



# Fault detection and identification using Bayesian recurrent neural networks

Weike Sun<sup>a</sup>, Antonio R.C. Paiva<sup>b,\*</sup>, Peng Xu<sup>b</sup>, Anantha Sundaram<sup>b</sup>, Richard D. Braatz<sup>a</sup>

<sup>a</sup> Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>b</sup> Corporate Strategic Research, ExxonMobil Research and Engineering, Annandale, NJ, USA

## ARTICLE INFO

### Article history:

Received 8 November 2019

Revised 12 May 2020

Accepted 24 June 2020

Available online 29 June 2020

### Keywords:

Fault detection

Fault identification

Recurrent neural networks

Variational dropout

Bayesian inference

Tennessee eastman process

## ABSTRACT

In the processing and manufacturing industries, there has been a large push to produce higher quality products and ensure maximum efficiency of processes, which requires approaches to effectively detect and resolve disturbances to ensure optimal operations. While many types of disturbances can be compensated by a control system, it cannot handle some large process disruptions. As such, it is important to develop monitoring systems to effectively detect and identify those faults such that they can be quickly resolved by operators. This article proposes a novel probabilistic fault detection and identification method which adopts a newly developed deep learning approach using Bayesian recurrent neural networks (BRNNs) with variational dropout. The BRNN model is general and can model complex nonlinear dynamics. Moreover, compared to traditional statistic-based data-driven fault detection and identification methods, the proposed BRNN-based method yields uncertainty estimates which allow for simultaneous fault detection of chemical processes, direct fault identification, and fault propagation analysis. The performance of the method is demonstrated and contrasted to (dynamic) principal component analysis, which is widely applied in the industry, in the benchmark Tennessee Eastman process (TEP) and a real chemical manufacturing dataset.

© 2020 Elsevier Ltd. All rights reserved.

## 1. Introduction

In industrial manufacturing processes, a *fault* is defined as any abnormal deviation from the normal operating conditions (NOC). Faults are a concern because even small faults in a complex industrial system can initiate a series of events that result in loss of efficiency and reliability. As a result, there is a need for techniques to improve the process's reliability and up-time. Effective fault detection and identification (FDI) is important for monitoring components for making appropriate maintenance decisions. First, fault detection determines whether a fault has occurred in the system (also characterized as anomaly detection in other applications). Then fault identification determines which observation variables are most relevant to diagnosing the fault detected, thereby helping operators to focus on specific subsystems. Systems that can accurately and promptly detect and identify faults can more effectively inform operators and engineers and significantly reduce the effort and time to recover the system.

A number of FDI methods have been proposed in the literature. Since analytical and knowledge-based methods are impractical in

most large-scale modern industrial processes, data-driven methods have dominated the literature for the past decade and have been effective in practice, taking advantage of increasing levels of instrumentation and widespread availability of sensor data (Qin, 2009; Ge et al., 2013; Yin et al., 2014; Chiang et al., 2000a). The choice of model used to characterize the NOC and deviations thereof is still a crucial aspect in these methods because the limitations of the model lead to decreased detection rates or increased occurrence of fault alarms.

A number of data-driven methods including principal component analysis (PCA) (Jolliffe, 2011), partial least squares (Kourtis and MacGregor, 1996), Fisher discriminant analysis (Fisher, 1936; Chiang et al., 2000a), and support vector machines (Chiang et al., 2004), have been applied for fault detection and identification in industry with varying degrees of success. The most widely used method is PCA which models the correlations between the process variables. PCA can detect faults effectively when the sensor measurements are highly correlated, which is often the case. For many processes, the temporal dynamics also need to be taken into consideration, especially when fast sampling rates are used, because the dynamics provide additional information through which to detect deviations from the NOC. To that end, DPCA has been proposed to handle serially correlated multivariate observations

\* Corresponding author.

E-mail address: [antonio.paiva@exxonmobil.com](mailto:antonio.paiva@exxonmobil.com) (A.R.C. Paiva).

(Ku et al., 1995). DPCA can be viewed as a multivariate autoregressive model with exogenous inputs (ARX). PCA and DPCA are both limited by the linear model structure and correlations in the process' dynamics. Methods for extending PCA to nonlinearities such as kernel PCA and neural network-based PCA (Lee et al., 2004; Hsieh, 2006) have been well studied only for static systems. As such, the development of approaches that can effectively model nonlinear system structure and dynamics has been an active research field.

Neural network (NN) based methods have also received significant attention because of their capability and flexibility for modeling complex structure and temporal dynamics. NN models have been used for fault detection in three general frameworks: (1) as a fault classification tool between normal and known faulty conditions (Zarei et al., 2014; Chine et al., 2016; Ince et al., 2016; Jia et al., 2016; Wu and Zhao, 2018; Hu and Jiang, 2019; Lee et al., 2017; Li et al., 2014; Zhang and Zhao, 2017), (2) as a model of the input-output variable relationships during NOC (Malhotra et al., 2015; Patan et al., 2008; Moustapha and Selmic, 2007; Talebi et al., 2009; Wang et al., 2017; Nie et al., 2018), or (3) as a generalization of the basic fault detection methods such as NN-based PCA (Kramer, 1992; Dong and McAvoy, 1996) using statistical indices to monitor the process. The first two approaches are dominant for NN-based fault detection. The first approach can be highly effective due to the up-front knowledge of specific fault conditions to detect. It can also be set up to classify each fault which directly enables fault diagnosis. On the other hand, training these NNs requires substantial amounts of data under fault conditions but these data are usually quite limited for chemical manufacturing processes compared to NOC data. Moreover, it is hard to assess the performance of these methods for fault conditions other than for which the classifier is explicitly trained for. In the second approach, NNs are used to model the process by capturing the nonlinear, multivariate, and temporal dependencies from inputs to outputs. In this approach, the NN models are typically trained on NOC data to predict the system outputs. This NN is then used for fault detection during runtime by comparing the predictions of the NN with the actual system output measurements, and a fault is detected if the difference is significantly large. This approach has the advantage that the NNs are trained using only NOC data, which is usually abundant, and that the NNs are not constrained by the type of fault because detection is marked from any significant deviation from the NOC. On the other hand, the model must accurately characterize these complex and potentially nonlinear structures between inputs and outputs in the process, or its fault detection will perform poorly as a result. Moreover, only faults that break the input-output relationships are considered, meaning that faults due to input disturbances will likely not be detected. In the third approach, NNs are used to account for nonlinearities in the process but, like other PCA-based methods, fail to appropriately model the process dynamics.

There are other challenges regarding both approaches, which have limited the application of NNs in industrial process monitoring. First, fault identification has not yet been properly addressed. Once a fault is detected, it is typically difficult to identify the input variables most relevant to the fault from a complex NN model. Even if an NN is trained to directly classify the fault, the underlying cause may still be unclear if there are multiple explanations for the observed fault type. Secondly, standard NNs are deterministic models which lack an estimate of the uncertainty in the model outputs. However, uncertainty and probabilistic estimates are important to assess the confidence level associated with the decision of detecting a fault and for fault identification. Lastly, NNs are prone to overfitting, meaning that they 'memorize' the particular characteristics of the training data that are not relevant for new data. This overfitting must be addressed to ensure good generalization

to the full space of operating conditions of a complex industrial process.

For fault identification, contribution plots (Miller et al., 1998) are one of the most popular techniques for providing information on the variables that are most strongly related to the faults. In the context of PCA-based methods, contribution plots are obtained by quantifying the contribution of each process variable to the individual scores of the PCA representation (Westerhuis et al., 2000). Methods based on the contribution of each process variable in the residual space have also been developed (Wise et al., 1989). However, the aforementioned limitations of PCA-based approaches will also be reflected in the identification procedure and those methods require extra processing steps after fault detection. Moreover, those methods only provide the relative contribution value of each variable which is not very useful. A more valuable and precise measure to aid operators in diagnosis would be the probable severity of each affected variable. On the other hand, it has been hard to extend contributions plots to NNs due to the complex and nonlinear relationships between predictions and model inputs.

This article proposes a novel end-to-end FDI framework, which adopts a recently developed Bayesian recurrent neural network (BRNN) architecture (Gal and Ghahramani, 2016b). The proposed FDI framework is fundamentally different from the two types of frameworks that have been previously used in the NN-based fault detection literature. The proposed framework uses BRNNs to model the joint distribution and dynamics between all process variables. This framework provides estimates of the prediction uncertainty, which capture both model uncertainty and the inherent noise in the data. The BRNN is realized using the variational dropout approach proposed in (Gal and Ghahramani, 2016a; 2016b) due to its simplicity, regularization capability, strong generalization ability, and scalability.

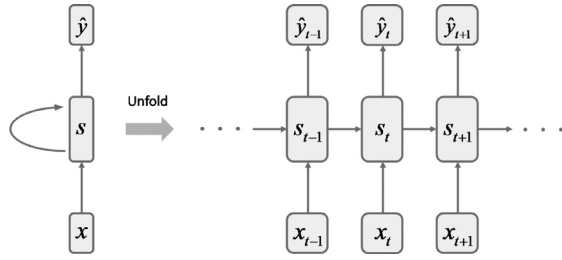
To the best of our knowledge, this work is the first time that Bayesian spatio-temporal models, and BRNNs in particular, have been successfully applied to FDI in the chemical manufacturing industry. The proposed approach tackles three key challenges typical of manufacturing systems: (1) nonlinearity, (2) non-Gaussian distributed variables, and (3) high degree of spatio-temporal correlations (i.e., temporal and sensor correlations). Furthermore, the probabilistic framework provided by BRNNs enables more sensitive and robust FDI. Fault identification through the proposed BRNN-based approach provides easily interpretable visualizations to the plant operators, for quick fault type categorization, analysis of the possible fault propagation path, and root cause determination using engineering judgment.

The remainder of this paper is organized as follows. Section 2 provides a brief introduction to RNNs and BRNNs, and describes the variational dropout approach used in this paper for inference in BRNNs. Section 3 presents the proposed BRNN-based FDI methodology. In Section 4, the effectiveness of the proposed approach is demonstrated and compared to (D)PCA-based methods in the Tennessee Eastman process and a real chemical manufacturing process, followed by the conclusion in Section 5.

## 2. Background

### 2.1. Recurrent neural networks

RNNs were developed in the 1980s (Rumelhart et al., 1986). Since then, RNNs have been shown to achieve state-of-the-art performance on a wide range of sequential data modeling tasks, including language modeling, speech recognition, image captioning, and music composition (Wu et al., 2016; Jozefowicz et al., 2016; Merity et al., 2016). Generally speaking, an RNN comprises an input layer, one or more hidden recurrent layers, and an output layer.



**Fig. 1.** A simple RNN structure with one recurrent layer and showing the unfolding in time of the sequence of its forward computation. The RNN includes the input variable  $\mathbf{x}_t$ , state variable  $\mathbf{s}_t$  and outputs  $\mathbf{y}_t$ . The state variable  $\mathbf{s}_t$  is calculated based on the previous state  $\mathbf{s}_{t-1}$  and the current input  $\mathbf{x}_t$ . The RNN output  $\mathbf{y}_t$  is then calculated based on the current state. In this way, the input sequence  $\mathbf{x}_t$  is mapped to output sequence  $\mathbf{y}_t$  with each  $\mathbf{y}_t$  depending on all previous inputs. The model parameters  $\omega = \{\mathbf{W}_s, \mathbf{U}_s, \mathbf{W}_y, \mathbf{b}_s, \mathbf{b}_y\}$  are shared at each time step.

The input layer corresponds directly to the input data, and hidden recurrent layers capture the state with the response of its nodes being added to the inputs on the next time step. At each time  $t$ , denote the input to the network as  $\mathbf{x}_t \in \mathbb{R}^{m_x}$ , the state (i.e., output of the hidden layer) as  $\mathbf{s}_t \in \mathbb{R}^{m_s}$ , and the RNN output as  $\mathbf{y}_t \in \mathbb{R}^{m_y}$ . They are represented as row vectors in the equations. Accordingly, the state and output layers have the general form

$$\begin{aligned} \mathbf{s}_t &= f_s(\mathbf{x}_t, \mathbf{s}_{t-1} | \theta_s) \\ \mathbf{y}_t &= \mathbf{W}_y \mathbf{s}_t + \mathbf{b}_y \end{aligned} \quad (1)$$

where the subscript  $s = 1, \dots, m_s$  is the index over hidden layer nodes,  $\theta_s$  and  $f_s$  denote the corresponding hidden layer parameter/weights and nonlinear operator for each node, and  $\mathbf{W}_y \in \mathbb{R}^{m_y \times m_s}$  and  $\mathbf{b}_y \in \mathbb{R}^{m_y}$  are the output layer parameters. The new state of the network depends on its value at the previous time step, emblematic of recurrent architectures. This dependency, and the unfolding through time, is depicted in Fig. 1. A linear output layer is commonly used for regression tasks.

In the simpler form of nodes, the state is computed as (Elman, 1990)

$$\mathbf{s}_t = \phi(\mathbf{W}_s \mathbf{x}_t + \mathbf{U}_s \mathbf{s}_{t-1} + \mathbf{b}_s) \quad (2)$$

where  $\mathbf{W}_s \in \mathbb{R}^{m_s \times m_x}$ ,  $\mathbf{U}_s \in \mathbb{R}^{m_s \times m_s}$ , and  $\mathbf{b}_s \in \mathbb{R}^{m_s}$  are the hidden layer parameters (denoted  $\theta$  above), and  $\phi$  is an element-wise activation function such as the logistic, hyperbolic tangent, or a rectifier linear function.

As can be explicitly observed from the mathematical formulation in Eq. 1, RNNs are essentially state-space models capable of modeling nonlinear dependencies. RNNs can capture complex nonlinear dynamics of a system in the state. Also, by appropriately training the parameters, RNNs can adapt to the right level of temporal depth. Thus, RNN models are considerably more powerful for modeling complex industrial processes in comparison to traditional statistical methods.

It is worth noting that different RNN architectures have been proposed (Jordan, 1997), with the formulation in Eq. 1 corresponding to Elman's architecture (Elman, 1990), which has been widely used in the recent deep learning RNN implementations and applications.

In order to learn the parameters of the RNN, an optimization problem is defined with regard to an appropriate loss function. For regression tasks, the loss function is typically chosen to be the mean squared loss,

$$J(\Theta) = \frac{1}{N} \sum_{t=1}^N \|\mathbf{y}_t - \hat{\mathbf{y}}_t\|_2^2, \quad (3)$$

or the cross-entropy loss for classification purposes,

$$J(\Theta) = - \sum_{t=1}^N \mathbf{y}_t \log \hat{\mathbf{y}}_t \quad (4)$$

where  $\Theta$  denotes the collection of all RNN model parameters, and  $\mathbf{y}_t$  is the desired output at time step  $t$ . In addition,  $L_2$  regularization terms are often added to help prevent overfitting, resulting in the overall minimization objective

$$L(\Theta) = J(\Theta) + \lambda (\|\mathbf{W}_s\|^2 + \|\mathbf{W}_y\|^2 + \|\mathbf{U}_s\|^2) \quad (5)$$

where  $\lambda$  is the regularization (aka weight decay) parameter.

