Intelligent state estimation for fault tolerant nonlinear predictive control

Anjali P. Deshpande [a], Sachin C. Patwardhan [b], Shankar S. Narasimhan [c]

[a] Systems and Control Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai 400076, India
[b] Department of Chemical Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai 400076, India
[c] Department of Chemical Engineering, Indian Institute of Technology, Madras, Chennai 600036, India

Keywords:
Extended Kalman filter
Generalized likelihood ratio method
Fault diagnosis and accommodation
Failure tolerance
Nonlinear model predictive control

A B S T R A C T

There is growing realization that on-line model maintenance is the key to realizing long term benefits of a predictive control scheme. In this work, a novel intelligent nonlinear state estimation strategy is proposed, which keeps diagnosing the root cause(s) of the plant model mismatch by isolating the *subset of active faults* (abrupt changes in parameters/disturbances, biases in sensors/actuators, actuator/sensor failures) and auto-corrects the model on-line so as to accommodate the isolated faults/failures. To carry out the task of fault diagnosis in multivariate nonlinear time varying systems, we propose a nonlinear version of the generalized likelihood ratio (GLR) based fault diagnosis and identification (FDI) scheme (NL-GLR). An active fault tolerant NMPC (FTNMPC) scheme is developed that makes use of the fault/failure location and magnitude estimates generated by NL-GLR to correct the state estimator and prediction model used in NMPC formulation. This facilitates application of the fault tolerant scheme to nonlinear and time varying processes including batch and semi-batch processes. The advantages of the proposed intelligent state estimation and FTNMPC schemes are demonstrated by conducting simulation studies on a benchmark CSTR system, which exhibits input multiplicity and change in the sign of steady state gain, and a fed batch bioreactor, which exhibits strongly nonlinear dynamics. By simulating a regulatory control problem associated with an unstable nonlinear system given by Chen and Allgower [H. Chen, F. Allgower, A quasi infinite horizon nonlinear model predictive control scheme with guaranteed stability, Automatica 34(10) (1998) 1205–1217], we also demonstrate that the proposed intelligent state estimation strategy can be used to maintain asymptotic closed loop stability in the face of abrupt changes in model parameters. Analysis of the simulation results reveals that the proposed approach provides a comprehensive method for treating both faults (biases/drifts in sensors/actuators/model parameters) and failures (sensor/ actuator failures) under the unified framework of fault tolerant nonlinear predictive control.

1. Introduction

The need to operate continuous processes over wide operating ranges and semi-batch/batch processes efficiently has motivated the development of nonlinear MPC (NMPC) techniques over last two decades. These techniques employ nonlinear models for prediction. The prediction model is typically developed once in the beginning of implementation of an NMPC scheme. However, as time progresses, slow drifts in unmeasured disturbances and changes in process parameters can lead to significant mismatch in plant and model behavior. Also, NMPC schemes are typically developed under the assumption that sensors and actuators are free from faults. However, *soft faults*, such as biases in sensors or actuators, are frequently encountered in the process industry. In addition to this, some sensor(s) and/or actuator(s) may fail during operation, which results in loss of degrees of freedom for control. Occurrences of such parametric changes, soft faults and failures progressively result in severe model-plant mismatch. This can lead to a significant degradation in the closed loop performance of the NMPC scheme and may also lead to instability. Thus, to arrest the degradation in controller performance, it is extremely important to isolate the *root causes* of the plant model mismatch and, if possible, compensate for them on-line.

The conventional approach to deal with the model-plant mismatch in the NMPC formulations is through the introduction of additional artificial states in the state observer [2–4]. The main limitation of this approach is that the number of extra states introduced cannot exceed the number of measurements. This implies that it is necessary to have *a priori* knowledge of which subset of faults are most likely to occur or which parameters are most likely to drift. In such a formulation where the state vector is permanently augmented with subset of parameters to be estimated, the state estimates can become biased when unanticipated abrupt

changes/faults occur. Moreover, the permanent state augmentation approach cannot systematically deal with the difficulties arising out of sensor biases or actuator/sensor failures. The difficulties encountered while selecting such a subset in design of extended Kalman filter (EKF) for a complex large dimensional system (Tennessee Eastman problem) have been highlighted by Ricker and Lee [2].

Attempts to develop fault-tolerant MPC schemes have mainly focused on dealing with sensor or actuator failures [5–7]. Yu et al. [6] have proposed to develop a failure tolerant cascaded Kalman filter with online tuning parameters. This approach involves the design of *main* and *auxiliary* Kalman filter (KF) based on *reliable set of measurements* and complete set of measurements, respectively. The auxiliary KF is used to remove the bias from the estimates given by the main KF. The steady state gain of auxiliary KF is modified online based on the failed measurements. Though this approach achieves fault tolerance while maintaining the integrity in the estimate of the lost output, the fault detection and isolation aspect does not feature in the formulation. Recently, Prakash et al. [8] have proposed an active fault tolerant linear MPC (FTMPC) scheme, which can systematically deal with soft faults in a unified framework. The FTMPC scheme is developed by integrating generalized likelihood ratio (GLR) method, a model based fault detection and identification (FDI) scheme, with the state space formulation of MPC based on Kalman filter. The GLR method performs fault identification using innovation sequence generated by the Kalman filter over a moving window of data in the past and this facilitates very close integration of the FDI and MPC schemes. The main limitation of these approaches arises from the use of linear perturbation model for performing control and diagnosis tasks. The use of linear models not only restricts its applicability to a narrow operating range but also limits the diagnostic abilities of fault detection and identification (FDI) components to only linear additive type faults. As a consequence, many faults that have a nonlinear effect on the system dynamics, such as abrupt changes in model parameters or unmeasured disturbances, have to be approximated as linear additive faults. Moreover, the FTMPC scheme does not deal with failures of sensors or actuators.

Recently, Mhaskar et al. [9,10] have presented an approach that deals with control system or actuator failure in nonlinear processes subject to constraints. They have presented an approach for design of robust hybrid predictive candidate controllers, which guarantees stability from an explicitly characterized set of initial conditions, subject to uncertainty and constraints. Reconfiguration or controller switching is done to activate or deactivate the constituent control configuration in order to achieve fault tolerance. The Fault tolerant controller uses the knowledge of the stability regions of the back up control configurations to guide the state trajectory within the stability regions of the back up control configurations to enhance the fault tolerance capabilities. Their approach, however, requires nonlinear system under consideration to have input affine structure. In another article, Mhaskar et al. [11] have presented an integrated fault detection and fault-tolerant control structure, for SISO nonlinear systems with input constraints subject to control failures. A bounded Lyapunov based controller has been developed, which depends on construction of control Lyapunov function. Upon failure of the primary controller, the faulty configuration is shut down and a well functioning fall back configuration is switched on. It may be noted that various control structures are developed by exploiting specific structural features of a nonlinear system, as no standard method is available for construction of these control Lyapunov functions. Also, these approaches, as proposed, do not address difficulties arising from abrupt changes in model parameters, mean shift in unmeasured disturbances, sensor/actuator biases and failed sensors.

Examination of various fault tolerant MPC/NMPC formulations proposed in literature reveals that the design of state observer is the key to integration of fault tolerance with predictive control. If it is desired to achieve tolerance with respect to a broad spectrum of faults (abrupt changes in unmeasured disturbance, parameter drifts, sensor/actuator biases) and sensor/actuator failures in a typical situation where the number of degrees of freedom available for observer design (synonymous with the number of measurements available for observer construction) is limited (i.e. far less than the number faults and failures to be dealt), then it becomes imperative to introduce some degree of intelligence in the state estimation to overcome these limitations [12]. In the present work, an intelligent nonlinear state estimation strategy is proposed, which keeps diagnosing the root cause(s) of the plant model mismatch by isolating the *subset of active faults* and auto-corrects the model on-line so as to accommodate the isolated faults. To carry out the task of fault diagnosis in multivariate nonlinear time varying systems, we propose a nonlinear version of the generalized likelihood ratio (GLR) based FDI scheme, which is referred to as nonlinear GLR (NL-GLR) in the rest of the text. The NL-GLR scheme, along with the fault location, also generates an estimate of the fault magnitude, which is used to correct the prediction model used in the proposed fault tolerant NMPC (FTNMPC) formulation. As the proposed NL-GLR scheme is computationally demanding, it is further simplified for online implementation (SNL-GLR). This simplification is based on linearization of nonlinear process model around a nominal trajectory. The significant contributions of the work described in this paper are

- Development of an active fault tolerant control scheme for nonlinear processes by suitably integrating a nonlinear version of the GLR method for FDI with a nonlinear model based controller.
- Development of fault/failure isolation strategy when multiple faults and failures occur simultaneously.
- Development of a comprehensive method for treating both faults (biases/drifts in sensors/actuators/model parameters) and failures (sensor/actuator failures) in fault diagnosis and accommodation.

The above contributions allow application of the fault tolerant scheme to nonlinear and time varying processes including batch and semi-batch processes. The proposed fault tolerant scheme also overcomes the limitation on the number of extra states that can be added to the state space model in NMPC for offset removal and allows bias compensation for more variables than the number of measured outputs. The advantages of the proposed state estimation and control scheme are demonstrated by conducting simulation studies on a benchmark CSTR system, which exhibits input multiplicity and change in the sign of steady state gain, and a fed batch bioreactor, which exhibits strongly nonlinear dynamics. By simulating regulatory control problem associated with a unstable nonlinear system given by Chen and Allgower [1], we also demonstrate that the proposed intelligent state estimation strategy can be used to recover closed loop stability in the face of abrupt changes in model parameters.

