE model over the chosen prediction horizon and solving QP for control move computation. Although similar in spirit with the recent work by Gattu and Zafiriou (1992), our proposed technique has advantages of broader applicability (for example, to inferential control problems) and theoretical rigor. Example application to the control problem of a bilinear paper machine was presented to point out the advantages and potential hazards of the proposed technique. A more elaborate application of the technique to an inferential control problem arising in batch pulp digesters can be found in Lee and Datta (1994).

While being effective for many practical applications, one inherent shortcoming of linearization-based approaches such as this is that the quality of the linearized model is dependent on the accuracy of the state estimates. Motivated by this, a more rigorous nonlinear state estimation technique utilizing NLP is currently being developed (Robertson and Lee, 1993).

Acknowledgment

J.H.L. gratefully acknowledges financial support from the National Science Foundation through Grant CTS-9209808 and from Pulp and Paper Research and Education Center at Auburn University. N.L.R. acknowledges financial support from the National Science Foundation through Grant CTS-9214983.

Literature Cited

Bequette, B. W. Nonlinear Control of Chemical Processes: A Review. *Ind. Eng. Chem.* **1991**, *30*, 1391–1413.

Biegler, L. T.; Rawlings, J. B. Optimization Approaches to Nonlinear Model Predictive Control. *Proc. CPC-IV*, San Padre Island, TX; 1991; pp 543–571.

Cox, H. On the Estimation of State Variables and Parameters for Noisy Dynamic Systems. *IEEE Trans. Autom. Control* **1964**, *AC-9*, 5–12.

Cutler, C. R.; Ramaker, B. L. Dynamic Matrix Control—A Computer Control Algorithm. *Proc. Automatic Control Conf.* San Francisco, CA; 1980; paper WP5-B.

Cutler, C. R.; Hawkins, R. B. Application of a Large Predictive Multivariable Controller to a Hydrocracker Second Stage Reactor. *Proc. Automatic Control Conf.* Atlanta, GA; 1988; pp 284–291.

Garcia, C. E. Quadratic/Dynamic Matrix Control of Nonlinear Processes: An Application to a Batch Reaction Processes. Presented at the AIChE Annual Meeting, San Francisco, CA, 1984.

Garcia, C. E.; Morari, M. Internal Model Control: 1. A Unifying Review and Some New Results. *Ind. Eng. Chem. Process Des. Dev.* **1982**, *21*, 308–323.

Garcia, C. E.; Morshedi, A. M. Quadratic Programming Solution of Dynamic Matrix Control (QDMC). *Chem. Eng. Commun.* **1986**, *46*, 73–87.

Garcia, C. E.; Prett, D. M.; Morari, M. Model Predictive Control: Theory and Practice—a Survey. *Automatica* **1989**, *25*, 335–348.

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

Goodwin, G. C.; Sin, K. S. *Adaptive Filtering, Prediction and Control*; Prentice-Hall: Englewood Cliffs, NJ, 1984.

Jang, S. S.; Joseph, B.; Mukai, H. Control of Constrained Multivariable Nonlinear Process Using a Two-Phase Approach. *Ind. Eng. Chem. Res.* **1987**, *26*, 2106–2114.

Jazwinski, A. H. *Stochastic Processes and Filtering Theory*; Academic Press, Inc.: San Diego, CA, 1970.

Kopp, R. E.; Oxford, R. J. Linear Regression Applied to System Identification and Adaptive Control System. *AIAA J.* **1963**, *1* (10), 2300–2306.

Li, W. C.; Biegler, L. T. Multistep Newton-Type Control Strategies for Constrained Nonlinear Processes. *Chem. Eng. Res. Des.* **1989**, *67*, 562–577.

Lee, J. H.; Datta, A. Nonlinear Inferential Control of Pulp Digesters. *AIChE J.* **1994**, *40* (1), 50–64.

Lee, J. H.; Yu, Z. Tuning of Model Predictive Controllers for Robust Performance. *Comput. Chem. Eng.* **1994**, *18* (1), 15–37.

Lee, J. H.; Gelormino, M. S.; Morari, M. Model Predictive Control of Multi-Rate Sampled-Data Systems: A State-Space Approach. *Int. J. Control* **1992**, *55*, 153–191.

Lee, J. H.; Morari, M.; Garcia, C. E. State-Space Interpretation of Model Predictive Control. *Automatica* 1994, in press.

Morshedi, A.; Cutler, C. R.; Skrovanek, T. A. Optimal Solution of Dynamic Matrix Control with Linear Programming Techniques (LDMC). *Proc. American Control Conf.* Boston, MA; 1985; pp 199–208.

Richalet, J.; Rault, A.; Testud, J. L.; Papon, J. Model Predictive Heuristic Control: Application to Industrial Processes. *Automatica* **1978**, *14*, 413.

Ricker, N. L. The Use of Quadratic Programming for Constrained Internal Model Control. *Ind. Eng. Chem. Process Des. Dev.* **1985**, *24*, 925–936.

Ricker, N. L. Model Predictive Control with State Estimation. *Ind. Eng. Chem. Res.* **1990**, *29*, 374–382.

Robertson, D.; Lee, J. H. Least Squares Formulation of State Estimation. *Proc. American Control Conf.*, San Francisco, CA; 1993; pp 389–394.

Rouhani, R.; Mehra, A. K. Model Algorithmic control (MAC): Basic Theoretical Properties. *Automatica* **1982**, *18*, 401.

Sistu, P. B.; Bequette, B. W. Nonlinear Predictive Control of Uncertain Processes: Application to a CSTR. *AIChE J.* **1991**, *37* (11), 1711–1723.

Ying, Y.; Rao, M.; Sun, Y. Bilinear Control Strategy for Paper-Making Process. *Chem. Eng. Commun.* **1992**, *111*, 13–28.

Received for review September 10, 1993
Revised manuscript received February 24, 1994
Accepted March 8, 1994[⊗]

[⊗] Abstract published in *Advance ACS Abstracts*, May 1, 1994.