Because the recurrence introduces dependencies between time steps, training RNNs involves backpropagation through time (BPTT) to compute the gradient update of the model parameters that minimizes the loss function (Werbos, 1974; 1990). BPTT corresponds to an unfolding of the network over a number of time steps, as depicted in Fig. 1. For BPTT, the difference between network outputs and target values is first calculated and stored for each time step in a forward pass, and then the weight gradient updates are calculated as the network is "rolled back". However, simple RNNs trained with BPTT can have difficulties learning long-range time dependencies due to the vanishing gradient problem (Bengio et al., 1994). To alleviate the vanishing gradient problem, recurrent node gating mechanisms have been recently developed. These gating mechanisms allow information and the gradients to flow through the unrolled network with minimal attenuation if determined to be necessary by BPTT. These gating mechanisms resulted in two popular variations on RNN hidden units: LSTM units (Hochreiter and Schmidhuber, 1997) and GRUs (Cho et al., 2014). RNNs with LSTM and GRU units have been reported to show salient performance (Graves et al., 2013; Cakir et al., 2015).

## 2.2. BRNNs

BRNNs combine statistical modeling of RNN parameters to obtain a probabilistic model of input-output mapping. As such, instead of point estimates, BRNNs can effectively perform Bayesian inference which provides probabilistic distributions over the outputs.

To realize that capability, BRNNs view the model parameters  $\omega = \{\mathbf{W}_s, \mathbf{W}_y, \mathbf{U}_s, \mathbf{b}_s, \mathbf{b}_y\}$  as random variables from a prior distribution  $p(\omega)$ . Expressing the functional dependence in Eq. 1 as  $\mathbf{s}_t = \mathbf{f}_s^\omega(\mathbf{x}_t, \mathbf{s}_{t-1})$  and  $\mathbf{y}_t = \mathbf{f}_y^\omega(\mathbf{s}_t)$ , the likelihood of the output for each data point is

$$p(\mathbf{y}_t | \omega, \mathbf{x}_t, \mathbf{s}_t, \tau) = N(\mathbf{y}_t | \mathbf{f}_y^\omega(\mathbf{f}_s^\omega(\mathbf{x}_t, \mathbf{s}_{t-1}), \tau^{-1} \mathbf{I}_D)) \quad (6)$$

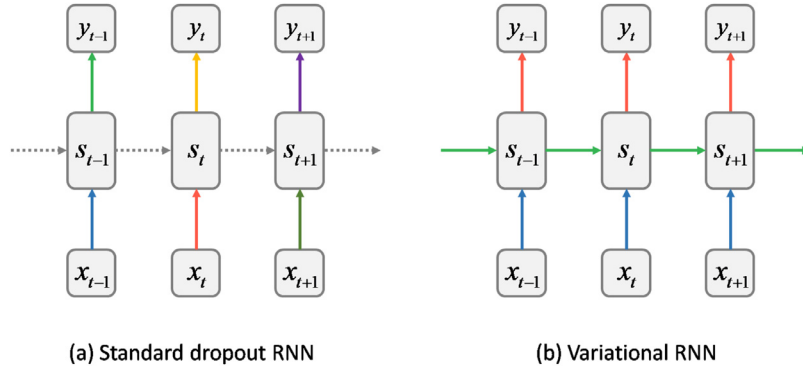
where  $\tau$  is the precision parameter that reflects the intrinsic noise in the data, and the likelihood function is assumed to have a normal distribution for simplicity. Note how the likelihood function is evaluated with respect to forward passes through the NN.

Then, given a training dataset comprising  $\mathbf{X}$  and  $\mathbf{Y}$ , learning entails estimating the posterior distribution  $p(\omega | \mathbf{X}, \mathbf{Y})$  over the space of parameters. With the updated distribution, the distribution of a predicted output  $\mathbf{y}^*$  can be obtained by integration

$$p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^* | \mathbf{x}^*, \omega) p(\omega | \mathbf{X}, \mathbf{Y}) d\omega \quad (7)$$

where the dependency on the precision parameter, state, and past inputs are not shown to simplify the expression. For the prior distribution, standard zero-mean Gaussian priors over the weight matrices  $p(\mathbf{W})$  and  $p(\mathbf{U})$  are typically chosen, with point estimates for the bias vectors assumed for simplicity. The uncertainty in the prediction will be directly reflected in the posterior distribution  $p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{X}, \mathbf{Y})$ .

In complex models such as NNs, however, the exact inference of the posterior is not possible. Moreover, traditional algorithms



**Fig. 2.** Illustration of the variational dropout technique (right) compared to standard dropout technique (left) for a simple RNN. Each graph shows units unfolded over time, with the lower level for inputs, middle level for state units, and upper level for output units. Vertical arrows represent the connections from inputs to outputs while horizontal arrows represent recurrent connections. The arrows with dashed grey lines represent the standard connection without dropout. Colored lines represent dropout connections with different colors for different dropout masks. (Left) In the standard dropout technique, no dropout is applied for the recurrent layers, while other connections have different dropout masks at different time steps. (Right) For the variational dropout approach proposed in (Gal and Ghahramani, 2016b), dropout is applied to both input, recurrent, and output layers with the same dropout mask at different time steps. Variational dropout is applied during both training and testing.

for approximating the Bayesian inference are generally not applicable to train RNNs having a large number of parameters or complex architectures. To overcome this limitation, several approximation inference methods have been proposed, including variational dropout (Gal and Ghahramani, 2016b; 2016a), Bayes by Back-Prop (Pawlowski et al., 2017; Fortunato et al., 2017), multiplicative normalizing flows (Louizos and Welling, 2017), and probabilistic backpropagation (Hernández-Lobato and Adams, 2015). Among all those techniques, the variational dropout technique proposed by Gal and Ghahramani (2016b) is adopted in this paper because of its simplicity and generalization capability. In particular, variational dropout can be applied to any RNN architecture without modification on the underlying NN structure, and only concurrent runs of the trained model are needed for online application. Details of this algorithm are reviewed in the next section.

### 2.3. Variational dropout as Bayesian approximation

Gal and Ghahramani (2016a) showed how dropout could be used as a general variational approximation to the posterior of Bayesian neural networks (BNNs), which can be applied directly to a variety of NN architectures. The main advantage of this ‘variational dropout’ approach is that it does not require significant modifications to the model architecture and training method, unlike other probabilistic approximation methods. Moreover, the uncertainty estimation incurs only the computation cost due to multiple stochastic forward passes through the network to generate samples of the posterior distribution.

Therefore, variational dropout is used here as a variational inference approach for BNNs. Variational inference is a technique used to approximate an intractable posterior distribution  $p(\omega|\mathbf{X}, \mathbf{Y})$  with a simpler parameterized distribution  $q(\omega)$ . Then, the integration in Eq. 7 can be approximated simply by MC integration using  $q(\omega)$ . Specifically, the approximation distribution is factorized over the weight matrices in  $\omega$ . For each row  $\mathbf{w}_k$ , variational dropout imposes a variation distribution comprising a mixture of two Gaussian distributions with small variances,

$$q(\mathbf{w}_k) = pN(\mathbf{w}_k|0, \sigma^2\mathbf{I}) + (1 - p)N(\mathbf{w}_k|\mathbf{m}_k, \sigma^2\mathbf{I}), \quad (8)$$

where  $p$  is the predefined dropout probability,  $\sigma^2$  is a small precision parameter, and  $\mathbf{m}_k$  is a variational parameter. The learning problem is then casted into an optimization problem by minimizing the KLD between  $q(\omega)$  and  $p(\omega|\mathbf{X}, \mathbf{Y})$ . It can be shown that optimizing the loss function using dropout is equivalent to minimizing  $\text{KL}(q(\omega)||p(\omega|\mathbf{X}, \mathbf{Y}))$  (Gal and Ghahramani, 2016a), which updates

the variational parameter. Although variational inference is a biased approximation, it has been shown to work well in practice.

Variational dropout requires caution when applied in the context of RNNs, however. Because of the recurrence, naïvely applying standard dropout (Srivastava et al., 2014) with different masks at each time step of an RNN can lead to model instabilities and disrupt an RNN’s capability to model a sequence (Pham et al., 2014; Pachitariu and Sahani, 2013). We use the approach in (Gal and Ghahramani, 2016b) to resolve these issues. Under these circumstances, variational dropout has been shown to also act as an effective regularization method for reducing overfitting by preventing co-adaptations in RNNs (Gal and Ghahramani, 2016b).

The implementation of BRNNs with variational dropout is relatively straightforward. During both training and testing, the variational approximation involves sampling the model distribution with regard to the variational distribution over the weights, which is implemented by dropping out (i.e., forcing to zero) randomly chosen inputs, outputs, and hidden states. This step results in multiple random realizations of the RNN model, each obtained by implicitly removing a portion of the inputs, outputs, or hidden states. However, as detailed in (Gal and Ghahramani, 2016b), it is crucial for RNNs that the dropout mask used for each model realization be kept fixed between time steps. In other words, the dropout mask of which elements are zeroed out is sampled and frozen for each time sequence sample. This sampling characteristic is contrasted to standard dropout in Fig. 2.

Variational dropout applied during testing can be viewed as an approximation to MC samples from the posterior predictive distribution,  $p(\omega|\mathbf{X}, \mathbf{Y})$ . Given a new observation  $\mathbf{x}^*$ , by forward passing it  $N$  times,  $N$  samples  $\{\hat{\mathbf{y}}^*(i)\}_{i=1, \dots, N}$  are collected of the approximate predictive posterior. The corresponding empirical estimators for the posterior predictive mean, standard deviation, and covariance are

$$E(\hat{\mathbf{y}}^*) \approx \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{y}}^*(i) \quad (9)$$

$$\text{std}(\hat{\mathbf{y}}^*) \approx \sqrt{\tau^{-1} + \frac{1}{N} \sum_{i=1}^N (\hat{\mathbf{y}}^*(i))^2 - E(\hat{\mathbf{y}}^*)^2} \quad (10)$$

$$\text{cov}(\hat{\mathbf{y}}^*) \approx \tau^{-1}\mathbf{I}_D + \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{y}}^*(i) \hat{\mathbf{y}}^{*\top}(i) - E(\hat{\mathbf{y}}^*)^\top E(\hat{\mathbf{y}}^*) \quad (11)$$



where  $\tau$  can be estimated as  $\tau = \frac{p_l^2}{2N\lambda}$  given a predefined regularization/weight-decay parameter  $\lambda$ , and prior length scale  $l$  (Gal and Ghahramani, 2016a). Higher order statistics can also be estimated using the samples by moment-matching.

Since the forward passes involve simply a number of independent and fixed realizations of the RNN model distribution, they can be done concurrently, thus making variational dropout a good candidate for online monitoring. In the next section, the proposed novel FDI scheme is explained in detail. While this methodology is described here in the context of chemical process monitoring, it can be observed that it could be readily extended to other manufacturing industries.

### 3. Methodology for FDI

The design of a FDI system generally begins with the development of a model to characterize the normal operating characteristics of a process. Historical data collected during the NOC are used to build the model, which means this learning problem is unsupervised. Then, an approach must be established to characterize the magnitude of the deviation from the NOC based on the developed model and to determine when deviations are considered to be outside of the NOC. For example, the  $T^2$  and  $Q$  statistics are commonly used to measure the distance of the observation to the NOC region in PCA-based models and thresholds thereon (Chiang et al., 2000a). Finally, given a new observation  $\mathbf{x}^*$ , these measures are calculated to determine whether  $\mathbf{x}^*$  deviates substantially from the NOC (fault detection) and, if that is the case, which variables are significantly affected (fault identification), thereby assisting in locating and troubleshooting the fault.

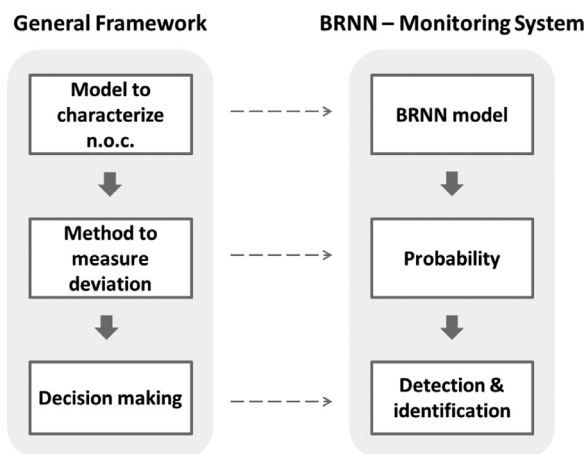
Specifically, this paper proposes using a BRNN with variational dropout to build the probabilistic model, denoted as  $f^\omega(\cdot)$ , and characterize the NOC and its intrinsic variability. As discussed earlier, BRNNs are capable of extracting the nonlinear spatial and temporal signatures in the data that are critical for characterizing complex chemical processes. Moreover, BRNNs provide probabilistic estimates of the likelihood of the observations with regard to its inferred posterior distribution of the variable values. These likelihood estimates lend themselves to be used to assess the current deviation level from the NOC region. Accordingly, observations are detected as faults whenever their deviation is above a threshold, determined such that the number of false alarms under the NOC does not exceed a predefined level. Fault identification then involves determining which process attributes are deviating significantly. This general framework for BRNN-based FDI is summarized in Fig. 3 and described in detail in the next sections.

#### 3.1. Fault detection

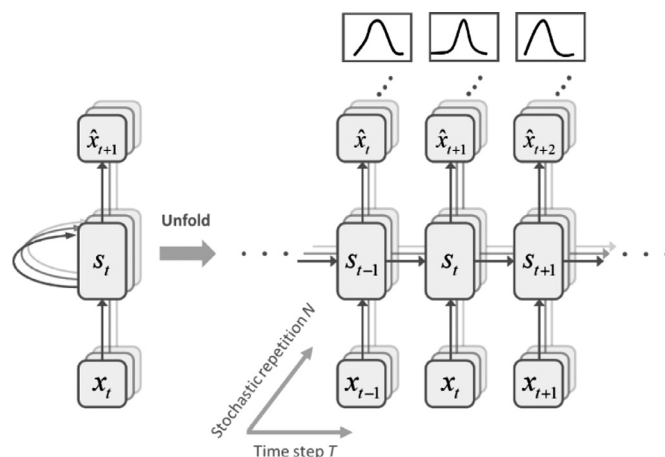
The first step toward fault detection is to learn a model to characterizing the NOC. This step involves training a BRNN with variational dropout to model the dynamics in time, correlations between sensors, and the prediction uncertainty resulting from model mismatch and inherent system variability/noise.