The rest of this article is organized as follows. To begin with, we develop the nonlinear version of GLR method. A fault tolerant NMPC formulation is presented in the subsequent section. We then proceed to present the results of simulation case studies. The main conclusions reached based on the analysis of these results are presented in the last section.

2. Fault diagnosis

In this section we develop an FDI method based on a nonlinear version of GLR scheme for diagnosing faults in nonlinear dynamic systems. To begin with, the method is described as applied once

when a single fault is detected for the first time. Modifications necessary for on-line implementation of the FDI scheme when multiple faults occur sequentially are described later.

2.1. Model for normal behavior

Consider a continuous time nonlinear stochastic system described by the following set of equations:

$$\mathbf{x}(k+1) = \mathbf{x}(k) + \int_{kT}^{(k+1)T} \mathbf{F}[\mathbf{x}(\tau), \mathbf{u}(k), \mathbf{p}, \mathbf{d}(k)]\,\mathrm{d}\tau \tag{1}$$

$$\mathbf{d}(k) = \bar{\mathbf{d}} + \mathbf{w}(k) \tag{2}$$

$$\mathbf{y}(k) = \mathbf{H}[\mathbf{x}(k)] + \mathbf{v}(k) \tag{3}$$

where $\mathbf{x} \in R^n, \mathbf{y} \in R^r$ and $\mathbf{u} \in R^m$ represent the state variables, measured outputs and manipulated inputs, respectively, and T represents sampling interval. The variables $\mathbf{p} \in R^p$ and $\mathbf{d} \in R^d$ represent the vector of parameters and unmeasured disturbance variables, respectively, which are likely to undergo deterministic changes. In addition, the unmeasured disturbances are also assumed to undergo random fluctuations. For mathematical tractability, these are simulated as piecewise constant between each sampling period and changing randomly from their nominal value at each sampling instant. Here, $\mathbf{v}(k)$ and $\mathbf{w}(k)$ are zero mean Gaussian white noise sequences with known covariance matrices. When process is not fully understood or when it is not possible to develop mechanistic models of each component of a system, it is often possible to develop grey box model by combining equations arising from first principles with some black box model components.

Eqs. (1) and (3) represent the normal or fault free behavior of the process, which can be used to develop a state estimator under normal operating conditions. In the present work, the state estimation is carried out using the standard linearized version of EKF [13] as follows:

$$\hat{\mathbf{x}}(k+1|k) = \hat{\mathbf{x}}(k|k) + \int_{kT}^{(k+1)T} \mathbf{F}[\mathbf{x}(\tau), \mathbf{m}(k), \bar{\mathbf{p}}, \bar{\mathbf{d}}]\,\mathrm{d}\tau \tag{4}$$

$$\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1) + \mathbf{L}(k)\gamma(k) \tag{5}$$

It may be noted that we distinguish between the controller output, $\mathbf{m}(k)$, and manipulated input, $\mathbf{u}(k)$, entering the process. Under ideal and fault-free conditions, the controller output equals the manipulated input. The innovation sequence under normal operating condition is computed as

$$\gamma(k) = \mathbf{y}(k) - \mathbf{H}[\hat{\mathbf{x}}(k|k-1)]$$

The state and innovation covariance estimates are updated as follows:

$$\mathbf{P}(k|k-1) = \mathbf{\Phi}(k)\mathbf{P}(k-1|k-1)\mathbf{\Phi}(k)^{\mathrm{T}} + \mathbf{\Gamma}_d(k)\mathbf{Q}\mathbf{\Gamma}_d(k)^{\mathrm{T}} \tag{6}$$

$$\mathbf{V}(k) = \mathbf{C}(k)\mathbf{P}(k|k-1)\mathbf{C}(k)^{\mathrm{T}} + \mathbf{R} \tag{7}$$

$$\mathbf{L}(k) = \mathbf{P}(k|k-1)\mathbf{C}(k)^{\mathrm{T}}[\mathbf{V}(k)]^{-1} \tag{8}$$

$$\mathbf{P}(k|k) = [\mathbf{I} - \mathbf{L}(k)\mathbf{C}(k)]\mathbf{P}(k|k-1) \tag{9}$$

where

$$\mathbf{\Phi}(k) = \exp[\mathbf{A}(k)T]; \quad \mathbf{A}(k) = \left[\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right]_{(\hat{\mathbf{x}}(k|k-1), \mathbf{m}(k-1), \bar{\mathbf{p}}, \bar{\mathbf{d}})}$$

$$\mathbf{C}(k) = \left[\frac{\partial \mathbf{H}(\mathbf{x})}{\partial \mathbf{x}}\right]_{(\hat{\mathbf{x}}(k|k-1))};$$

$$\mathbf{B}_d(k) = \left[\frac{\partial \mathbf{F}}{\partial \mathbf{d}}\right]_{(\hat{\mathbf{x}}(k|k-1), \mathbf{m}(k-1), \bar{\mathbf{p}}, \bar{\mathbf{d}})}; \quad \mathbf{\Gamma}_d(k) = \int_0^{\mathrm{T}} \exp[\mathbf{A}(k)\tau]\mathbf{B}_d(k)\delta\tau$$

Here, $\bar{\mathbf{p}}$ and $\bar{\mathbf{d}}$ are assumed to be the nominal values of parameters and the unmeasured disturbances, respectively. In remainder of the text, we refer to this EKF as *normal EKF*.

2.2. Fault detection

When process starts behaving abnormally, the first task is to detect the deviations from the normal operating conditions. To simplify the task of fault detection, it is further assumed that, under normal operating conditions, the innovation sequence $\{\gamma(k)\}$ is a zero mean Gaussian white noise sequence with covariance $\mathbf{V}(k)$. The whiteness of innovation sequence generated by the *normal EKF* is taken as an indicator of absence of plant-model mismatch. A significant and sustained departure from this behavior is assumed to result from model plant mismatch. To detect such departures systematically, a simple statistical test namely fault detection test (FDT) as given in Prakash et al. [14] is modified based on the innovations obtained from the normal EKF. This test is applied at each time instant to estimate the time of occurrence of a fault. The test statistic for this purpose is given as follows:

$$\epsilon(k) = \gamma(k)^{\mathrm{T}}\mathbf{V}(k)^{-1}\gamma(k) \tag{10}$$

Since it is assumed that innovation sequence is a zero mean Gaussian white noise process, the above test statistic follows a central chi-square distribution [15,16] with r degrees of freedom, which can be used to fix the threshold. Here r is the number of measurements. If FDT is rejected, the occurrence of a fault is further confirmed by examining innovation sequence in the time interval $[t, t+N]$. The test statistic given by Eq. (11) is used for this purpose, which follows a central chi-square distribution with $r(N+1)$ degrees of freedom:

$$\epsilon(t, N) = \sum_{k=t}^{t+N} \gamma(k)^{\mathrm{T}}\mathbf{V}(k)^{-1}\gamma(k) \tag{11}$$

If this test statistic exceeds the threshold, the occurrence of the fault or failure is confirmed. Window size for FDI computations is the tuning parameter. A large size of window reduces false alarms and also improves the magnitude estimates of the fault. However a very high value of N may result in operating the process in a degraded mode for a long time, which may have a deteriorating effect on the performance.

2.3. Fault and failure models

Once the occurrence of a fault is confirmed, the next step is to isolate the fault and estimate its magnitude. To identify the fault(s) that might have occurred, it is necessary to develop a model for each hypothesized fault or failure that describes its effect on the evolution of the process variables. A fault can either develop as an abrupt (step-like) change or as a slow drift from its nominal value. For example, an abrupt change in jth parameter can be modelled as

$$\mathbf{p}_{p_j}(k) = \bar{\mathbf{p}} + b_{p_j}\mathbf{e}_{p_j}\sigma(k-t) \tag{12}$$

Here, b_{p_j} represents change in the parameter value from its nominal value, e_{p_j} represents parameter fault vector with jth element equal to unity and all other elements equal to zero and $\sigma(k-t)$ represents a unit step function defined as

$$\sigma(k-t) = 0 \text{ if } k < t; \quad \sigma(k-t) = 1 \text{ if } k \geqslant t$$

Similarly, if bias occurs in jth sensor at instant t, then, subsequent to the occurrence of bias in the sensor, the behavior of measured outputs is modeled as follows:

$$\mathbf{y}_{y_j}(k) = \mathbf{H}[\mathbf{x}(k)] + b_{y_j}\mathbf{e}_{y_j}\sigma(k-t) + \mathbf{v}(k) \tag{13}$$

On the other hand, if jth unmeasured disturbance changes as a slow drift, then the corresponding fault model

$$\mathbf{d}_{d_j}(k) = \bar{\mathbf{d}} + \mathbf{w}(k) + b_{d_j}\mathbf{e}_{d_j}\varsigma(k-t)$$
$$\varsigma(k-t) = 0 \text{ if } k < t; \quad \varsigma(k-t) = t \text{ if } k \geqslant t \tag{14}$$

Here, b_{d_j} represents the magnitude of the unmeasured disturbance variable change and e_{d_j} represents the corresponding fault location vector.

When an actuator or a sensor fails abruptly, then the models for failure modes have to be developed in a different manner [17]. For example, if jth actuator is stuck abruptly at instant t, then plant input $\mathbf{u}(k)$ subsequent to the failure (denoted as $\mathbf{u}_{u_j}(k)$) can be represented as

$$\mathbf{u}_{u_j}(k) = \mathbf{m}(k) + [b_{u_j} - \mathbf{e}_{u_j}^{\mathrm{T}}\mathbf{m}(k)]\mathbf{e}_{u_j}\sigma(k-t) \tag{15}$$

where b_{u_j} represents constant value at which the jth actuator is stuck. Model given in Eq. (15), indicates that though the controller manipulates $\mathbf{m}(k)$ in the usual manner, the signal going to the plant from an actuator becomes constant due to some fault in the actuator.