The BRNN model is trained directly on historical NOC data. Specifically, this step involves setting a training problem wherein the BRNN model uses the past context (as captured by its state) and current observation to predict the next observation, as depicted in Fig. 4. During training, BPTT is applied to batches of time subsequences with one variational dropout mask sampled per sequence, as explained in Section 2.3.

After training, the model output  $\hat{\mathbf{x}}_{t+1}$  from the BRNN is sampled from the posterior predictive distribution for next observation  $\mathbf{x}_{t+1}$  via variational dropout model realizations. That is, at each time step  $t$ , the stochastic forward pass is repeated  $N$  times, each with a different dropout mask, and the predictive distribution of



**Fig. 3.** General procedure for process monitoring system development (left) versus procedure for developing BRNN-based FDI system (right). The general framework to establish a monitoring system begins with a model to characterize NOC behavior, such as using the BRNN model to learn the NOC pattern from the training data. Then, the method to measure the deviation of a particular observation to the NOC region is chosen. In our case, the process observations are compared to the BRNN posterior predictive distributions. Finally, the decision will involve determining whether the acquired observation is from the NOC or not (i.e., compare deviations of observations for fault detection and assess which variables significantly deviate from the NOC for identification).



**Fig. 4.** Depiction of BRNN model using variational dropout (left) for FDI. The BRNN model uses the current observation and state to predict the next system observation. The BRNN model is unrolled in two dimensions (right): the time of the computation involved in its forward computation and the stochastic repetition by variational dropout. At each time step, stochastic variational dropout is applied  $N$  times and the corresponding MC prediction samples  $\{\hat{\mathbf{x}}_t(i)\}_{i=1,\dots,N}$  are used to approximate the posterior predictive distribution for that time step. For the next time step, the same procedure is repeated and MC samples  $\{\hat{\mathbf{x}}_{t+1}(i)\}_{i=1,\dots,N}$  are collected and used to approximate the distribution.

the output for  $t + 1$  is approximated based on the MC samples of the BRNN model,  $\{\hat{\mathbf{x}}_{t+1}(i)\}_{i=1,\dots,N}$ . Then, when the true observation  $\mathbf{x}_{t+1}$  is available, it is compared to the predictive distribution and deemed as abnormal if it significantly deviates from the predictive distribution. Finally, the true observation is fed into the BRNN model and the procedure is repeated for the next time step.

Notice that the predictive distribution is evolving, which provides an adaptive decision boundary for the next measurement. This adaptive decision boundary is calculated based on all the useful past system information, which takes into consideration both the spatial and temporal correlations in the data. Further combined with the potential ability to model nonlinear correlations,

this property increases both the detection sensitivity and robustness because of the increasing accuracy in modeling NOC pattern.

Depending on the complexity of the system and observed properties of the predictive distribution from the BRNN model, below is a description of two methodologies to quantify the deviation magnitude of each observation to its corresponding predictive distribution. The first method is faster and simpler to implement, but is limited to Gaussian predictive posterior distributions. The second method approximates the posterior distribution non-parametrically and is much more flexible, but requires tuning an additional density estimation parameter.

### 3.1.1. Method 1: Squared Mahalanobis distance for Gaussian predictive distributions

If the predictive distribution is Gaussian, or well approximated as such, the squared Mahalanobis distance can be used to characterize the magnitude of the deviation. First, the MC samples at time  $t$  of the predictive distribution,  $\{\hat{\mathbf{x}}_t(i)\}_{i=1,\dots,N}$ , are used to approximate the sample mean  $\boldsymbol{\mu}_t$  and covariance  $\mathbf{S}_t$ :

$$\boldsymbol{\mu}_t \approx \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{x}}_t(i) \quad (12)$$

$$\mathbf{S}_t \approx \tau^{-1} \mathbf{I}_D + \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{x}}_t(i)^\top \hat{\mathbf{x}}_t(i) - \boldsymbol{\mu}_t^\top \boldsymbol{\mu}_t. \quad (13)$$

Then, when the true observation  $\mathbf{x}_t$  is available, the squared Mahalanobis distance is calculated as

$$M^2 = (\mathbf{x}_t - \boldsymbol{\mu}_t)^\top \mathbf{S}_t^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_t). \quad (14)$$

A larger value of  $M^2$  indicates that the observation  $\mathbf{x}_t$  is far away from the predicted mean and there is a higher likelihood that it corresponds to a fault. The detection threshold  $M_{th}^2$  is determined with regard to a chosen maximum FAR  $\alpha$  on a validation dataset. That is, the threshold is the  $100(1 - \alpha)^{th}$  percentile of the  $M^2$  statistic in the MC samples of the validation dataset. Therefore, any data point with  $M^2$  exceeding the threshold ( $M^2 > M_{th}^2$ ) should be detected as a fault.

### 3.1.2. Method 2: Local density ratio (LDR) for non-Gaussian predictive distributions

If the predictive distribution is not well characterized by a Gaussian distribution (e.g., is multi-modal), then non-parametric methods are necessary to quantify the abnormality of each observation. For those cases, a LDR method is proposed, which is closely related to the so-called local outlier factor (Breunig et al., 2000). The LDR statistic quantifies the abnormality of each new observation with respect to its predictive distribution using an estimate of the density around the observation based on its  $k$ NN.

The  $k$ NN local density estimate  $\hat{f}(\mathbf{x})$  can be calculated as (Duda et al., 2001)

$$\hat{f}(\mathbf{x}) = \frac{k}{\sum_{p \in N_k(\mathbf{x})} d(p, \mathbf{x})} \quad (15)$$

where  $N_k(\mathbf{x})$  denotes the set of  $k$ NN of  $\mathbf{x}$  in  $\{\hat{\mathbf{x}}_t(i)\}_{i=1,\dots,N}$  and  $d(p, \mathbf{x})$  is the Euclidean distance between  $\mathbf{x}$  and a point  $p \in N_k(\mathbf{x})$ . Intuitively, the points close to its  $k$ NN will have high local density values, whereas points in more sparsely sampled or spread out areas will have low density.

Then, the LDR for an observation  $\mathbf{x}_t$  is defined as

$$\text{LDR}(\mathbf{x}_t) = \frac{\frac{1}{k} \sum_{p \in N_k(\mathbf{x}_t)} \hat{f}(p)}{\hat{f}(\mathbf{x}_t)} \quad (16)$$

which is the ratio of the averaged local density of the  $k$ NN of  $\mathbf{x}_t$  in  $\{\hat{\mathbf{x}}_t(i)\}_{i=1,\dots,N}$  to the local density of  $\mathbf{x}_t$ . A larger value of the

$\text{LDR}(\mathbf{x}_t)$  means that the observed point is far away from the samples of the prediction posterior and thus indicates higher likelihood that the observation  $\mathbf{x}_t$  is abnormal.

The number of  $k$ NN specifies the smallest number of data points in a cluster that will be considered as abnormal and is crucial for the algorithm to perform properly. In general, this number defines a tradeoff, because a small value of  $k$  will result in large fluctuations, whereas a very large value of  $k$  will reduce the detection sensitivity. As recommended in (Breunig et al., 2000), a minimum and maximum  $k$  can be chosen and, for each observation, the final value can be set equal to the maximum of LDR over  $k$ . The detection threshold  $\text{LDR}_{th}$  is obtained similarly to  $M_{th}^2$ .

### 3.2. Fault identification

Once a fault is detected, the next goal is to identify the main variables associated with the fault. Without using labeled fault examples, this step involves determining the observation variables with the abnormal deviations, which are most relevant to locate and troubleshoot the fault.

BRNN fault identification is obtained by applying the fault detection approach but independently for each variable. To determine which variables deviate abnormally, each observation variable is compared to its corresponding predicted marginal posterior distribution estimated from the BRNN samples. More specifically, the observation  $x_t^j$ , corresponding to the  $j^{th}$  system variable at time  $t$ , is compared to the predictive posterior distribution characterized by the samples  $\{\hat{x}_t^j(i)\}_{i=1,\dots,N}$ . This variable-wise comparison allows the identification of variables with values in low probability areas and thus more likely to be relevant for diagnosing the fault.

The marginal posterior distributions used for identification still take into consideration the spatial and temporal correlations in past data observations. Hence, the marginal distribution for each variable also evolves with dynamics that depend on other variables and past observations. This analysis sacrifices some information with regard to the complete joint distribution, as considered during fault detection, but is necessary to obtain variable specificity.

As with fault detection, two methodologies to quantify the fault identification deviation are described here, depending on the properties or assumptions placed on the predictive distribution. The same considerations apply to these methodologies.

#### 3.2.1. Method 1: Standard deviation for Gaussian predictive distributions

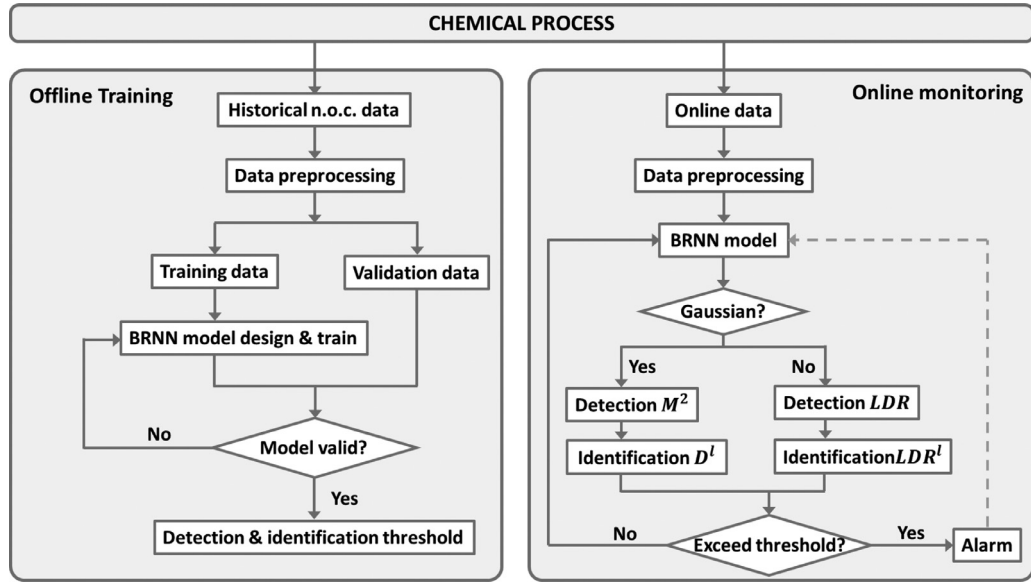
Assuming that the predictive distribution can be approximated by a Gaussian distribution, the number of standard deviations of each variable to its predictive mean can be used to measure the deviation. Using the same MC samples of the posterior predictive distribution generated for fault detection at time  $t$ ,  $\{\hat{\mathbf{x}}_t(i) = \{\hat{x}_t^l(i)\}_{l=1,\dots,m_x}\}_{i=1,\dots,N}$ , the mean  $\mu_t^l$  and standard deviation  $\sigma_t^l$  of each variable  $\hat{x}_t^l$  can be estimated by

$$\mu_t^l \approx \frac{1}{N} \sum_{i=1}^N \hat{x}_t^l(i) \quad (17)$$

$$\sigma_t^l \approx \sqrt{\tau^{-1} + \frac{1}{N} \sum_{i=1}^N (\hat{x}_t^l(i))^2 - (\mu_t^l)^2} \quad (18)$$

The deviation for each variable  $D^l$ ,  $l \in \{1, \dots, m_x\}$ , is then calculated from

$$D^l = \frac{x_t^l - \mu_t^l}{\sigma_t^l} \quad (19)$$



**Fig. 5.** Flowchart of the BRNN-based FDI methodology. The offline training stage (left) and the online monitoring stage (right) are shown in the figure. The procedure starts with offline training, and then the offline-trained model is used during online monitoring. The choice of statistics for detection and identification is made at design time.

The  $D^l$  can be either negative or positive, unlike  $M^2$  which can only be positive. Under a Gaussian approximation, variables are identified as significantly affected by the disturbance based only on whether  $D^l$  has a large magnitude (i.e., absolute value). Still, the sign of the deviation (positive or negative) can be helpful to operators because the sign explicitly indicates whether the variable is significantly higher or lower than expected. For a predefined significance level, the NOC validation dataset can be used to determine thresholds  $\{D_{th}^l\}_{l=1,\dots,N}$  such that variables with  $(D^l > D_{th}^l)$  are explicitly highlighted as abnormal.

### 3.2.2. Method 2: LDR for non-Gaussian predictive distributions

For more general distributions, and similarly to the fault detection procedure, the LDR can be used element-wise for fault identification by considering each variable separately in the calculation of the LDR. Given the true measurement  $\mathbf{x}_t = \{x_t^l\}_{l=1,\dots,m_x}$  and MC samples from predictive distribution  $\{\{\hat{x}_t^l(i)\}_{l=1,\dots,m_x}\}_{i=1,\dots,N}$ , the LDR for variable  $l$  can be calculated by

$$\hat{f}(x_t^l) = \frac{k}{\sum_{p^l \in N_k(x_t^l)} d(p^l, x_t^l)} \quad (20)$$

$$LDR^l(x_t^l) = \frac{\frac{1}{k} \sum_{p^l \in N_k(x_t^l)} \hat{f}(p^l)}{\hat{f}(x_t^l)} \quad (21)$$

where  $p^l$  is one of the  $k$ NN of  $x_t^l$  and  $d(p^l, x_t^l)$  is the Euclidean distance between the  $p^l$  and  $x_t^l$  sample. The same rule for selecting the number of nearest neighbors  $k$  discussed with regard to fault detection can be used here.

The variables associated with a large value of  $LDR^l$  can be explicitly selected as significantly affected by the fault. This is done similarly as for fault detection using the NOC validation dataset to determine a threshold  $LDR_{th}^l$  above which variables are considered abnormal. Alternatively, the variables can simply be sorted from the largest to the smallest such to emphasize the system variables that deviate the most.

### 3.2.3. Fault identification plots

The fault identification statistics of each variable can be visualized by plotting their values over time. The resulting plots are

visually similar to contribution plots (Miller et al., 1998; Zhu and Braatz, 2014). Their interpretation and analysis, however, are fundamentally different and are referred to here as *identification plots*. The main distinction is that the statistics in identification plots are specific to the current status of each variable and its dynamics, rather than as a relative component of a global statistic. These plots provide greater specificity in the analysis and allows the interpretation of the status of each variable directly.

### 3.3. FDI Scheme

For completeness, the overall methodology is summarized in Fig. 5. Although the figure explicitly shows the two methods for detecting and identifying faults, this decision of which method to use is actually done at the design stage rather than during operations. In either case, the BRNN model with variational dropout is crucial to the methodology by providing samples that characterize the uncertainty and directly enable both FDI. The  $M^2$  or LDR statistics are used to detect the fault in the system, while the  $D^l$  or  $LDR^l$  statistics are used to identify the impacted variables useful for locating the fault and possible root cause analysis. Minimal computation is needed for fault identification, having to calculate only some additional statistics on the same samples.