When jth sensor fails abruptly at instant t, it is often observed that we get a constant reading close to the value measured by digital to analog converter before the failure occurs. Thus, if jth sensor fails at instant t, we propose to model the behavior of the measurement vector subsequent to the failure as follows:

$$\mathbf{y}_{y_j}(k) = \mathbf{H}[\mathbf{x}(k)] + [b_{y_j} - \mathbf{e}_{y_j}^{\mathrm{T}}\mathbf{H}[\mathbf{x}(k)]]\mathbf{e}_{y_j}\sigma(k-t) + \mathbf{v}(k) \tag{16}$$

where b_{y_j} represents constant value at which the jth sensor reading is stuck. According to Eq. (16), the measurement coming from a particular sensor in a plant gives a signal with constant mean when a sensor fails, though the true plant output is changing.

2.4. Review of linear GLR method

In this work, a new approach has been proposed for fault identification based on EKF. As motivation for developing the new approach is derived from the version of linear GLR, method proposed by Narasimhan and Mah [18] and Willsky and Jones [19], a brief review of their method is presented here. The linear GLR method makes use of the innovation sequences generated by the normal Kalman filter and Kalman filters obtained under different fault assumptions. Let $\{\gamma(t)\ldots\gamma(t+N)\}$ represent the sample of innovation vectors generated by the *normal* Kalman filter over a window for time $[t, t+N]$ after a fault is detected. This innovation sequence obtained from the *normal* Kalman Filter is viewed as a Gaussian random process with unknown means $\mu(k;t)$ and covariance matrices $\mathbf{V}(k;t)$. The hypothesis H_0 for absence of a fault in the observed data can be written as

$$H_0 : \mu(k;t) = 0$$

which is referred to as null hypothesis and the alternate hypothesis H_1 for the presence of a fault in the observed data can be written as

$$H_1 : \mu(k;t) = b_{f_j}\mathbf{G}_{f_j}(k;t)\mathbf{e}_{f_j} + \mathbf{g}_{f_j}(k;t)$$
$$k \in [t, t+N] \quad \text{and} \quad f \in \mathbf{p}, \mathbf{d}, \mathbf{y}, \mathbf{u}$$

where b_{f_j} refers to magnitude of a fault f_j, $\mathbf{G}_{\mathrm{f}}(k;t)$ and $\mathbf{g}_{f_j}(k;t)$ represent fault signature matrix and fault signature vector, respectively, which describe the effect of fault f_j on the innovations. Fault identification (fault location and magnitude estimation) is carried out by maximizing the log-likelihood function

$$T = \sup_{b_{f_j} f_j} T_{f_j} \tag{17}$$

$$T_{f_j} = \sum_{k=t}^{t+N} \gamma^{\mathrm{T}}(k)\mathbf{V}(k)^{-1}\gamma(k) - \inf_{b_{f_j}} \sum_{k=t}^{t+N} \gamma_{f_j}^{\mathrm{T}}(k)\mathbf{V}(k)^{-1}\gamma_{f_j}(k) \tag{18}$$

where $\gamma_{f_j}(k) = [\gamma(k) - b_{f_j}\mathbf{G}_{f_j}(k;t)e_{f_j} - \mathbf{g}_{f_j}(k;t)]$ represents the innovation sequence generated by the *fault mode* Kalman filter developed under the assumption that fault f_j has occurred. Thus, the fault isolation can be viewed as finding the observer that best explains data in window $[t, t+N]$. It may be noted that the first term in Eq. (18) is same for all hypothesized faults. Thus, once a fault is detected, then the fault type together with its magnitude (i.e. b_{f_j} and f_j) is determined by solving the following set of optimization problems:

$$\inf_{b_{f_j} f_j} \sum_{k=t}^{t+N} \gamma_{f_j}^{\mathrm{T}}(k)\mathbf{V}(k)^{-1}\gamma_{f_j}(k) \tag{19}$$

2.5. Nonlinear GLR method

In the GLR method, linearity of the process model/observers and additive nature of faults can be exploited to develop recursive relationships between the innovation sequences generated by the *normal* Kalman filter and *fault mode* Kalman filters developed under different fault assumptions (ref. [14] for details). However, for a general nonlinear system governed by Eqs. (1) and (3), most of the faults affect system dynamics in a nonlinear manner and similar recurrence relationships cannot be derived. Thus, to develop a nonlinear analog of the GLR method, we formulate a separate EKF for each hypothesized fault model over the time window $[t, t+N]$, with the assumption that a fault has occurred at time instant t. We then pose the problem of fault isolation as "finding the *fault mode observer* that best explains the measurement sequence $\{\mathbf{y}(t)\ldots\mathbf{y}(t+N)\}$ collected over a window for time $[t, t+N]$".

To understand the proposed FDI method, consider an observer developed under the assumption that an actuator has failed. Assuming that actuator j has failed at instant t, the process behavior over window $[t, t+N]$ can be described as follows:

$$\mathbf{x}_{u_j}(i+1) = \mathbf{x}_{u_j}(i) + \int_{iT}^{(i+1)T} \mathbf{F}[\mathbf{x}_{u_j}(\tau), \mathbf{u}_{u_j}(i), \mathbf{p}, \mathbf{d}]\,\mathrm{d}\tau \tag{20}$$

$$\mathbf{y}_{u_j}(i) = \mathbf{H}[\mathbf{x}_{u_j}(i)] + \mathbf{v}(i) \tag{21}$$

where $\mathbf{u}_{u_j}$ is given by Eq. (15). Assuming that a fault has occurred at t, the corresponding fault mode observer can be formulated as follows:

$$\hat{\mathbf{x}}_{u_j}(i|i-1) = \hat{\mathbf{x}}_{u_j}(i-1|i-1) + \int_{(i-1)T}^{(i)T} \mathbf{F}[\mathbf{x}_{u_j}(\tau), \mathbf{m}_{u_j}(i-1), \bar{\mathbf{p}}, \bar{\mathbf{d}}]\,\mathrm{d}\tau \tag{22}$$

$$\mathbf{m}_{u_j}(i) = \mathbf{m}(i) + [b_{u_j} - \mathbf{e}_{u_j}^{\mathrm{T}}\mathbf{m}(i)]\mathbf{e}_{u_j} \tag{23}$$

$$\hat{\mathbf{x}}_{u_j}(i|i) = \hat{\mathbf{x}}_{u_j}(i|i-1) + \mathbf{L}_{u_j}(i)\gamma_{u_j}(i) \tag{24}$$

$$\gamma_{u_j}(i) = \mathbf{y}(i) - \mathbf{H}[\hat{\mathbf{x}}_{u_j}(i|i-1)] \tag{25}$$

$$\hat{\mathbf{x}}_{u_j}(t|t) = \hat{\mathbf{x}}(t|t) \tag{26}$$

where $i \in [t, t+N]$ and $\mathbf{L}_{u_j}(i)$ represents the Kalman Gain matrices for this *fault mode observer* computed using equations of the form (6)–(9). For each hypothesized fault, a separate fault mode observer is developed in a similar manner over window $[t, t+N]$.

The next step is to generate estimates of the parameters of the fault model for each hypothesized fault. Taking motivation from linear GLR method, the fault magnitude estimation problem is formulated as a nonlinear optimization problem as follows:

$$\min_{b_{f_j}}(\mathbf{J}_{f_j}) = \sum_{i=t}^{t+N} \gamma_{f_j}^{\mathrm{T}}(i)\mathbf{V}_{f_j}(i)^{-1}\gamma_{f_j}(i) \tag{27}$$

where $\gamma_{f_j}(i)$ and $\mathbf{V}_{f_j}(i)$ are the innovations and the innovations covariance matrices, respectively, computed using the *fault mode* observer corresponding to fault f_j. The *fault mode observer* that best explains the measurement sequence $\{\mathbf{y}(t)\ldots\mathbf{y}(t+N)\}$ is one for which the value of $(\mathbf{J}_{f_j})$ is minimum. Thus, the fault f_j that corresponds to

$$\min_{f_j\in(\mathbf{p},\mathbf{d},\mathbf{y},\mathbf{u})}[(\mathbf{J}_{f_j})] \tag{28}$$

is isolated as the fault that has occurred at time t and its corresponding magnitude estimate $\hat{b}_{f_j}$ is taken as fault magnitude. This proposed approach for fault identification, which is motivated by linear GLR method, is referred to as Nonlinear GLR (NL-GLR) method in the rest of the text.

2.6. Simplification of NL-GLR method

The NL-GLR method proposed above involves solving multiple nonlinear optimization problems, which are subjected to nonlinear ODE constraints. Such NLPs are notoriously difficult to solve and computationally demanding from the viewpoint of on-line implementation. To simplify the task of on-line fault isolation, we propose a simplified version of the proposed NL-GLR method. This simplification is based on linearization of the nonlinear process model along a *nominal trajectory* defined as follows:

$$\{[\hat{\mathbf{x}}(i|i),\mathbf{m}(i-1),\bar{\mathbf{d}},\bar{\mathbf{p}}]: i\in[t-1,t+N-1]\}$$

which is generated using the *Normal EKF* under the assumption that no fault has occurred over window $[t,t+N]$. Now, for small magnitude faults, the system dynamics under faulty conditions $\{\mathbf{x}_{f_j}(i): i\in[t,t+N]\}$ can be viewed as deviation from the nominal trajectory generated by *Normal EKF*. Under the hypothesis of occurrence of fault f_j, let the deviation in the state estimates from the *nominal trajectory* be represented as

$$\delta\mathbf{x}_{f_j}(i)=\mathbf{x}_{f_j}(i)-\hat{\mathbf{x}}(i|i) \tag{29}$$

Then, using the Taylor series expansion in the neighborhood of the nominal trajectory and neglecting higher order terms, a time varying linear perturbation model in the neighborhood of the nominal trajectory can be obtained as follows:

$$\delta\mathbf{x}_{f_j}(i+1)=\varkappa(i)+\mathbf{\Phi}(i)\delta\mathbf{x}_{f_j}(i)+\mathbf{\Gamma}_u(i)\delta\mathbf{m}(i)+\mathbf{\Gamma}_d(i)\delta\mathbf{d}+\mathbf{\Gamma}_p(i)\delta\mathbf{p}+\mathbf{\Gamma}_d(i)\mathbf{w}(i) \tag{30}$$

$$\delta\mathbf{y}_{f_j}(i)=\mathbf{C}(i)\delta\mathbf{x}_{f_j}(i)+\mathbf{v}(i) \tag{31}$$

$$\delta\mathbf{m}(i)=\mathbf{m}(i)-\mathbf{m}(i-1) \tag{32}$$

$$\mathbf{y}_{f_j}(i)=\mathbf{C}(i)\hat{\mathbf{x}}(i|i)+\delta\mathbf{y}_{f_j}(i) \tag{33}$$

$$i\in[t-1,t+N-1]$$

where $\delta\mathbf{d}$ and $\delta\mathbf{p}$ represent vectors of abrupt changes in unmeasured disturbances and model parameters from their nominal values, respectively. The time varying vector $\varkappa(i)$ and matrices $\mathbf{\Phi}(i),\mathbf{\Gamma}_u(i),\mathbf{\Gamma}_d(i),\mathbf{\Gamma}_p(i),\mathbf{C}(i)$ appearing in the above set of equations are computed by linearizing the *normal* process model along the nominal trajectory as follows:

$$\varkappa(i)=\int_0^{\mathrm{T}}\exp(\mathbf{A}(i)q)\mathbf{F}[\hat{\mathbf{x}}(i|i),\mathbf{m}(i-1),\bar{\mathbf{p}},\bar{\mathbf{d}}]\,\mathrm{d}q \tag{34}$$

$$\mathbf{\Phi}(i)=\exp[\mathbf{A}(i)T];\quad \mathbf{A}(i)=\left[\frac{\partial\mathbf{F}}{\partial\mathbf{x}}\right]_{(\bullet)};\quad \mathbf{C}(i)=\left[\frac{\partial\mathbf{H}}{\partial\mathbf{x}}\right]_{(\bullet)} \tag{35}$$

$$\mathbf{\Gamma}_u(i)=\int_0^{\mathrm{T}}\exp(\mathbf{A}(i)q)\mathbf{B}_u(i)\,\mathrm{d}q;\quad \mathbf{B}_u(i)=\left[\frac{\partial\mathbf{F}}{\partial\mathbf{m}}\right]_{(\bullet)} \tag{36}$$

$$\mathbf{\Gamma}_p(i)=\int_0^{\mathrm{T}}\exp(\mathbf{A}(i)q)\mathbf{B}_p(i)\,\mathrm{d}q;\quad \mathbf{B}_p(i)=\left[\frac{\partial\mathbf{F}}{\partial\mathbf{p}}\right]_{(\bullet)} \tag{37}$$

$$\mathbf{\Gamma}_d(i)=\int_0^{\mathrm{T}}\exp(\mathbf{A}(i)q)\mathbf{B}_d(i)\,\mathrm{d}q;\quad \mathbf{B}_d(i)=\left[\frac{\partial\mathbf{F}}{\partial\mathbf{d}}\right]_{(\bullet)} \tag{38}$$

$$(\bullet)\equiv(\hat{\mathbf{x}}(i|i),\mathbf{m}(i-1),\bar{\mathbf{p}},\bar{\mathbf{d}}) \tag{39}$$

Based on the above perturbation model a separate linearized observer is formulated over a time window $[t,t+N]$ for each hypothesized fault. For example, consider a case where jth actuator has failed. Let b_{u_j} denote the magnitude at which the jth input is stuck. Assuming that the fault has occurred at t, the linearized observer can be formulated as follows:

$$\delta\hat{\mathbf{x}}_{u_j}(i|i-1)=\varkappa(i-1)+\mathbf{\Phi}(i-1)\delta\hat{\mathbf{x}}_{u_j}(i-1|i-1)+\mathbf{\Gamma}_u(i-1)\delta\mathbf{m}_{u_j}(i-1) \tag{40}$$

$$\hat{\gamma}_{u_j}(i)=\delta\mathbf{y}(i)-\mathbf{C}(i)\delta\hat{\mathbf{x}}_{u_j}(i|i-1) \tag{41}$$

$$\delta\hat{\mathbf{x}}_{u_j}(i|i)=\delta\hat{\mathbf{x}}_{u_j}(i|i-1)+\mathbf{L}(i)\hat{\gamma}_{u_j}(i) \tag{42}$$

$$\delta\mathbf{m}_{u_j}(i)=\mathbf{m}_{u_j}(i)-\mathbf{m}(i-1) \tag{43}$$

$$=\delta\mathbf{m}(i)+[b_{u_j}-\mathbf{e}_{u_j}^{\mathrm{T}}\mathbf{m}(i)]\mathbf{e}_{u_j} \tag{44}$$

$$\delta\mathbf{y}(i)=\mathbf{y}(i)-\mathbf{C}(i)\hat{\mathbf{x}}(i|i) \tag{45}$$

for $i\in[t,t+N]$ starting from the initial condition $\delta\hat{\mathbf{x}}_{u_j}(t-1\mid t-1)=\bar{0}$. It may be noted that the Kalman gain matrices $\{\mathbf{L}(i): i\in[t,t+N]\}$ obtained from the *normal EKF* are used for state correction. Also, the time varying matrices $\mathbf{\Phi}(i),\mathbf{\Gamma}_u(i)$, and $\mathbf{C}(i)$ have to be computed only once by this approach, which significantly reduces on-line computational burden. The fault magnitude for each hypothesized fault is estimated from the following optimization problem:

$$\min_{b_{f_j}}(\mathbf{J}_{f_j})=\sum_{i=t}^{t+N}\hat{\gamma}_{f_j}^{\mathrm{T}}(i)[\mathbf{V}(i)]^{-1}\hat{\gamma}_{f_j}(i) \tag{46}$$

$$f\in y,p,d,u \tag{47}$$

where $\mathbf{V}(i)$ is the covariance matrix for the innovations from the *normal EKF* and $\gamma_{f_j}(i)$ is the innovation sequence from the linearized observer under fault hypothesis f_j. The fault isolation can now be carried out by finding fault f_j that corresponds to minimum value of $\mathbf{J}_{f_j}$.

Once a fault f_j is isolated, a refined estimate of the fault magnitude is generated by formulating a nonlinear optimization problem (27) as described in the previous sub-section. This simplification of NL-GLR, referred to as SNL-GLR in the rest of the text, reduces the on-line computational burden significantly. The nonlinear optimization is carried out only once for refinement of fault magnitude estimate for the fault that has been isolated.

2.7. Multiple simultaneous faults

In the previous section, the proposed FDI method has been described for the case when a single (root cause) fault occurs. If multiple (root cause) faults occur simultaneously (i.e. at the same time instant), the above formulation can be extended to isolate and estimate magnitudes of multiple simultaneous faults as follows. In this case, we propose to enumerate all possible combinations of multiple faults that can occur simultaneously and develop fault mode observers for each hypothesized combination. For example if simultaneous faults are to be hypothesized in jth sensor and lth parameter, then the observer for this combination of faults can be formulated as follows:

$$\hat{\mathbf{x}}_{(p_l y_j)}(i|i-1)=\hat{\mathbf{x}}_{(p_l y_j)}(i-1|i-1)+\int_{(i-1)T}^{iT}\mathbf{F}[\hat{\mathbf{x}}_{(p_l y_j)}(t),\mathbf{m}(i),(\bar{\mathbf{p}}+b_{p_l}\mathbf{e}_{p_l}),\bar{\mathbf{d}}]\,\mathrm{d}t \tag{48}$$

$$\hat{\mathbf{x}}_{(p_l y_j)}(i|i)=\hat{\mathbf{x}}_{(p_l y_j)}(i|i-1)+\mathbf{L}_{(p_l y_j)}(i)\gamma_{(p_l y_j)}(i) \tag{49}$$

$$\gamma_{(p_l y_j)}(i)=\mathbf{y}(i)-[\mathbf{H}[\hat{\mathbf{x}}_{(p_l y_j)}(i)]+b_{y_j}\mathbf{e}_{y_j}] \tag{50}$$

$$\text{for } i\in[t,t+N]\quad\text{with}\quad \hat{\mathbf{x}}_{(p_l y_j)}(t-1|t-1)=\hat{\mathbf{x}}(t-1|t-1) \tag{51}$$

Fault magnitude estimation problem for each hypothesized combination is then formulated and solved similar to formulation (27) discussed in Section 2.5. However, when multiple simultaneous faults are hypothesized together with single faults the fault models have unequal number of unknown parameters (i.e. different degrees of freedom). Consequently, the fault isolation step cannot be carried out using the minimum value of $\mathbf{J}_{f_j}$ as described in Section 2.5. To

alleviate this difficulty, we propose to use Akaike Information Criterion (AIC) for fault isolation when multiple simultaneous faults are hypothesized together with single faults. Thus, the test statistic used for fault isolation is as follows:

$$\min_{f_j \in \text{ all hypothesized faults}} (\mathbf{AIC})_{f_j} = N \ln\left[\frac{1}{N}(\mathbf{J}_{f_j})\right] + 2\phi \tag{52}$$

where $\mathbf{J}_{f_j}$ represents the value of the prediction error term obtained after solving the magnitude estimation problem and ϕ represents the total number of parameters estimated when fault f_j has occurred. The fault, i.e. either a single fault or a set of simultaneous faults, that yields minimum value for AIC is isolated as the fault that has occurred. It may be noted that

- The proposed fault isolation strategy based on AIC can also be employed when fault models with different number of unknown parameters (e.g. step jump in a parameter and slow drift in the parameter) are hypothesized.
- The number of faults that can be hypothesized to occur simultaneously cannot exceed the number of measurements due to observability considerations.