## 4. Case studies

In this section, the effectiveness of the proposed BRNN-based FDI method is demonstrated in two case studies: the benchmark Tennessee Eastman process synthetic dataset and a real dataset from a chemical plant.

For comparison, results are also shown for PCA (Jackson and Mudholkar, 1979; Kourti and MacGregor, 1996) and DPCA (Ku et al., 1995) FDI methods. For each method, both models with and without dimension reduction are considered, and identified by prefix 'r-' or 'f-', respectively. For the models with reduced dimension, parallel analysis (Downs and Vogel, 1993) is used to determine the number of PCs  $a$  to retain in the model. These (D)PCA-based methods are commonly accepted benchmark methods for algorithm comparison in the FDI community (Chiang et al., 2000a; Yin et al., 2014; De Ketelaere et al., 2015; Venkatasubramanian et al., 2003). DPCA in particular provides an interesting contrast to the

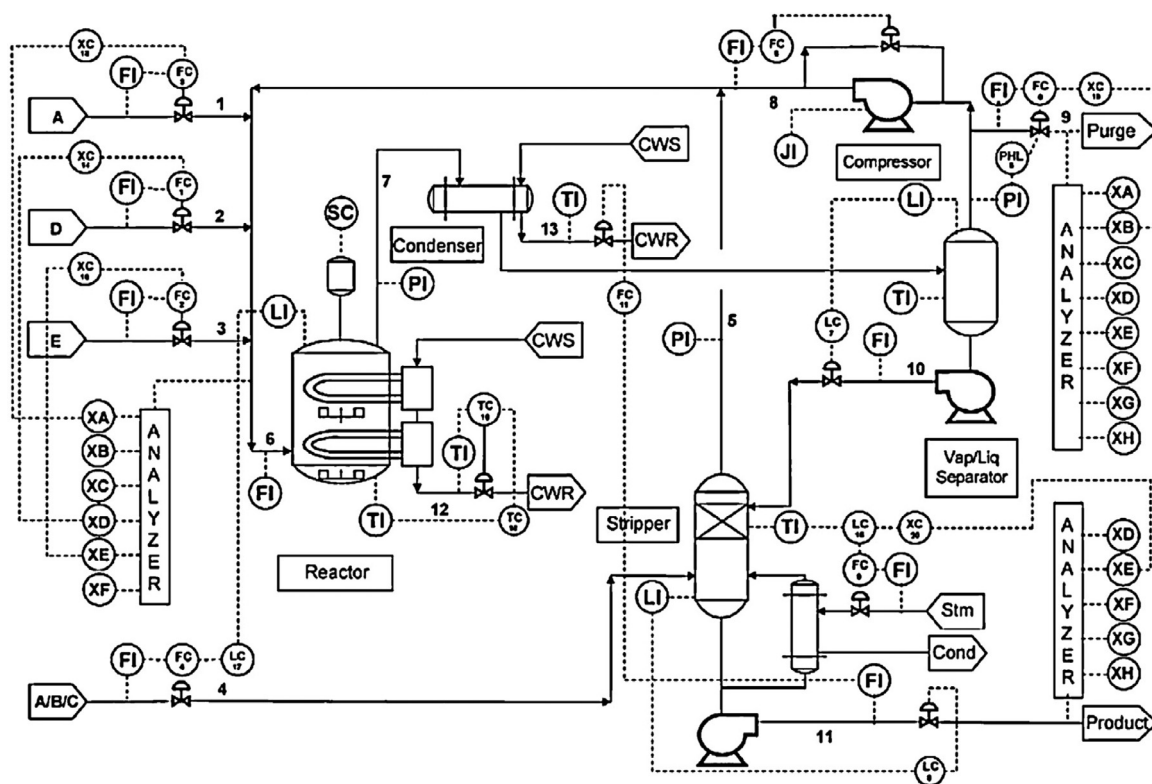


Fig. 6. A process flowsheet for the TEP with the second control structure in (Downs and Vogel, 1993).

proposed BRNN method because DPCA also models both spatial and temporal correlations, albeit in a limited form. As mentioned in the introduction, DPCA is limited to linear dynamics and correlations and scale poorly with increased temporal memory depth. The proposed BRNN method does not have these limitations.

In both case studies, the BRNN model constructions were implemented in TensorFlow (Abadi et al., 2015), and a number of BRNN model configurations and hyperparameters were tested. The choices included different recurrent node types (regular RNN, GRU, and LSTM cells), activation functions (i.e., linear, sigmoid, hyperbolic tangent, and rectifier linear), number of recurrent nodes/states  $m_s$ , number of recurrent layers, regularization hyperparameter values  $\lambda$ , dropout probabilities  $p_d$ , and RNN training parameters (e.g., learning rate). BRNN models were trained for each variation of these configurations and hyperparameters. The final model configuration and hyperparameters were selected as the ones that gave the maximum likelihood on the validation dataset. Only the results for the final BRNN model are shown in the next sections.

#### 4.1. Tennessee Eastman process

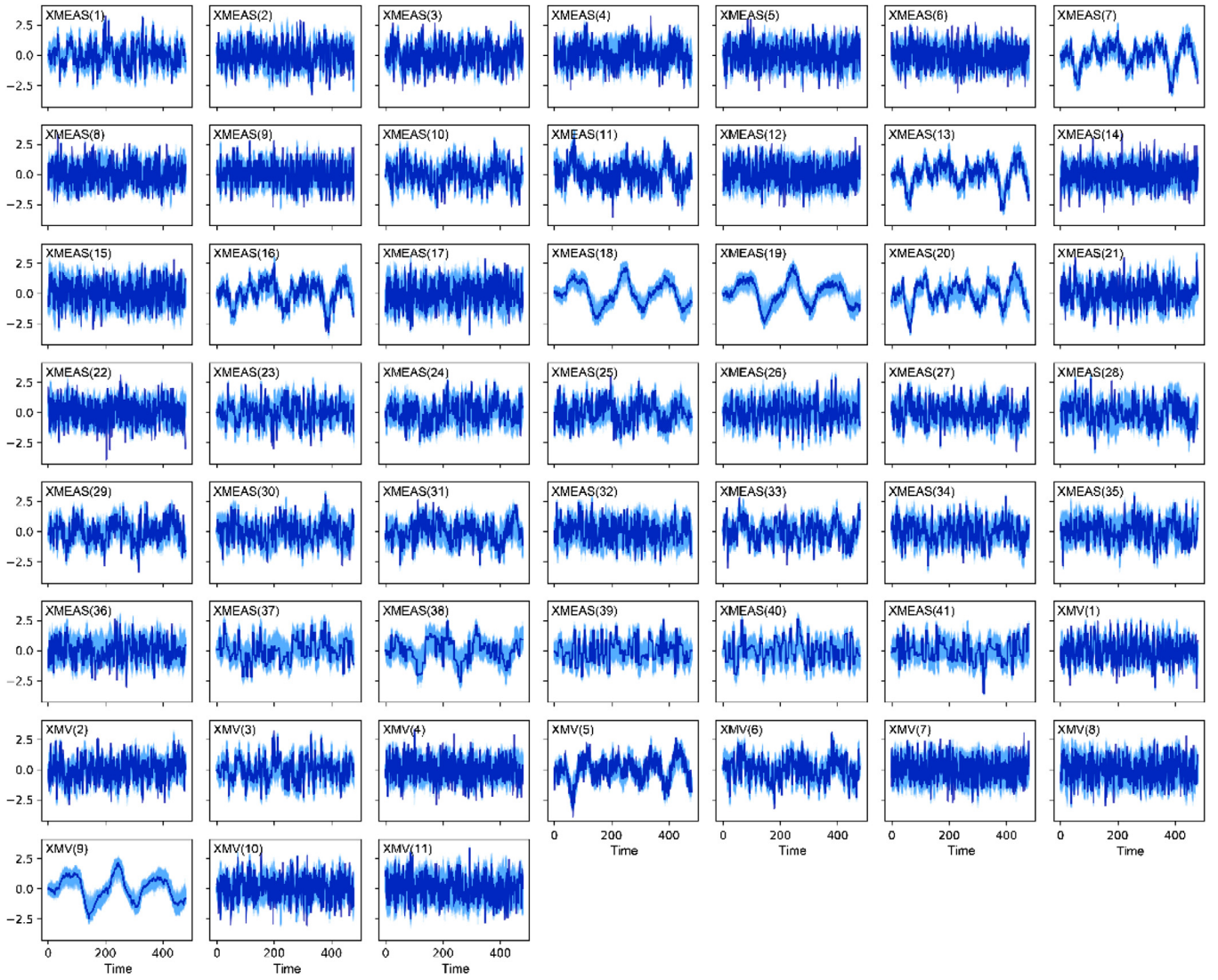
The Tennessee Eastman process (TEP) is a well-known benchmark by the Eastman Chemical Company for process monitoring and control studies. It is based on a realistic industrial process with properly modified components, kinetics, and operating conditions (Downs and Vogel, 1993). In this study, the second plant-wide control strategy was utilized, with the process flowsheet as shown in Fig. 6. The process contains eight components (A, B, C, D, E, F, G, and H) and five major units (a reactor, condenser, compressor, separator and stripper).

In this case study,  $m_x = 52$  variables are used to construct the monitoring system, of which 41 are sensor measurements (XMEAS(1)–XMEAS (41)) and 11 are manipulated variables

(XMV(1)–XMV(11)). During the NOC, the system is operating under one production mode and the sampling period is set to 3 min. The training data contains 480 samples and the validation data contains 960 samples. The TEP simulation contains 21 preprogrammed faults with different disturbance types and locations. Once a fault is introduced in the system, the system will either behave normally if the control system is effective in controlling the disturbance, or it will evolve outside the NOC region. For each set of data with a fault condition, the simulator first runs for 160 time points in the normal state, and then the corresponding fault disturbance is introduced with the simulator continuing to run for another 800 samples. The dataset used in this case study can be downloaded from the website of Prof. Richard Braatz (Chiang et al., 2000b). For further details about the TEP dataset, the reader is referred to Downs and Vogel (1993) or Chiang et al. (2000a).

Although a number of BRNN model architecture variations were tried as previously mentioned, the final BRNN model used in this case study contains one recurrent layer with regular RNN cell and linear activation function. A linear dense layer is used for the output layer, as is commonly done for regression tasks. This structure means that, in this case, the BRNN model implements a probabilistic linear state-space model. Although this architecture is simpler, it is also easier to train and achieved better performance than more complex structures and neuron types. The results are likely due to the fact that the inherent correlations and dynamics in the NOC data of TEP are well modeled as linear (Sun, 2020). The final model has  $m_s = 80$  hidden nodes in the recurrent layer, and is trained with regularization parameter  $\lambda = 10^{-4}$  and dropout rate  $p_d = 0.1$ . Given the linear structure of the model and by the central limit theorem, the predictive distribution by the final BRNN model is well approximated by the Gaussian distribution in this case. Thus, the  $M^2$  and  $D^I$  statistics are used for FDI. In any event, the results are quite similar for a number of configurations of the hyperparameters within a reasonable tuning range.





**Fig. 7.** BRNN model outputs for TEP NOC training data. The plot shows all 52 variables in TEP. The dark blue lines are the TEP measurements and the light blue lines correspond to the BRNN model predictive distribution outputs for the NOC data. For measurements under the NOC, the dark blue lines should lie within the predictive distribution. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

For the reduced dimensionality (D)PCA models used in the comparison, the number of PCs determined by parallel analysis is  $a = 12$  for r-PCA and  $a = 25$  for r-DPCA (with lag = 1). The fault detection procedure for r-PCA and r-DPCA use both the  $T^2$  and  $Q$  statistics, whereas f-PCA and f-DPCA use only the  $T^2$  statistic for fault detection, which plays the same role as the  $M^2$  statistic. Contribution plots are used for (D)PCA-based fault identification. For implementation details on the (D)PCA methods, the reader is referred to (Russell et al., 2000), (Zhu and Braatz, 2014), or (Chiang et al., 2000a).

The false alarm rate (FAR) and fault detection rate (FDR) are used to evaluate the fault detection performance of different algorithms:

$$\text{FAR} = \frac{\text{\# of samples with alarm during NOC}}{\text{total \# of samples during NOC}} \quad (22)$$

$$\text{FDR} = \frac{\text{\# of samples with alarm after the fault is introduced in the system}}{\text{total \# of samples after the fault is introduced in the system}} \quad (23)$$

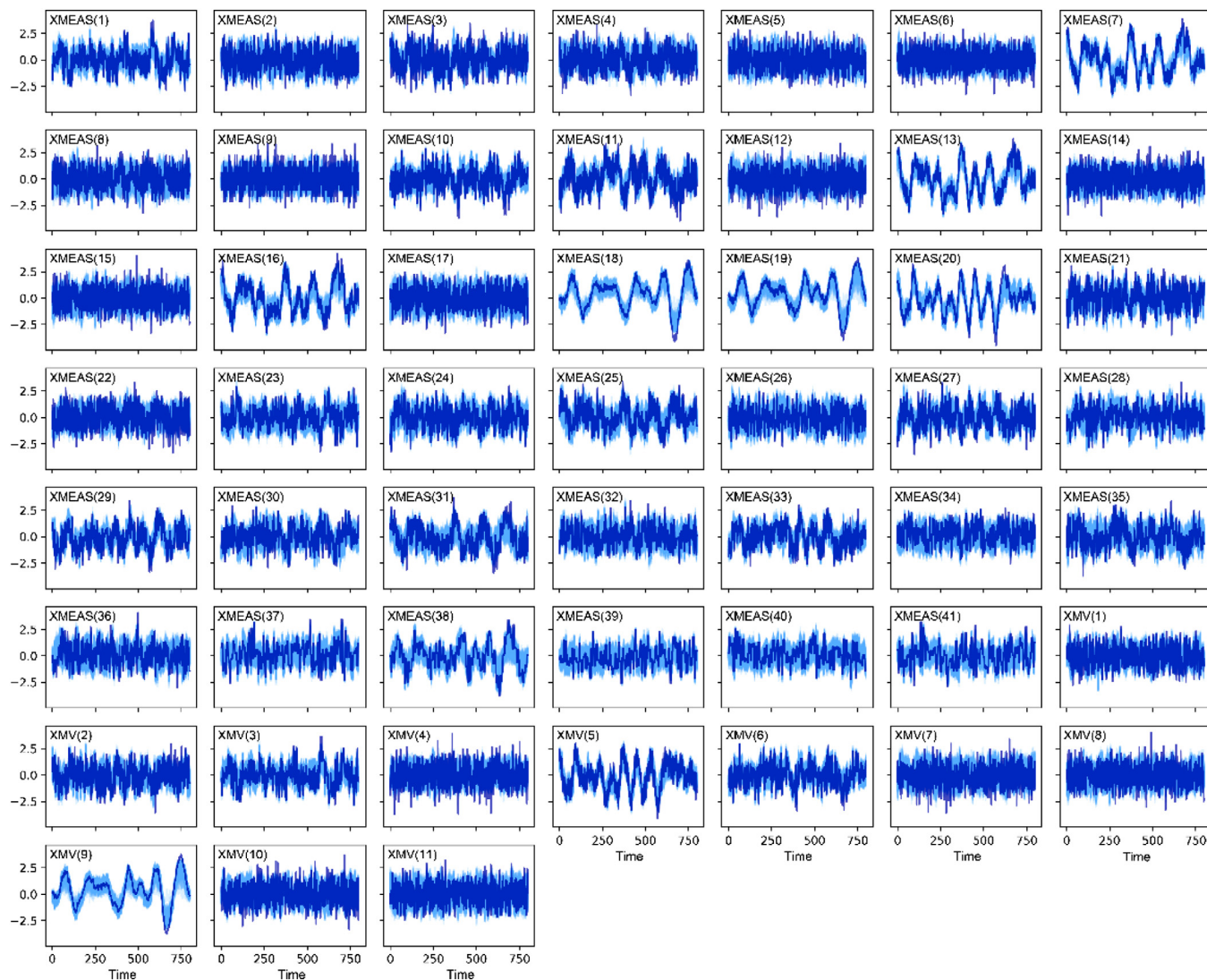
In words, the FAR corresponds to the frequency of spurious detection of faults under NOC, and FDR is the sample frequency of a fault being detected when a fault situation is present.