Remark 1. It may be noted that the choice of window length N determines the trade-off between 'delay in diagnosis' and 'accuracy of diagnosis'. A large value of N results in less false alarms and smaller variance errors in fault magnitude estimate. However, choosing larger N also introduces a longer delay in fault identification. On the other hand, choosing N to be small reduces delay in fault isolation. However, it can increase false alarms and results in larger variance errors in fault magnitude estimates. Based on simulation studies, Prakash et al. [14] have suggested that the window length N can be chosen approximately equal to half the time required for the estimator to converge after a change occurs.

3. Intelligent state estimation for fault tolerant NMPC

NMPC techniques use nonlinear model for prediction, which is typically developed once in the beginning of implementation of an NMPC scheme. However, as time progresses, slow changes in unmeasured disturbances and/or process parameters and *faults* such as biases in sensors or actuators results in significant mismatch in plant and model behavior (*behavior mismatch*). In addition, hard failures, like failures of actuators and sensors can lead to significant structural plant model mismatch (*structure mismatch*). The conventional approach to deal with the *behavior mismatch* in the NMPC formulations is through the introduction of additional artificial states in the state observer [2–4]. The main limitation of this permanent augmentation approach is that the number of extra states that can be introduced cannot exceed the number of measurements. This implies that it is necessary to have *a priori* knowledge of a subset of faults that are most likely to occur or a subset of parameters that are most likely to drift. In such a formulation, the state estimates can become biased when unanticipated faults and/or parameter drifts occur. When NMPC formulation is used for inferential control of some unmeasured quality variables, the biased state estimates can have detrimental effect on the closed loop performance. The accuracy of the state estimates, which is the prime concern in the inferential control formulation, can be maintained only if identical model is used for fault diagnosis and control and the model is corrected at the correct location when a fault or abrupt change occurs [12]. Moreover, the permanent augmentation of state space model cannot systematically deal with the difficulties arising out of sensor biases and sensor/actuator failures.

In this section, we describe the integration of the conventional state space based NMPC formulation with NL-GLR or SNL-GLR scheme, which is capable of generating unbiased state estimates by intelligently correcting the state estimator. To begin with, we describe the modifications necessary in the state estimator used for FDI a well as NMPC, when a fault is detected for the first time by FDI component. Modifications necessary for dealing with recurrence of the fault, occurrence of another fault at subsequent time instants and drifting (non-stationary) changes in unmeasured disturbances/model parameters are described later. We then proceed to propose NMPC formulation that can deal with behavioral as well as structural changes in the model and state estimator. A schematic representation of the proposed FTNMPC scheme is shown in Fig. 1.

3.1. On-line modifications to state estimator and predictor

Consider a situation where FDT has been rejected at time instant t and subsequently FCT has been rejected at time $t+N$ for the first time. Further assume that at instant $t+N$ a fault f_j has been isolated using NL-GLR/SNL-GLR method and the fault magnitude has been estimated using data collected in the interval $[t, t+N]$. During the interval $[t, t+N]$, the NMPC formulation is based on the prediction model given by equations

$$\hat{\mathbf{x}}(k+l+1|k) = \hat{\mathbf{x}}(k+l|k) + \int_{(k+l)T}^{(k+l+1)T} \mathbf{F}[\hat{\mathbf{x}}(t), \mathbf{m}(k+l|k), \bar{\mathbf{p}}, \bar{\mathbf{d}}, t]dt \tag{53}$$

However after the identification of the fault at instant $t+N$, the model for $k \geqslant t+N$ is modified as follows:

- *Step jump in unmeasured disturbance/model parameter*: When FDI component isolates abrupt change in unmeasured disturbance, the prediction equations in the state estimator and future predictions in NMPC are modified as follows:

$$\hat{\mathbf{x}}(k|k-1) = \hat{\mathbf{x}}(k-1|k-1) + \int_{(k-1)T}^{kT} \mathbf{F}[\hat{\mathbf{x}}(t), \mathbf{m}(k-1), \bar{\mathbf{p}}, \bar{\mathbf{d}} + \hat{b}_{d_j}\mathbf{e}_{d_j}]\,\mathrm{d}t \tag{54}$$

$$\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1) + \mathbf{L}(k)\gamma(k) \tag{55}$$

$$\hat{\mathbf{x}}(k+l+1|k) = \hat{\mathbf{x}}(k+l|k) + \int_{(k+l)T}^{(k+l+1)T} \mathbf{F}[\hat{\mathbf{x}}(k+l|k), \mathbf{m}(k), \bar{\mathbf{p}}, \bar{\mathbf{d}} + \hat{b}_{d_j}\mathbf{e}_{d_j}]\,\mathrm{d}t$$

If an abrupt change is detected in a parameter, then the state estimator and predictor can be modified in analogous manner.

- *Sensor faults*: If jth sensor bias is isolated, the measured output is compensated as

$$\mathbf{y}_c(k) = \mathbf{y}(k) - \hat{b}_{y_j}\mathbf{e}_{y_j} \tag{56}$$

and used in FDI as well as NMPC formulation for computing innovation sequence.

- *Compensation for actuator bias*:

$$\hat{\mathbf{x}}(k|k-1) = \hat{\mathbf{x}}(k-1|k-1) + \int_{(k-1)T}^{kT} \mathbf{F}[\hat{\mathbf{x}}(t), \mathbf{m}(k-1) + \hat{b}_{u_j}\mathbf{e}_{u_j}, \bar{\mathbf{p}}, \bar{\mathbf{d}}]\,\mathrm{d}t \tag{57}$$

$$\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1) + \mathbf{L}(k)\gamma(k) \tag{58}$$

$$\hat{\mathbf{x}}(k+l+1|k) = \hat{\mathbf{x}}(k+l|k) + \int_{(k+l)T}^{(k+l+1)T} \mathbf{F}[\hat{\mathbf{x}}(t), \mathbf{m}(k+l|k) + \hat{b}_{u_j}\mathbf{e}_{u_j}, \bar{\mathbf{p}}, \bar{d}]\,\mathrm{d}t \tag{59}$$

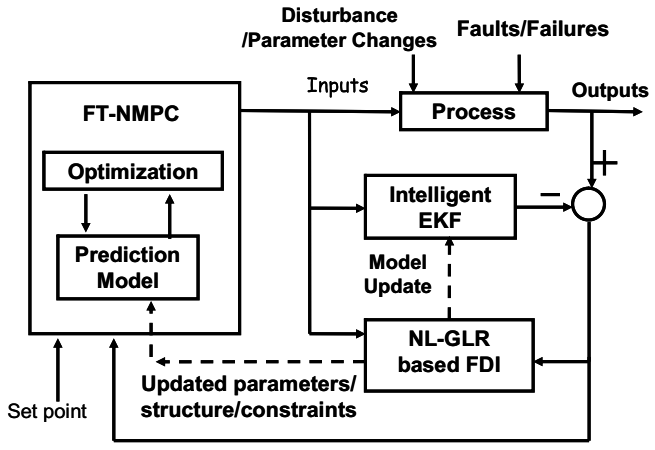

Fig. 1. Fault tolerant NMPC (FTNMPC): schematic representation.

- *Failed actuator*: In state estimation and prediction, the failed actuator is treated as constant $\mathbf{m}_j(k) = \hat{b}_{u_j}$ for $k \geqslant t+N$, where $\hat{b}_{u_j}$ is the estimate of stuck actuator signal for jth actuator.
- *Failed sensor*: When the FDI component isolates a sensor failure, the failed sensor measurement is removed from the set of measurements used for state estimation.

After a fault, say f_j, is diagnosed and corrections are made to the state estimator, it becomes necessary to correct the state estimates while re-starting the state estimator at $k = t+N$ based on the modified model. The state vector and state error covariance matrix estimated at $k = t+N$ in the magnitude refinement step is used to re-start the state estimation with the modified model and state observer by setting

$$\hat{\mathbf{x}}(t+N|t+N) = \hat{\mathbf{x}}_{f_j}(t+N|t+N)$$
$$\mathbf{P}(t+N|t+N) = \mathbf{P}_{f_j}(t+N|t+N)$$

and these values are used subsequently for state propagation and covariance update.