The dataset contains three types of faults: controllable faults, back-to-control faults, and uncontrollable faults. Controllable faults are disturbances that can be well compensated by the control system, and therefore the disturbance does not significantly affect the process state. In these situations, since the operator is not required to intervene, the FDR should ideally be as low as the FAR to avoid distracting the operators. Back-to-control faults are disturbances that are large enough to cause the system to initially deviate from the NOC, but for which the control system is able to compensate at least some aspects of the disturbance after some time. The process measurements return to the normal region after some time, but certain manipulated or input variables remain outside the normal regime. These represent sub-optimal or off-spec conditions that ought to be handled by an operator and to be detected accordingly. Moreover, the FDI result should accurately reflect the

**Table 1**

TEP fault detection percentage results. The FAR is shown for the NOC (in the first row), and the FDR is given for the 21 fault conditions.

Type NOC	Fault ID	BRNN 4.75	r-PCA ( $a = 12$ ) 5.00	f-PCA ( $a = 52$ ) 4.88	r-DPCA ( $a = 25$ ) 5.00	f-DPCA ( $a = 104$ ) 5.00
controllable faults	IDV(3)	5.00	7.00	19.75	6.12	22.25
	IDV(9)	5.00	7.88	15.25	8.87	21.37
	IDV(15)	7.12	10.62	26.87	11.13	36.63
back to control faults	IDV(4)	100.00	98.88	100.00	100.00	100.00
	IDV(5)	100.00	32.62	100.00	34.50	100.00
	IDV(7)	100.00	100.00	100.00	100.00	100.00
uncontrollable faults	IDV(1)	99.75	99.75	100.00	99.75	99.25
	IDV(2)	99.00	98.75	99.12	98.62	99.12
	IDV(6)	100.00	100.00	100.00	100.00	100.00
	IDV(8)	98.12	98.00	98.25	97.75	98.38
	IDV(10)	87.38	54.13	93.50	55.75	94.63
	IDV(11)	74.75	74.25	87.25	80.75	92.75
	IDV(12)	99.75	99.00	100.00	99.25	100.00
	IDV(13)	95.75	95.50	95.75	95.50	96.25
	IDV(14)	100.00	100.00	100.00	100.00	100.00
	IDV(16)	90.38	46.50	95.50	48.50	97.00
	IDV(17)	96.13	93.13	97.75	94.78	98.12
	IDV(18)	90.63	90.38	91.50	90.50	92.87
	IDV(19)	88.25	25.12	96.00	34.00	99.50
	IDV(20)	78.63	58.25	92.13	61.75	92.37
	IDV(21)	48.00	48.50	61.62	47.88	59.38

**Fig. 8.** BRNN model outputs for TEP NOC validation data.

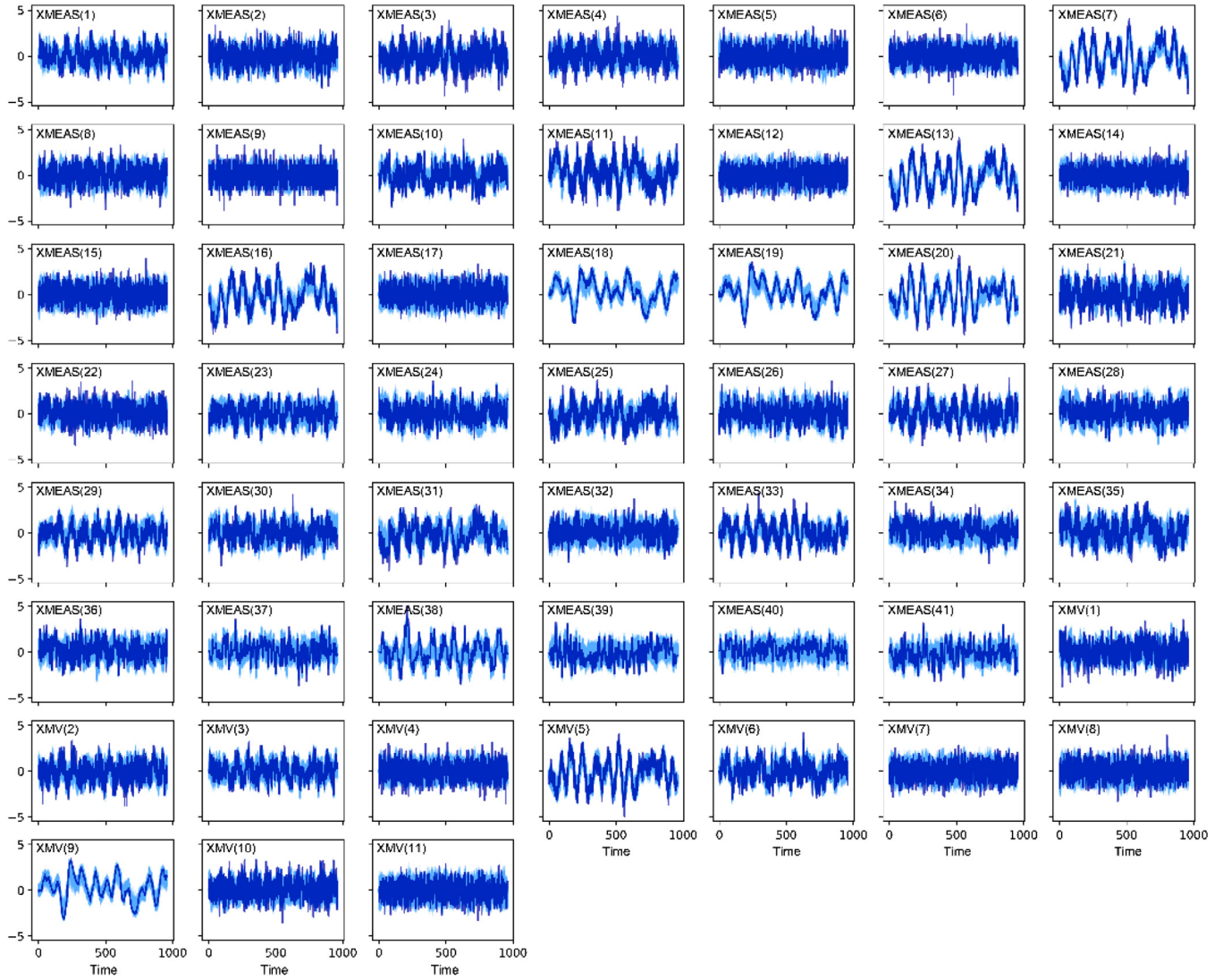


Fig. 9. BRNN model outputs for TEP Fault 3.

system state, such that its evolution back to control is apparent. Finally, uncontrollable faults are faults that cannot be handled adequately by the control system and require operator intervention. For both back-to-control and uncontrollable faults, the fault detection algorithm should ideally yield high FDR to notify the operator that the system has been disturbed outside the original NOC. It is worth noting that this “classification” is based on prior knowledge of the faults and used here only to facilitate the interpretation of the results; it was not used anywhere in the model training.

For fault identification, the proposed BRNN-based identification plots are compared with the (D)PCA-based contribution plots, which are shown for a representative fault of each of the aforementioned types. Note that, ideally, fault identification should accurately pinpoint the variables that are affected by the fault to provide the operators with specific information for them to analyze the situation and quickly diagnose the underlying root cause. Accordingly, it should be verified that no variable should be identified as abnormal for controllable faults. For back-to-control faults, the abnormal variables should first be identified, and only the corresponding tuned manipulated variables should remain identified once the system is back to control. For uncontrollable faults, the identification procedure should correctly locate the abnormal variables as soon as they are outside the NOC regime.

#### 4.1.1. Training and validation results on the NOC data

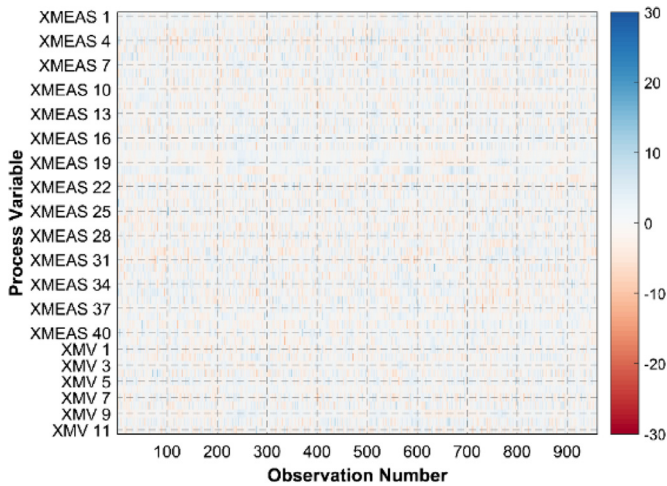
The training and validation results are first shown to demonstrate how the posterior predictions of the BRNN model characterize the NOC. The training results of the total 52 variables with centered and normalized values are shown in Fig. 7. The dark blue lines are the real data and the light blue lines are the posterior prediction samples by the BRNN model with  $N = 400$  model samples by variational dropout. The results do not differ significantly for  $N > 100$ . When the real measurements are within the predictive distribution (in dark and light blue in the figures, respectively), the system is considered normal. This condition can be observed in Fig. 7, which indicates that the trained BRNN model accurately captures the NOC pattern.

Then, to validate the model, the trained BRNN model is applied to a separate NOC validation dataset. These results are shown in Fig. 8. As observed in the training results, the real validation measurements lie within the predictive distribution. This result indicates that the model generalizes well, meaning that it is able to capture the normal pattern without overfitting to the training data, which is crucial to avoiding high FARs.

#### 4.1.2. Fault detection results

The fault detection results for the 21 predefined faults are shown in Table 1. The results are grouped according to one of the



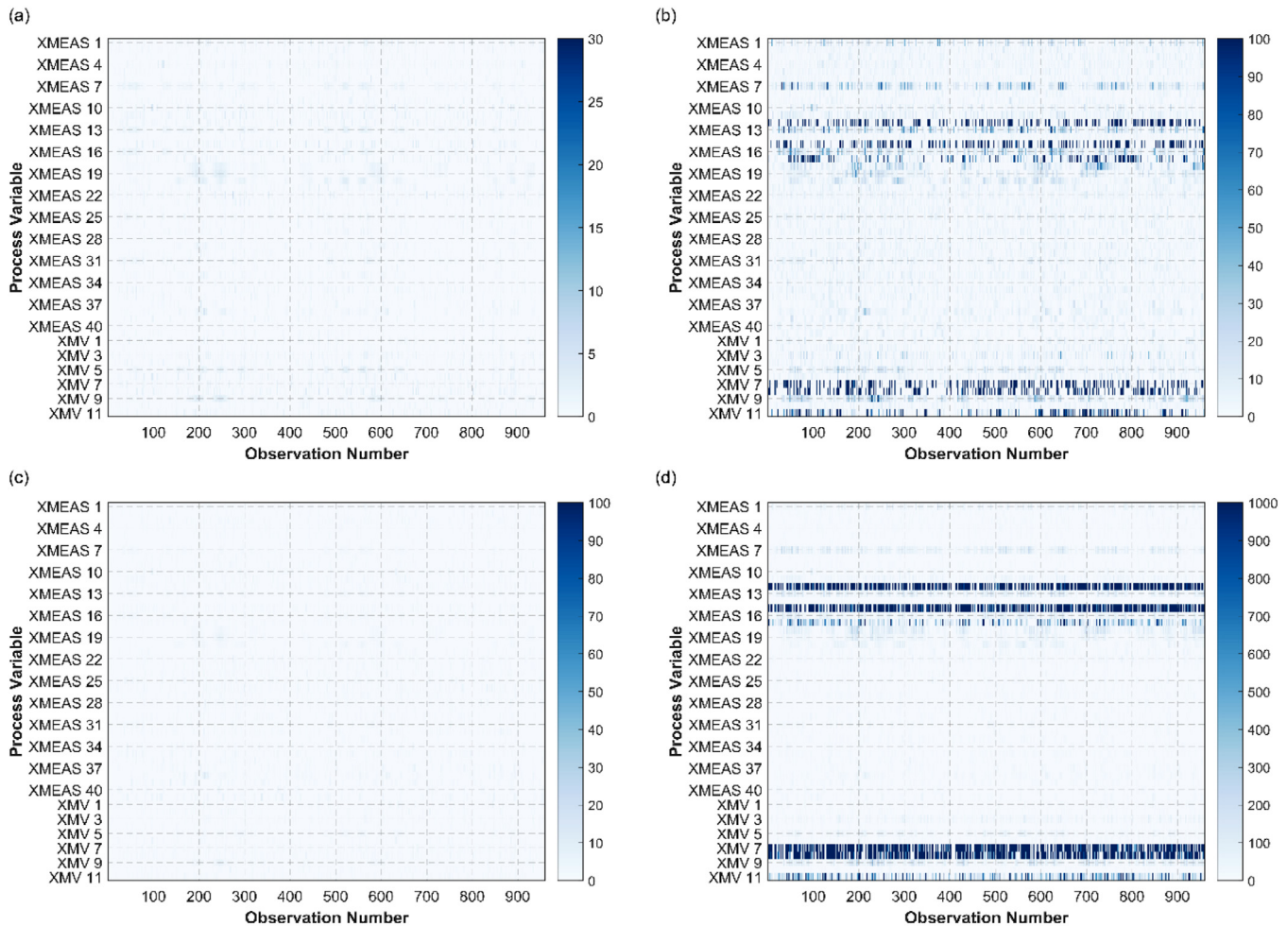


**Fig. 10.** Fault identification plot of the BRNN- $D^l$  statistic for Fault 3. The  $\{D^l\}_{l=1,\dots,52}$  values for the 960 timesteps are color coded in the identification plot. Variables with dark blues have high values of  $D^l$ , meaning that the variable has positively deviated from the NOC region. Conversely, variables with dark red have low values of  $D^l$  and have negatively deviated from the NOC region. A light color means the variable is not significantly affected by the disturbance. As expected, no variable significantly deviates from the NOC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

above-mentioned three types of faults. For all algorithms, the FDRs are estimated with regard to the threshold estimated for a FAR of 5%, and validated on NOC data as shown on the first row of the table.

As shown in Table 1, the proposed BRNN-based method yields close to 5% FDR on controllable faults, which is almost as low as the pre-determined FAR level. On the other hand, (D)PCA-based methods are overly sensitive in these cases, especially the models without model reduction, f-PCA and f-DPCA. These results show that (D)PCA-based methods cannot accurately differentiate the controllable faults from the other cases, because they do not appropriately characterize the dynamics of NOC, such as to determine if the situation is ultimately controllable. As previously explained, controllable faults should not trigger an alert because they are handled directly by the control system. The ability of the fault detection approach to differentiate between these situations is of crucial practical importance because alerts due to these situations will often be perceived as false alarms and can erode an operator's confidence in the method and the significance of its alerts. The BRNN method is observed to be more robust to controllable fault than the (D)PCA methods.

For both back-to-control and uncontrollable faults, the BRNN method reliably detected faults with high FDRs. Full PCA and DPCA models with the squared Mahalanobis distance were also able to detect the back-to-control and uncontrollable faults with high FDR. However, the (D)PCA models were overly sensitive for general fault



**Fig. 11.** Contribution plots for Fault 3 from (a) r-PCA, (b) f-PCA, (c) r-DPCA, and (d) f-DPCA. The plot shows the contribution factor and with the darkness of the blue color indicating the amount of deviation of the variable from the NOC region. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



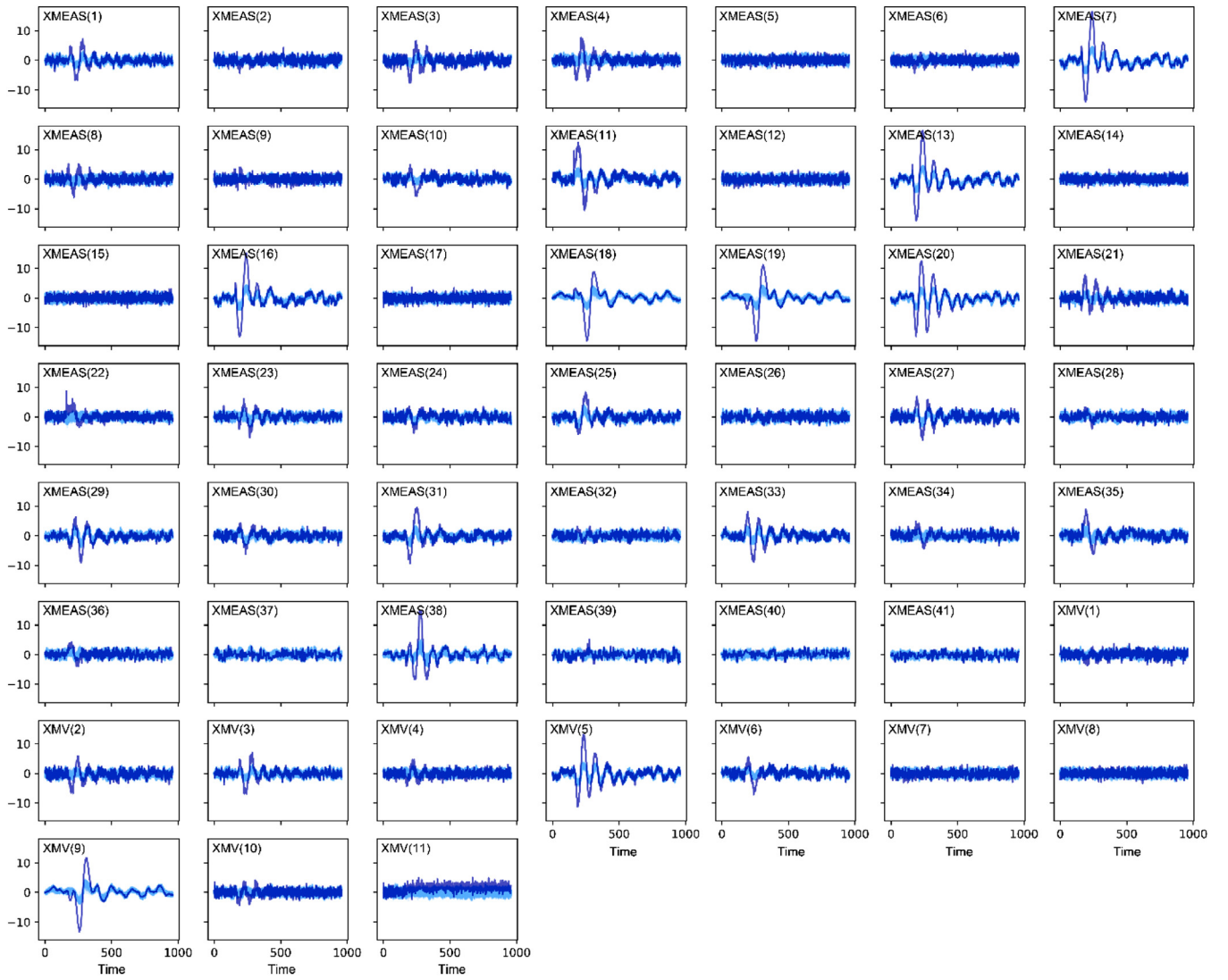


Fig. 12. BRNN model outputs for TEP Fault 5.

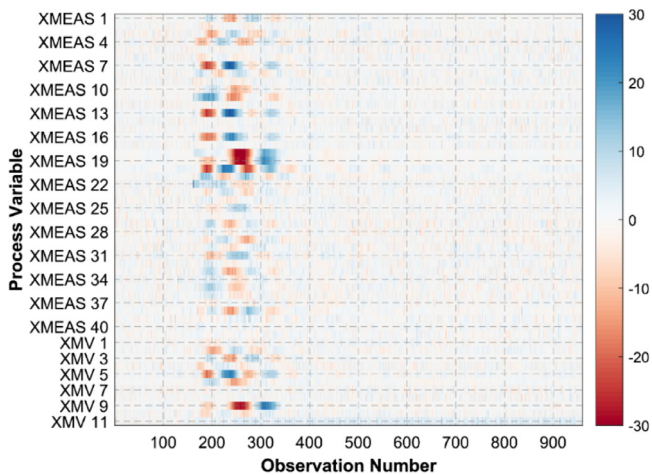


Fig. 13. Fault identification plot by BRNN- $D^I$  for Fault 5. The switch between dark blue and red colors shows that the system is undergoing large fluctuation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

detection purposes because they overreacted to controllable faults. Compared to the BRNN method, (D)PCA models emphasized higher

sensitivity to disturbances at the expense of an increased likelihood of unwarranted alerts. The reduced dimensionality (D)PCA models (i.e., r-PCA and r-DPCA), with the number of PCs determined by parallel analysis, responded more reasonably to controllable faults but also yield much worse performance compared to the BRNN method. In fact, they fail to reliably detect several back-to-control and uncontrollable faults (Faults 5, 16, and 19, for example).

It is insightful to consider how the temporal dynamics interact with the detection approach to lead to the measured FDR results. If a disturbance is such that the measurements oscillate around the NOC region, there will be moments in time that are momentarily indistinguishable from those in the NOC region. Since the BRNN is trained such that its state characterizes the NOC distribution in state space, it is understandable that some of these time points may not be detected as faulty. This observation explains the slightly lower FDR of the BRNN method for those cases. Unlike the BRNN, the (D)PCA methods do not model internal system dynamics under the NOC. Hence, since the internal dynamics are not considered when explaining the observed data, these models do not have this detection ambiguity. This lack of ambiguity is achieved at the expense of the inability by (D)PCA to assess whether a fault is controllable. In summary, the fault detection results indicate the BRNN has high detection accuracy and is able

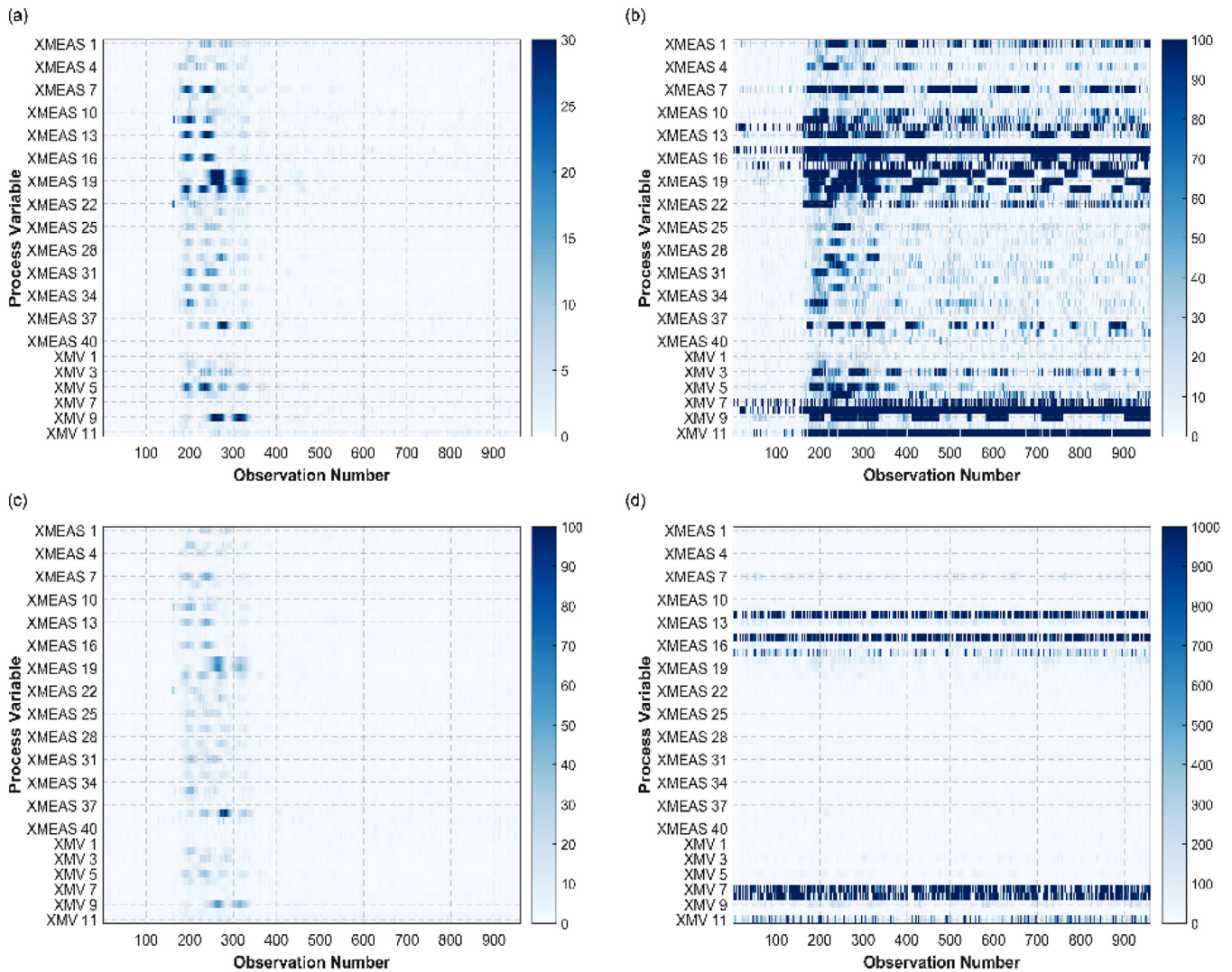


Fig. 14. Contribution plots for Fault 5 from (a) r-PCA, (b) f-PCA, (c) r-DPCA, and (d) f-DPCA.

to more robustly detect faults when operator intervention is truly necessary.

Specific FDI results are presented and discussed in detail below. Faults 1, 3, and 5 are presented because they are representative of each type. There was no significant differences between fault within the same type. The use of the BRNN contribution plot results for fault propagation analysis and one example is given for Fault 6.

#### 4.1.3. Controllable fault: Fault 3

Fault 3 is considered first as a demonstrative controllable fault. For Fault 3, the D feed temperature in Stream 2 has a step change at the 160<sup>th</sup> time point. Since this change in feed temperature is handled immediately and directly by the control system, the process is not driven outside its normal operating state. In this case, a data-driven fault detection algorithm should not trigger the alarm (beyond the chosen FAR).

The prediction results by the BRNN model are shown in Fig. 9. Similarly to the NOC case, the dark blue lines (i.e., real measurements) are within the distribution high-likelihood area characterized by the light blue lines, which indicates that the system is operating under the NOC.

Fault identification results by the BRNN and (D)PCA methods are shown in Figs. 10 and 11, respectively. The color indicates the

deviation from the NOC over time for each of the 52 variables. The  $D^I$  statistic (c.f. Eq. 19) is used in the BRNN identification plot. As shown in Fig. 10, the BRNN model identifies that no variable has its normal operating dynamics significantly affected by Fault 3, as is expected. In contrast, the contribution plots in Fig. 11bd, by the full PCA and DPCA models, incorrectly identify several variables as being affected by the disturbance even before the introduction of the disturbance (at the 160<sup>th</sup> sample). This further demonstrates the oversensitiveness of those models. The r-PCA and r-DPCA models have identification results that are similar to those of the BRNN model and are somewhat robust to controllable faults, but at the expense of robustness in fault detection.

In summary, for controllable faults, BRNN-based FDI is robust and successfully characterizes those disturbances as corresponding to NOC. r-PCA and r-DPCA methods gave similar fault identification results but have lower detection rates (c.f. Table 1). The f-PCA and f-DPCA models were clearly oversensitive, have high FARs, and incorrectly characterized the controllable faults in the contribution plots.

#### 4.1.4. Back-to-control fault: Fault 5

Fault 5 is a representative example of a back-to-control fault. This fault involves a step change in condenser cooling water inlet temperature. This step change requires a step change in the





Fig. 15. BRNN model outputs for TEP Fault 1.