3.2. Correction for drifting disturbances, parameters and multiple sequential faults

The main concern with the above approach is that the magnitude and the position of the fault may not be accurately estimated. Thus, there is a need to introduce integral action in such a way that the errors in estimation of fault magnitude or position can be corrected in the course of time. Furthermore, other faults may occur at subsequent times. Thus, in the on-line implementation of NMPC, application of FDI method resumes at $t+N+1$. The FDI method may identify a fault in the previously identified location or a new fault/parameter change/mean shift in unmeasured disturbance may be identified. In either case, the Eqs. (54)–(59) can be modified using cumulative estimate of the corresponding biases [14], which are computed as

$$\tilde{b}_{f_j} = \sum_{l=1}^{n_f} \hat{b}_{f_j}(l) \quad \text{with initial value } \hat{b}_{f_j}(0) = 0 \tag{60}$$

where $f \in [u, y]$ denotes the fault type occurring at jth position and n_f represents the number of times a fault of type f was confirmed and isolated in the jth position. Similarly, cumulative unmeasured disturbance vector, $\tilde{\mathbf{d}}(t_{d_l})$, can be defined as follows:

$$\tilde{\mathbf{d}}(t_{d_l}) = \tilde{\mathbf{d}}(t_{d_{l-1}}) + \sum_{j=1}^{d} \hat{b}_{d_j}(t_{d_l})\mathbf{e}_{d_j}$$
$$\tilde{\mathbf{d}}(0) = \bar{\mathbf{d}} \tag{61}$$

where t_{d_l} represent the last time instant when unmeasured disturbance fault was isolated, $t_{d_{l-1}}$ represent the time instant previous to t_{d_l} when such fault was isolated and $\hat{\mathbf{d}}(t_{d_l})$ represents the fault vector (point) estimate at time instant t_{d_l}. Cumulative parameter vector can be defined in a similar manner as follows:

$$\tilde{\mathbf{p}}(t_{p_l}) = \tilde{\mathbf{p}}(t_{p_{l-1}}) + \sum_{j=1}^{p} \hat{b}_{p_j}(t_{p_l})\mathbf{e}_{p_j}$$
$$\tilde{\mathbf{p}}(t_{p_l}) = \bar{\mathbf{p}} \tag{62}$$

The cumulative bias estimates given by Eqs. (60)–(62) are used in Eqs. (54)–(59) in place of the point estimates $\hat{b}_{f_j}\mathbf{e}_{f_j}, \bar{\mathbf{d}} + \hat{b}_{d_j}\mathbf{e}_{d_j}$ and $\bar{\mathbf{p}} + \hat{b}_{p_j}\mathbf{e}_{p_j}$. The use of cumulative bias estimates can be looked upon as a method of introducing integral action to account for plant model mismatch, in which some of the states (cumulative bias estimates) are integrated at much slower rate and at irregular sampling intervals. It may be noted that the use of cumulative bias estimates to *correct* the EKF also implies that the definition of *normal behavior model* keeps changing as and when faults are detected and subsequently the model is compensated for the faults. Thus, after sufficiently long time, the *normal behavior* model used for state estimation and fault diagnosis can be represented as follows:

$$\mathbf{x}(k+1) = \mathbf{x}(k) + \int_{kT}^{(k+1)T} \mathbf{F}[\mathbf{x}(t), \tilde{\mathbf{u}}(k), \tilde{\mathbf{p}}(t_{p_l}), \tilde{\mathbf{d}}(t_{d_l}) + \mathbf{w}_d(k)]\,\mathrm{d}t \tag{63}$$

$$\tilde{\mathbf{u}}(k) = \mathbf{m}(k) + \sum_{j=1}^{m} \tilde{b}_{u_j}\mathbf{e}_{u_j} \tag{64}$$

$$\mathbf{y}(k) = \mathbf{H}[\mathbf{x}(k)] - \sum_{j=1}^{r} \tilde{b}_{y_j}\mathbf{e}_{y_j} + \mathbf{v}(k) \tag{65}$$

provided no actuator/sensor failures are diagnosed. When a new fault is detected, the on-line diagnosis problem is now formulated as follows:

$$\inf_{\hat{b}_{f_j}(t_{f_{l+1}}), f_j} \sum_{k=t_{f_{l+1}}}^{t_{f_{l+1}}+N} \gamma_{f_j}^{\mathrm{T}}(k)\mathbf{V}(k)^{-1}\gamma_{f_j}(k) \tag{66}$$

$$\hat{\mathbf{x}}_{f_j}(i|i-1) = \hat{\mathbf{x}}_{f_j}(i-1|i-1) + \int_{(i-1)T}^{(i)T} \mathbf{F}[\mathbf{x}_{f_j}(\tau), \tilde{\mathbf{u}}(i-1), \tilde{\mathbf{p}}(t_{p_l}), \tilde{\mathbf{d}}(t_{d_l}), \hat{b}_{f_j}(t_{f_{l+1}})\mathbf{e}_{f_j}]\,\mathrm{d}\tau \tag{67}$$

$$\hat{\mathbf{x}}_{f_j}(i|i) = \hat{\mathbf{x}}_{f_j}(i|i-1) + \mathbf{L}_{f_j}(i)\gamma_{f_j}(i) \tag{68}$$

$$\gamma_{f_j}(i) = \mathbf{y}(i) - \left\{\mathbf{H}[\hat{\mathbf{x}}_{f_j}(i|i-1)] - \sum_{j=1}^{r} \tilde{b}_{y_j}\mathbf{e}_{y_j}\right\} \tag{69}$$

$$\hat{\mathbf{x}}_{f_j}(t|t) = \hat{\mathbf{x}}(t|t) \tag{70}$$

where $i \in [t_{f_l}, t_{f_{l+1}} + N]$ and $\hat{b}_{f_j}(t_{f_{l+1}})\mathbf{e}_{f_j}$ influence the system dynamics through the cumulative bias expressions given by Eqs. (60)–(62). In abstract form, if θ represents the set of all corrections that are made to the model subsequent to diagnosis, then the above formulation followed by the model correction (i.e. fault accommodation) step is equivalent to a slow rate recursion of the form:

$$\boldsymbol{\theta}(t_{f_{l+1}}) = \Psi[\boldsymbol{\theta}(t_{f_l}), \Omega(t_{f_{l+1}}, t_{f_{l+1}} + N)]$$
$$\Omega(t_{f_{l+1}}, t_{f_{l+1}} + N) = \{[\mathbf{y}(i), \mathbf{u}(i)] : i \in [t_{f_l}, t_{f_{l+1}} + N]\}$$

where t_{f_l} represent the last time instant when a fault was isolated, $t_{f_{l-1}}$ represent the time instant previous to t_{f_l} when a fault was isolated and $\Psi[\cdot]$ represents update rule through NL-GLR. This is tantamount to using all the data collected after each fault detection for updating the model. As a consequence, the use of cumulative bias estimate improves parameter/bias estimates and reduces the variance errors if a fault is isolated in the same location multiple times. In fact, Eqs. (60)–(65) together represent a multi-rate model with expected values of unmeasured disturbances and parameters changing at a significantly slower and irregular sampling rates.

The minimum gap between two such changes equals the window length used for fault isolation. Thus, this model effectively separates unmeasured disturbances into two components: (a) stationary colored noise modelled through extended Kalman filter; (b) non-stationary low frequency mean changes captured through $\tilde{\mathbf{d}}(t_{d_l})$ and $\tilde{\mathbf{p}}(t_{p_l})$. The above self-adapting form of model with slowly time varying parameters is used in the proposed NMPC formulation.

Remark 2. It may be noted that Eqs. (61) and (62) slowly model drifting disturbances as sequence of step changes. However, it is likely that in some situations drift model given Eq. (14) is more appropriate and has to be used. The mechanism for model correction has to be suitably modified when a fault is modelled as a drift. If the mean value of some unmeasured disturbance/parameter changes continuously and at a much faster rate, then the time scale separation that is implicit in the proposed formulation may not be acceptable. In such a situation, that specific parameter or unmeasured disturbance variable can be included in the state vector and its value can be estimated together with the other states. While such permanent augmentation will reduce a degree of freedom available for diagnosis, the proposed approach can still be used for diagnosing remaining faults without requiring any significant modifications.

3.3. Fault tolerant NMPC formulation

At any sampling instant k, the nonlinear model predictive control problem is formulated as a constrained optimization problem whereby the future manipulated input moves denoted as

$$\{\mathbf{m}(k|k), \mathbf{m}(k+1|k) \ldots \mathbf{m}(k+N_p-1|k)\}$$

are determined by minimizing an objective function involving predicted controller errors. Typical objective function used in an NMPC formulation is of the form

$$\min_{\mathbf{m}(k|k),\mathbf{m}(k+1|k)\ldots\mathbf{m}(k+N_p-1|k)} \mathbf{J} = \mathbf{J}_{\mathbf{e}} + \mathbf{J}_{\Delta\mathbf{m}} \tag{71}$$

$$\mathbf{J}_{\mathbf{e}} = \sum_{l=1}^{N_p} \mathbf{e}_{\mathrm{f}}(k+l|k)^{\mathrm{T}} \mathbf{W}_E \mathbf{e}_{\mathrm{f}}(k+l|k) \tag{72}$$

$$\mathbf{J}_{\Delta\mathbf{m}} = \sum_{l=0}^{q-1} \Delta\mathbf{m}(k+l|k)^{\mathrm{T}} \mathbf{W}_m \Delta\mathbf{m}(k+l|k) \tag{73}$$

subject to following constraints:

$$\begin{aligned}
&\mathbf{e}_{\mathrm{f}}(k+l|k) = \mathbf{y}_r(k) - \hat{\mathbf{y}}(k+l|k)\\
&\Delta\mathbf{m}(k+l|k) = \mathbf{m}(k+l|k) - \mathbf{m}(k+l-1|k)\\
&\hat{\mathbf{x}}(k+l+1|k) = \hat{\mathbf{x}}(k+l|k)\\
&\qquad + \int_{(k+l)T}^{(k+l+1)T} \mathbf{F}[\hat{\mathbf{x}}(t), \mathbf{m}(k+l|k)\\
&\qquad + \sum_{j=1}^{m} \tilde{b}_{u_j} \mathbf{e}_{u_j}, \tilde{\mathbf{p}}(t_{p_l}), \tilde{\mathbf{d}}(t_{d_l}), t]\,\mathrm{d}t
\end{aligned} \tag{74}$$

$$\tilde{\mathbf{y}}(k+l|k) = \mathbf{G}[\hat{\mathbf{x}}(k+l|k)] \tag{75}$$

$$\begin{aligned}
&\mathbf{m}^L \leqslant \mathbf{m}(k+l|k) + \sum_{j=1}^{m} \tilde{b}_{u_j} \mathbf{e}_{u_j} \leqslant \mathbf{m}^U\\
&\Delta\mathbf{m}^L \leqslant \Delta\mathbf{m}(k+l|k) \leqslant \Delta\mathbf{m}^U
\end{aligned}$$

where $l \in [0, N_p]$. Here, N_p represents prediction horizon, q represents control horizon, $\mathbf{y}_r(k)$ represents the future setpoint trajectory and $\tilde{\mathbf{y}} = \mathbf{G}[\mathbf{x}]$ represents the vector of controlled outputs, which may, in general, differ from the measured outputs $\mathbf{y} = \mathbf{H}[\mathbf{x}]$. It may be noted that constraints on manipulated input are modified to accommodate bises in manipulated inputs. While the above modified model can deal with faults, actuator failure may require additional measure such as modification of the control objective to accommodate the failure. For example, if dimension of the setpoint vector equals the number of manipulated inputs and an actuator failure is diagnosed, then, the NMPC objective function is modified by relaxing setpoint on one of the controlled outputs.

In an ideal situation where all the behavioral changes in the plant are detected and isolated by the proposed FDI scheme, NMPC formulated using model (71)–(75) can provide offset free control. However, since we are dealing with a stochastic system, all faults/changes that occur in the plant may not get diagnosed correctly. To achieve offset free control in such a scenario, the prediction equations can be modified as follows:

$$\begin{aligned}
\tilde{\mathbf{x}}(k+l+1|k) = \hat{\mathbf{x}}(k+l|k) + \int_{(k+l)T}^{(k+l+1)T} &\mathbf{F}[\hat{\mathbf{x}}(t), \mathbf{m}(k+l|k)\\
&+ \sum_{j=1}^{m} \tilde{b}_{u_j} \mathbf{e}_{u_j}, \tilde{\mathbf{p}}(t_{p_l}), \tilde{\mathbf{d}}(t_{d_l}), ,t]\,\mathrm{d}t
\end{aligned} \tag{76}$$

$$\hat{\mathbf{x}}(k+l+1|k) = \tilde{\mathbf{x}}(k+l+1|k) + \mathbf{L}(k)\gamma(k) \tag{77}$$

$$\hat{\mathbf{y}}(k+l|k) = \mathbf{G}[\hat{\mathbf{x}}(k+l|k)] + \varepsilon(k) \tag{78}$$

$$\varepsilon(k) = \mathbf{y}(k) - \hat{\mathbf{y}}(k|k) \tag{79}$$

where $\hat{\mathbf{y}}(k/k)$ is computed using Eq. (75). When the sets of measured and controlled outputs are identical, this simple modification in the prediction equation can eliminate offset without requiring state augmentation [20].

3.4. Closed loop stability

The above finite horizon formulation of NMPC does not guarantee closed loop stability even under nominal conditions. Chen and Allgower [1] have shown that inclusion of terminal weighting in the NMPC objective function (quasi-infinite horizon formulation) can guarantee asymptotic closed loop stability under nominal conditions in the absence of any unmeasured disturbances. In the quasi-infinite horizon formulation, the NMPC objective function is modified by including the following additional term:

$$\mathbf{J} = \mathbf{J}_{\mathbf{e}} + \mathbf{J}_{\Delta\mathbf{m}} + \mathbf{J}_{\infty} \tag{80}$$

$$\mathbf{J}_{\infty} = \mathbf{x}(k+N_p|k)^{\mathrm{T}} \mathbf{W}_{\infty} \mathbf{x}(k+N_p|k) \tag{81}$$

where $\mathbf{W}_{\infty}$ represents the terminal state penalty matrix, which is computed by solving an appropriate Lyapunov equation (ref. Appendix). In addition, the predicted state $\mathbf{x}(k+p|k)$ is constrained to lie within a terminal set Ω_x defined as

$$\Omega_x := \{\mathbf{x} \in R^n | \mathbf{x}^{\mathrm{T}} \mathbf{W}_{\infty} \mathbf{x} \leqslant \alpha\}$$

in the neighborhood of the operating steady state. If the Jacobian linearization of the nonlinear system to be controlled is stabilizable at the operating steady state, then, it has been shown that feasibility of the open loop quasi-infinite horizon control problem at time $t = 0$ implies nominal asymptotic stability of the closed loop system.

It may be noted that, the terminal region Ω_x and the penalty matrix $\mathbf{W}_{\infty}$ are functions of the model parameters. As a consequence, if the model parameters/unmeasured disturbances undergo abrupt and large changes during plant operation, the closed loop stability can no longer be guaranteed using $(\mathbf{W}_{\infty}, \Omega_x)$ computed initially. We propose a remedy to this problem whereby we recompute the terminal region Ω_x and the penalty matrix $\mathbf{W}_{\infty}$ every time the FDI component diagnoses an abrupt change in model parameters/unmeasured disturbance. Under the ideal condition where a fault is correctly isolated and its magnitude is accurately estimated, this pro-active measure can ensure nominal closed loop stability under certainty equivalence control.

It may be noted that Chen and Allgower [1] assume exact knowledge of the complete state vector in their formulation. In our formulation, on the other hand, we make use of the state estimate $\hat{\mathbf{x}}(k|k)$ to initialize prediction in the NMPC formulation. The stability of NMPC and EKF pair is still an open issue as the separation principle does not hold in the nonlinear case.

4. Simulation case studies

Simulation studies are carried out to evaluate the proposed intelligent state estimation (referred to as *intelligent EKF* in the rest of the text) and FTNMPC schemes by simulating control problems associated with the following highly nonlinear systems:

- CSTR exhibiting input multiplicity [20,21].
- Unstable nonlinear system described in Chen and Allgower [1].
- Fed-batch bioreactor [22].

The performance of the *conventional* NMPC (CNMPC) that employs *conventional EKF* for state estimation is compared with the performance of the proposed FTNMPC scheme under different fault scenarios. In all the three case studies the CNMPC formulation is based on the nominal model given by Eqs. (1)–(3) and state estimator given by Eqs. (6)–(9). The future trajectory predictions in CNMPC formulation are carried out using Eq. (52). In addition, a model-plant mismatch compensation scheme similar to (77)–(79) has been used in CNMPC formulations used in the CSTR and fed-batch bioreactor case studies for eliminating offset.

4.1. CSTR with input multiplicity

The system under consideration consists of a CSTR in which a reversible exothermic reaction of type $A \rightleftharpoons B$ is carried out. The nominal parameters and the optimum operating steady state used in the simulation studies can be found in Li and Biegler [21] and Economou [23].

The dynamic model used for simulating the CSTR system is as follows [20]:

$$\frac{dC_a}{dt} = \frac{F}{hA_c}(C_{ai} - C_a) - K_1 C_a + K_2 C_b \tag{82}$$

$$\frac{dC_b}{dt} = -\frac{F_i}{hA_c} C_b + K_1 C_a - K_2 C_b \tag{83}$$

$$\frac{dT}{dt} = \frac{1}{hA_c} F_i (T_i - T) + \frac{-H_r}{\rho C_p}(K_1 C_a - K_2 C_b) \tag{84}$$

$$\frac{dh}{dt} = \frac{1}{A_c}(F_i - k\sqrt{h}) \tag{85}$$

$$K_1 = k_f \exp(-E_f/T); \quad K_2 = k_b \exp(-E_b/T) \tag{86}$$

In the present work, the concentration of component $B(\mathbf{y}_1 = C_b)$ and level $(\mathbf{y}_2 = h)$ in the CSTR are treated as two controlled outputs of the system. The inlet flow rate $(\mathbf{u}_1 = F_i)$ and inlet feed temperature $(\mathbf{u}_2 = T_i)$ are used as manipulated variables. The constraints imposed on manipulated inputs are as follows:

$$0 \leqslant F_i \leqslant 2 \quad \text{and} \quad 300 \leqslant T_i \leqslant 550$$

The inlet concentration (C_{ai}) is treated as unmeasured disturbance and it is assumed to be corrupted with a zero mean white noise signal of standard deviation 0.05 mol/m^3. The sampling interval is chosen as 0.4 min. This system exhibits input multiplicity and change in the sign of the steady state gain in the operating region. For a fixed value of flow rate, the concentration (C_b) as a function of inlet flow temperature has a well defined maximum. Thus, the objective is to control the concentration (C_b) at the optimum operating point of the system where the conversion is maximum. Regulating the CSTR at the optimum point is a challenging task as the steady state gain reduces to zero at the peak and changes its sign across the peak [20,21,23].