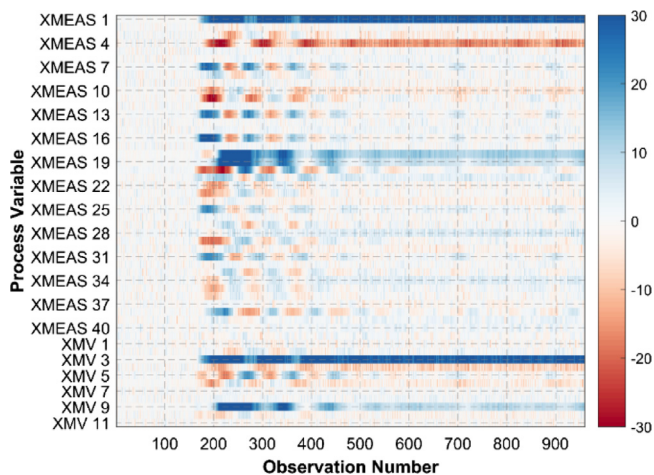


Fig. 16. Fault identification plot by BRNN- $D^l$  for Fault 1. The root cause for this uncontrollable fault can be assessed by looking at the variables that are persistently off the NOC region.

condenser cooling water flow rate XMV(11) by the control system. While the fault is ultimately controllable, the fault causes the sys-

tem at first to operate off-spec, or at least sub-optimally. In this particular, immediately after the fault occurs, the system oscillates with about 32 variables exhibiting this similar transient oscillation behavior. The process returns to control after about 10 hours, at which point the sensor measurements XMEAS(1)–XMEAS(41) are back to their pre-disturbance set-points, and only the manipulated variable XMV(11) remains outside the NOC regime, tuned so as to compensate the step change in condenser cooling water inlet temperature.

The BRNN results for all of the variables are shown in Fig. 12. As expected, when the fault is introduced, several measurements in dark blue lines deviate from the posterior predictive distribution under the NOC shown in light blue. After about 200 data points the system returns to control, verified by the fact that all the system measurements (XMEAS(1)–XMEAS(41)) are back within the predictive NOC region while only the manipulated variable XMV(11) maintains a systematic deviation off-the-center of the BRNN model prediction distribution. These results show that the BRNN model is able to correctly identify the NOC pattern and how the deviation from the predictive distribution accurately locates the faulty variable under disturbance. The BRNN model is also able to better assess the state of the system, distinguishing the back to control faults from the uncontrollable faults by showing the tran-

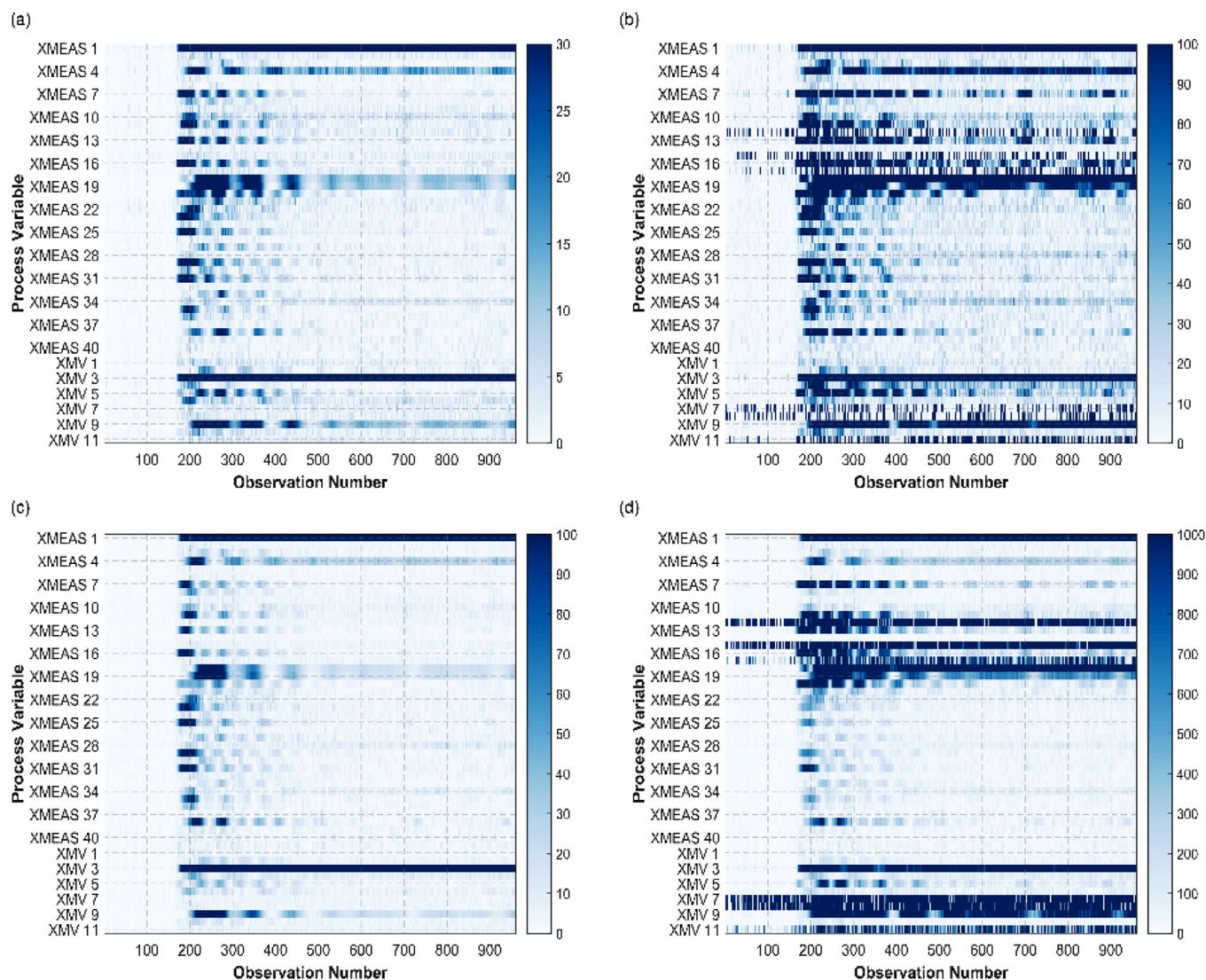


Fig. 17. Contribution plots for Fault 1 from (a) r-PCA, (b) f-PCA, (c) r-DPCA, and (d) f-DPCA.

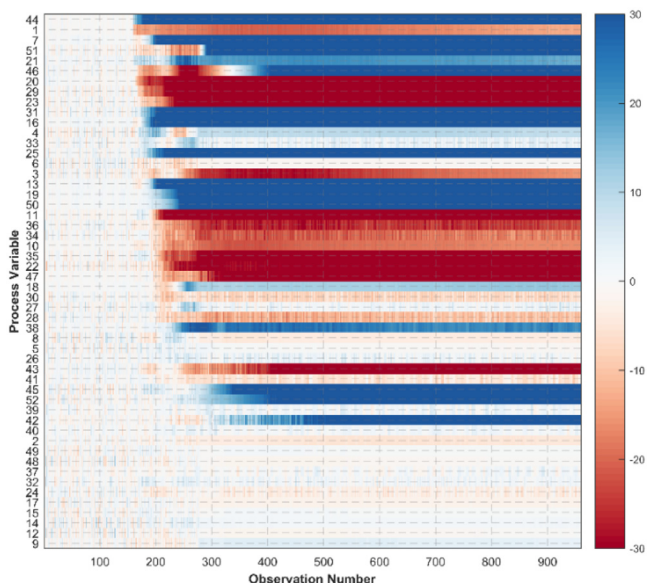


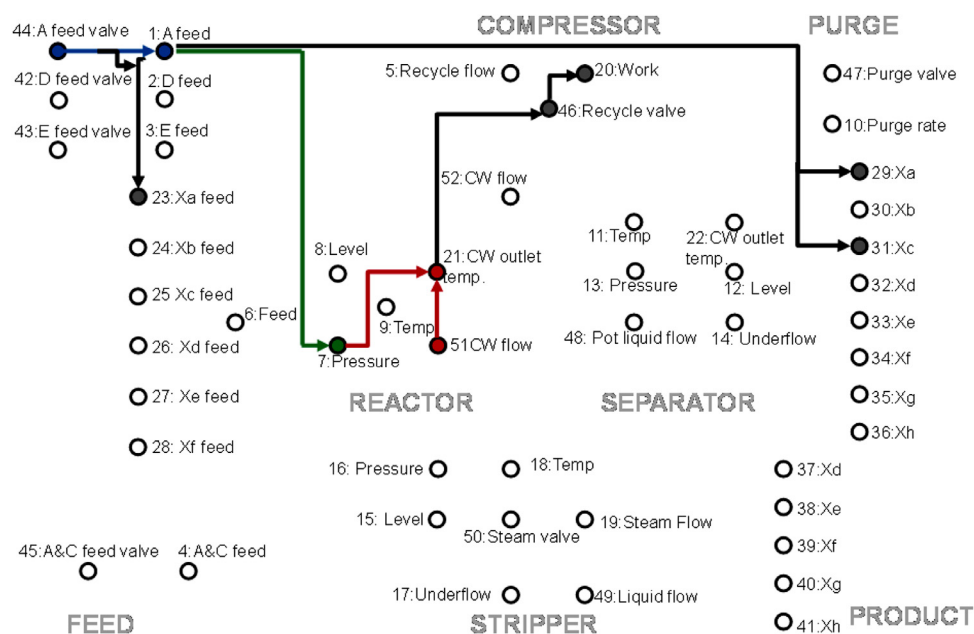
Fig. 18. Sorted fault identification plot according to the detected deviation occurrence time of BRNN- $D^f$  for Fault 6.

sient deviation of the process variables and their return to the NOC region.

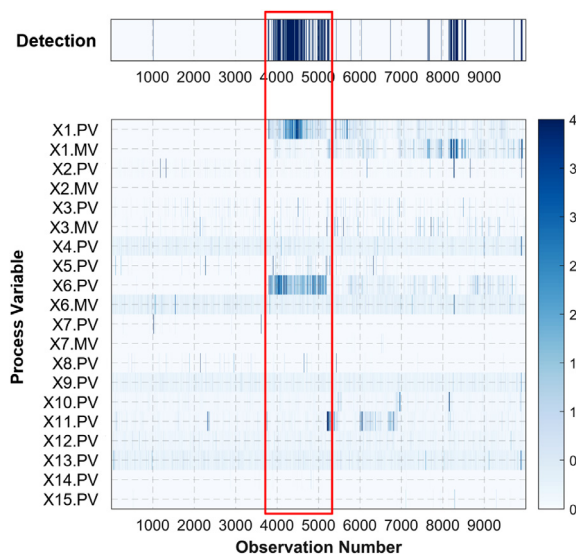
The fault identification plot by BRNN is shown in Fig. 13. This example showcases the typical pattern of back to control faults, with several measurements outside the predictive region after the fault is introduced and only the manipulated variables deviating once the system is back to steady state. In this case, about 32 variables are affected once the disturbance is introduced to the system, and the color switches between blue and red, indicating the system is oscillating. The plot also clearly shows how, after the 360<sup>th</sup> time point, all system variables except XMV(11) are undoubtedly back to normal. XMV(11) remains consistently above the predictive mean after the fault as that is forced by the controller to compensate for the fault. However, the magnitude of the deviation of XMV(11) is relatively small, indicating that the disturbance is no longer critical.

The BRNN identification plot also contains crucial information for locating the likely root cause of the fault. The plot clearly shows that XMEAS(22) is the first variable positively deviated from the predictive distribution, which indicates the higher than normal separator cooling water outlet temperature. Combined with the fact that the condenser cooling water flow rate is increased to compensate for the disturbance to the system, one would rea-





**Fig. 19.** Fault 6 propagation path at the 180<sup>th</sup> data point (1 h after the fault occurs). Colored nodes indicate that the corresponding variable has been detected as deviating significantly from the NOC.



**Fig. 20.** FDI by the BRNN model on the Fault 1 testing data. The red box indicates the period with the foaming event recorded by the operator. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

son that the root cause is the increase in the condenser cooling water temperature. After the condenser cooling water temperature increases, the outlet stream from the condenser to the separator also increases the temperature, resulting in an increase in the temperature in the separator, which finally results in the increase in separator cooling water outlet temperature.

For comparison purposes, the contribution plots by PCA and DPCA methods are shown in Fig. 14. The r-PCA and r-DPCA model results in Fig. 14ac fail to identify the consistent deviation in XMV(11), which clearly explains their detection results for this fault. These results demonstrate again that the r-PCA and r-DPCA models can exhibit much lower detection and identification sensitivity than the BRNN method. The results of the f-PCA and f-DPCA models in Fig. 14bd clearly identify the deviations in several vari-

ables. However, the f-PCA and f-DPCA are oversensitive and identify variables in an unspecific manner, which prevents those statistics from being used, at least directly, by operators for diagnosing the root cause of the fault.

For the back-to-control fault, BRNN FDI had high accuracy and robustness. Moreover, this method yielded more specific information for evaluating the state of the system. By inspecting the identification plot, operators have a clear view about which variables are affected by the disturbance and are able to assess the type of the fault occurring and the current stage of the system.