In this case study, we hypothesize ten different faults consisting of single faults such as (a) biases in two actuators, (b) biases and failures of two sensors, (c) step jump in inlet concentration (C_{ai}), (d) step change in reaction rate parameter and simultaneous faults as (e) simultaneous occurrence of step changes in C_{ai} and k_f (e) simultaneous occurrence of bias in level sensor and inlet concentration (C_{ai}). The controlled outputs are concentration, C_b, and reactor level. The tuning parameters used in the controller formulation and SNL-GLR method are given in Tables 1 and 2 respectively.

4.1.1. Optimum seeking control in presence of parametric faults

In this sub-section, it is assumed that measured outputs are same as controlled outputs, i.e.

$$\mathbf{G}[\mathbf{x}] = \mathbf{H}[\mathbf{x}] = [0 \quad 1 \quad 0 \quad 1]\mathbf{x}$$

and measurements of C_b and h are assumed to be corrupted with a zero mean white noise signals with standard deviations 0.005 mol/m^3 and 0.002 m, respectively. The control problem is to regulate the system at the optimum operating point in the face of abrupt changes in parameters and unmeasured disturbances. It may be noted that the location of the maximum conversion point is a function of model parameters and unmeasured disturbances. Under nominal operating conditions, the optimum operating point is located at $C_b = 0.5088$ mol/m^3 for $h = 0.16$ m. However, when there is a significant shift in the mean value of model parameters or unmeasured disturbances, the maximum concentration of C_b predicted by the nominal model is different than the maximum achievable output in the plant. Patwardhan and Madhavan [24] have discussed two possible situations arising due to shift in the optimum point: (a) sub-optimal operation when the maximum attainable conversion in the plant shifts above the nominal maximum; (b) unattainable setpoint when the maximum shifts below the nominal maximum. The later situation results in a steady state offset and may lead to input saturation and loss of control. In this section, we demonstrate that the proposed FTNMPC formulation, in combination with on-line steady state optimization, can be used to track the changing optimum operating point.

To begin with, we demonstrate performance of our scheme when two faults occur sequentially. Initially, the process is controlled at the nominal operating point. At $t = 26.4$ min, the reaction rate parameter k_f is changed from 1 to 1.3. This increases the maximum attainable concentration form $C_b = 0.5088$ mol/m^3 to $C_b = 0.5738$ mol/m^3. The proposed SNL-GLR method correctly isolates this fault and magnitude estimated is $\hat{k}_f = 1.2263$ which is further refined using NL-GLR to $\hat{k}_f = 1.2945$. The optimum concentration operating point computed based on the refined estimate of k_f is $C_b = 0.5728$ mol/m^3. Thus, the concentration setpoint is changed to $C_b = 0.5728$ mol/m^3 subsequent to fault diagnosis. Subsequent to this fault a step jump of 0.2 is given in inlet concentration (C_{ai}) at $k = 306$. This increases the maximum attainable concentration from $C_b = 0.5738$ mol/m^3 to $C_b = 0.6886$ mol/m^3. The proposed SNL-GLR method correctly isolates this fault and magnitude estimated after refinement is 0.2018. The optimum concentration operating point computed based on this estimate

Table 1
CSTR example: controller tuning parameters

Prediction horizon	12
Control horizon	3
Error weighting matrix	$\begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$
Set-point	[0.5088 0.16]

Table 2
CSTR example: SNL-GLR tuning parameters

Window for fault confirmation	60
Level of significance for fault detection	0.5
Level of significance for fault confirmation	0.01

is $C_b = 0.6883$ mol/m^3. Thus, the concentration setpoint is changed to $C_b = 0.6883$ mol/m^3 subsequent to fault diagnosis. The CNMPC, however, is unaware of the nature and type of unmeasured disturbance and attempts to achieve the original setpoint of $C_b = 0.5088$ mol/m^3. This results in suboptimal operation of the CSTR. Fig. 2 compares the performances of both the controllers in the presence of multiple sequential faults. A better insight into their behavior is obtained when we compare the state estimation errors generated by conventional EKF and the proposed intelligent EKF (see Fig. 3). It can be seen from Fig. 3 that both the conventional EKF and Intelligent EKF generate biased estimates of C_a and C_b immediately after the fault occurs. However, as soon as the fault is correctly diagnosed and compensated, the states estimated using intelligent EKF are close to their true values and bias in state estimation is eliminated. On the other hand, the bias in the estimates of C_b persists in case of conventional EKF even when C_b is directly measured. This can be attributed to persistent plant model mismatch that develops subsequent to occurrence of abrupt changes in parameters.

In another simulation run, to evaluate the performance of the proposed state estimation and control scheme when multiple faults occur simultaneously, we introduce a bias of $(-0.02$ m) in level sensor and step jump of magnitude +0.1 in the inlet concentration (C_{ai}) at $t = 26.4$ min. The proposed SNL-GLR method correctly isolates this simultaneous fault and refined fault magnitudes estimated are -0.0196 m and 0.0955 mol/m^3. The optimum concentration operating point computed based on these estimates is $C_b = 0.5574$ mol/m^3 (true optimum point under the changed conditions is $C_b = 0.5597$ mol/m^3). Thus, the concentration setpoint is changed to $C_b = 0.5574$ mol/m^3 subsequent to fault diagnosis and FTNMPC shifts the average steady state concentration to 0.552. The CNMPC that employs conventional EKF for state estimation, however, attempts to reject these abrupt changes as an unmeasured disturbance to achieve the original setpoint of $C_b = 0.5088$ mol/m^3, which results in suboptimal operation. In addition, the biased level sensor gives rise to an offset in true value of reactor level and the setpoint in the case of CNMPC. Figs. 4 and 5 compare the performances of both the controllers and state estimators, respectively. Similar to the sequential fault case, the states estimated using intelligent EKF move close to their true values and bias in state estimation is eliminated soon after the fault is correctly diagnosed and compensated.

4.1.2. Estimator reconfiguration on sensor failure

In this subsection, we assume that the reactor temperature measurements are also available together with measurements of C_b and h, i.e.

$$\mathbf{H}[\mathbf{x}] = [0 \quad 1 \quad 1 \quad 1]\mathbf{x} \tag{87}$$

while the controlled outputs are

$$\mathbf{G}[\mathbf{x}] = [0 \quad 1 \quad 0 \quad 1]\mathbf{x}$$

i.e. C_b and h. The temperature measurements are assumed to be corrupted with a zero mean white noise signal with standard deviation 0.02 °C. We begin simulations under the scenario that all three sensors are functioning well with reactor operating at a suboptimal operating point $[0.4088 \quad 0.16]$. At sampling instant $k = 71$ ($t = 28.4$ min) setpoint is changed to a new value $[0.5088 \quad 0.16]$ and just prior to this at $k = 66$ ($t = 26.4$ min), a failure in the sensor for C_b is simulated by holding sensor output constant at subsequent time instants. Fig. 6 shows the state estimation errors for the proposed intelligent EKF before and after fault compensation. As can be observed from this figure, the state estimates become biased when concentration sensor fails. However, in the Intelligent EKF scheme, after the failure has been identified by the SNL-GLR method, the state estimator is reconfigured using only temperature and level measurements, i.e. by setting

$$\mathbf{H}[\mathbf{x}] = [0 \quad 0 \quad 1 \quad 1]\mathbf{x} \tag{88}$$

This measure eliminates the bias in the estimate of C_b and enables FTNMPC to track this setpoint change using correctly estimated concentration C_b (see Fig. 7) from the available level and temperature measurements. Before the failure is isolated and compensated, the controller attempts to increase reactor concentration by increasing the throughput and thereby increasing the reactor level. However, FTNMPC is able to recover the level to the desired setpoint subsequent to the fault accommodation.

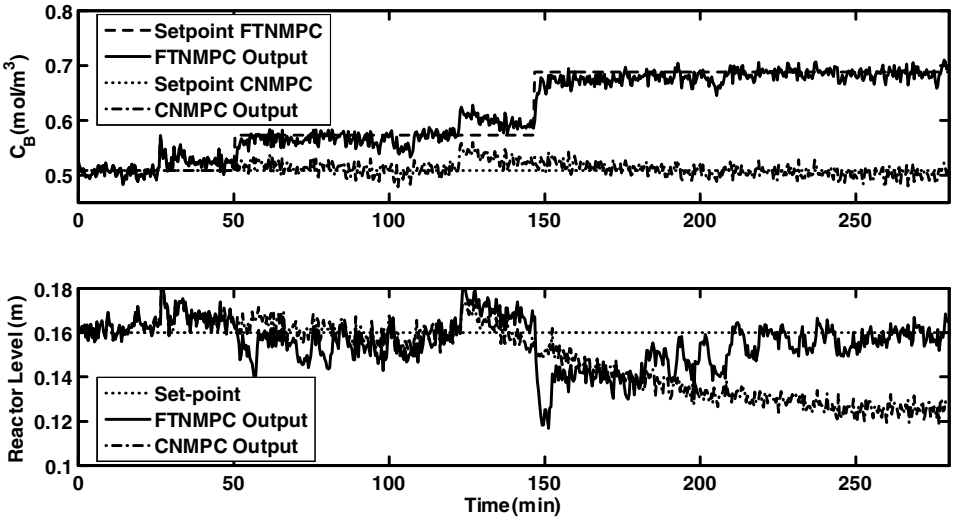

Fig. 2. CSTR example: comparison of controlled outputs of FTNMPC and CNMPC under multiple sequential faults.