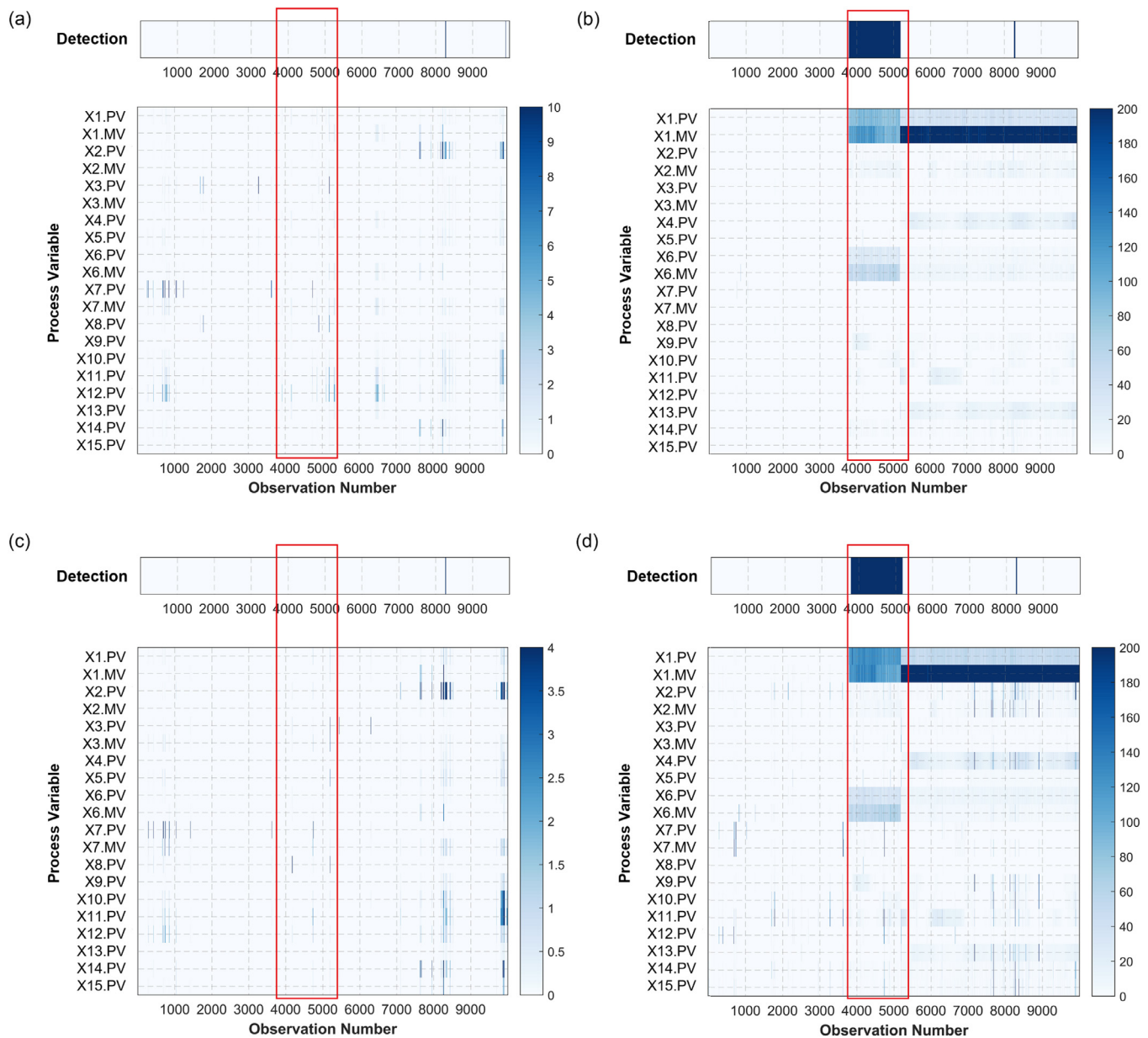
#### 4.1.5. Uncontrollable fault: Fault 1

An uncontrollable fault is now considered. Fault 1 involves a step change in the A/C feed ratio in Stream 4, which results in an increase in the C feed and a decrease in the A feed. This subsequently leads to a decrease in feed A in the recycle Stream 5 and the controller reacts by increasing the A feed flow in Stream 1. These two effects conflict with each other, thereby shifting the system to an uncontrollable operating situation.

The BRNN model output results are shown in Fig. 15. After the fault is introduced to the system, more than half of the variables are observed to deviate significantly from the BRNN predictive NOC region. All of the (D)PCA methods are also capable of detecting this fault.

The corresponding BRNN fault identification statistics are shown in Fig. 16. Since the system is seriously affected by the disturbance and several variables associated with material balances (e.g., composition, pressure) change significantly, this fault is easily detected. The long-term and uncontrollable nature of the fault on these measurements and manipulated variables can also be observed in the identification plot, making the fault easy to diagnose based on those variables.

As before, the contribution plots by (D)PCA methods are shown in Fig. 17. The r-PCA and r-DPCA models, in Fig. 17ac, both give somewhat results similar to those of the BRNN model in Fig. 16. However, both of them fail to identify the continued deviation in XMV(4) (shown between XMV(3) and XMV(5) in Fig. 17) for instance, which is the manipulated variable for total feed flow in Stream 4 and clearly plays a central role in the fault. In con-



**Fig. 21.** FDI by (D)PCA methods for Fault 1: (a) r-PCA, (b) f-PCA, (c) r-DPCA, and (d) f-DPCA. The red box indicates the period with the foaming event recorded by the operator. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

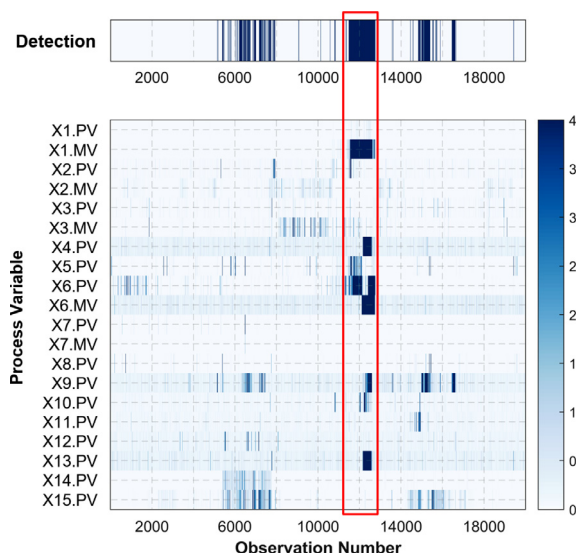
trast, it is clearly identifiable from the BRNN results in Fig. 16 that XMV(4) has negatively deviated from the NOC region. The contribution plots of f-PCA and f-PCA in Fig. 17bd show the identification of the involved variables, but again highlight several other variables that are unrelated to the fault and operating within their normal pattern (such as XMV(11)). As previously observed, this again shows that the f-PCA and f-DPCA models are overly sensitive and their identification results require substantial additional processing such that operators cannot directly use them to diagnose the fault.

In summary, the BRNN model is able to accurately and robustly detect and identify uncontrollable faults. Perhaps most crucially, BRNN identification plots provide clear information that is directly useful for root cause analysis. While several (D)PCA models are also able to detect uncontrollable faults, their identification results are less accurate and precise than for the BRNN model.

#### 4.1.6. Fault propagation path analysis: Fault 6

This section shows how the accuracy and specificity of the BRNN identification statistics can be used for fault propagation path analysis. The key observation is that the chronological sequence of events of when each variable deviates significantly from its NOC is useful information to understand the start and evolution of the disturbance through the process (Chiang and Braatz, 2003). The BRNN method can extract this information with a high degree of temporal precision. This information can then be combined with expert knowledge of the process to examine the propagation of the fault through the system.

This approach is exemplified here using Fault 6, which is an uncontrollable fault induced by a loss of feed A in Stream 1. The loss of component A thus causes the control system to increase the manipulated variable XMV(3) in order to increase A in the system and attempt to compensate for the disturbance. However, since there is no component A in Stream 1, the control system fails to take the



**Fig. 22.** FDI by BRNN for Fault 2. The red box indicates the period with the foaming issue as recorded by the operator. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

system back to NOC. Due to the severity of this fault, a large portion of system variables is affected.

The temporal sequence of the fault through the process is achieved by sorting the identification plot according to the time when each variable significantly deviates from the NOC. For this approach, one needs to estimate the threshold used to determine when the deviation is significant. In our case, this is estimated using the NOC validation set and determined to be  $D_{th}^l = 4.8$ . Then, once a fault is detected, the process variables are sorted according to the time index at which its  $D^l$  statistic first exceeds the threshold, yielding the sorted identification plot shown in Fig. 18. The y-axis numbers 1–52 correspond to [XMEAS(1), ..., XMEAS(41), XMV(1), ..., XMV(11)].

A diagram of the fault propagation path is then obtained by combining the timing results of the sorted BRNN identification plot with the knowledge of the process, as demonstrated in Fig. 19 for Fault 6. When the fault is introduced in the system, XMEAS(1) and XMV(3) are affected and deviates from the NOC first. Then, after a few minutes, the reactor pressure measurement XMEAS(7) is affected. Then the reactor cooling water system is also affected due to the change in the mass inside the reactor and both XMEAS(21) and XMV(10) deviate from the NOC. The diagram in Fig. 19 highlights that, after 1 h, the fault has already propagated to the final product and the concentration of A and C have been affected, thus clearly showing the impact of the fault in the system at that point in time.

The approach outlined here shows how the properties of the BRNN method can be used to easily determine and visualize the fault propagation path. This information is crucial to operators to accurately diagnose the fault and determine which parts of the process have been affected.

#### 4.2. Real industrial dataset

The next case study further demonstrates the efficiency of the proposed BRNN method on a real dataset from a chemical manufacturing process. The use of this method for real-time FDI is a promising application for the next generation of process monitoring systems in chemical plants. The complex nature of real chemical manufacturing processes and their intricate control system dynamics, make BRNN the best-suited tool to extract and

recognize these patterns from data in comparison to traditional methods.

The dataset pertains to the operation of an amine tower. The column experienced foaming issues resulting in faults that decrease the efficiency of the process. There are a total of 20 sensor measurements with a sampling time of  $t = 1$  min. A total of two months of data are available. Two events have been recorded by operators as a result of the foaming issue in the tower. However, it is also possible that additional disturbances are encountered during the two-month operating window that have been previously missed.

The final BRNN model uses standard RNN cells with the sigmoid activation function. There is one hidden recurrent layer with 40 units (i.e., 'state variables'). The dropout probability is set to  $p_d = 0.1$  and the regularization parameter to  $\lambda = 10^{-5}$ . The BRNN model using LSTM or GRU cells yield similar performance in spite of the higher complexity and thus those results are omitted. Similar to the TEP data, the results are similar for a number of configurations of the hyperparameters within a reasonable tuning range.

For comparison, the PCA and DPCA models, with the number of PCs determined by parallel analysis and full models without dimension reduction, are also applied. The number of PCs from parallel analysis is determined to be  $a = 6$  for r-PCA and  $a = 9$  for r-DPCA.

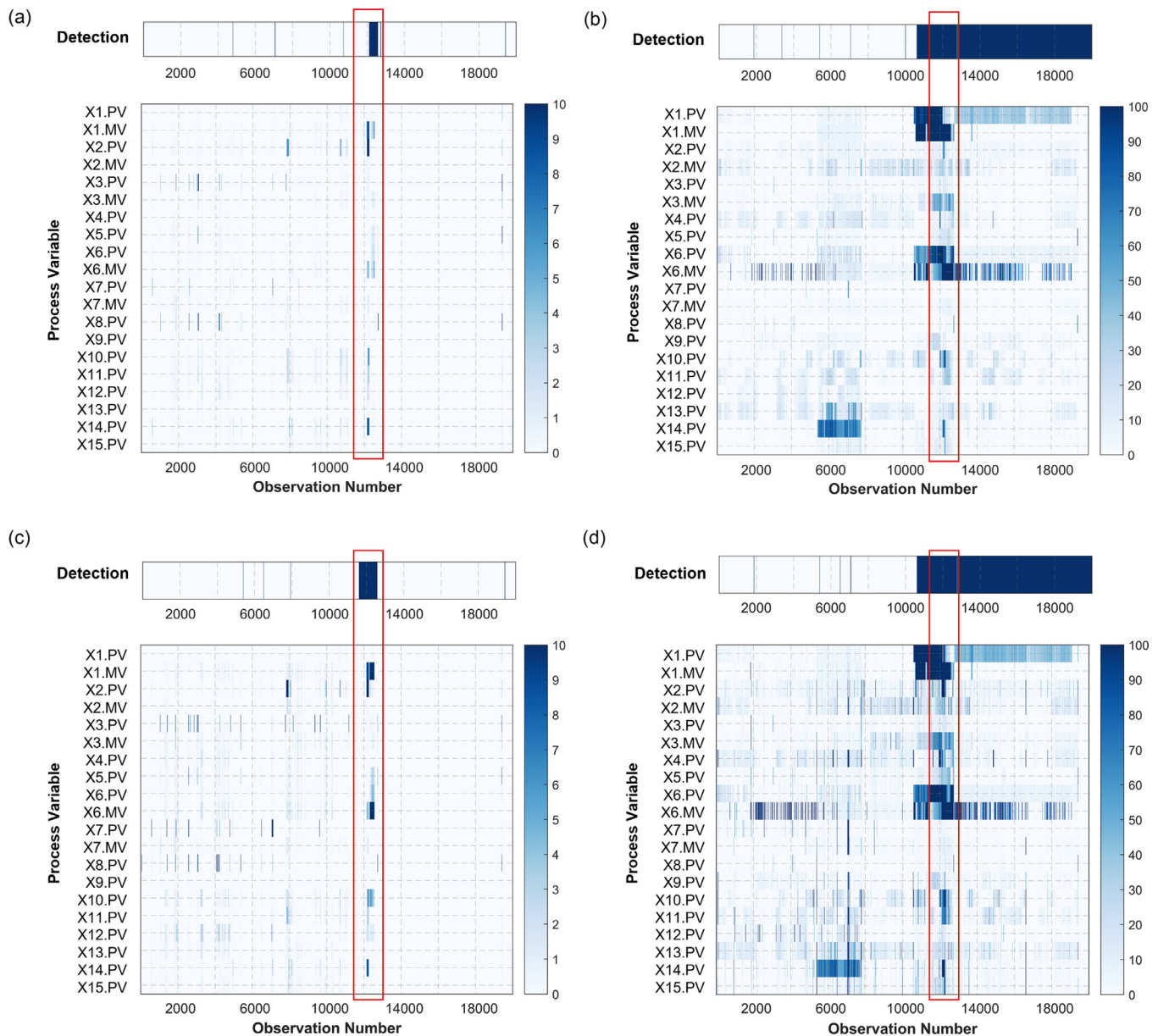
Due to the sensitivity of the data, the actual measurement values and the BRNN model predictions are omitted and only the detection and identification results are shown. The posterior predictive distribution is observed to be multi-modal and thus the LDR statistics are used for FDI. The number of kNN is set to the range of 10 to 20. The dataset is divided into a 35-day training dataset, a 17-day validation dataset, and two testing datasets. The first testing dataset spanned 7 days containing Fault 1, and the second testing dataset spanned 14 days containing Fault 2.

##### 4.2.1. Fault 1 results

The BRNN FDI results for Fault 1 are shown in Fig. 20. The BRNN model successfully detects the documented event, marked by the red box in the figure. The model also accurately pinpoints the variables that are most affected by the foaming issue, X1.PV and X6.PV. Moreover, it also highlights several points after the 8000<sup>th</sup> time point that may have been originally missed. During these later periods, the X1.MV sensor measurement is identified by the BRNN method. This is subsequently verified to have been the result of large unexplained fluctuations in that variable and that the BRNN has performed as expected.

For comparison, the corresponding FDI results by (D)PCA methods are shown in Fig. 21. The r-PCA and r-DPCA models simply fail to detect the fault. In the contribution plots, the r-PCA and r-DPCA models also fail to identify any variable that is noticeably affected by the foaming event. While (D)PCA models with reduced dimensionality determined by parallel analysis have been widely applied (Chiang et al., 2000a; Yin et al., 2014; De Ketelaere et al., 2015; Valle et al., 1999), they are incapable of accurately detecting the main fault in this case. For f-PCA and f-DPCA models, the  $T^2$  statistic is able to detect the documented fault. The contribution plots also identify X1.PV and X6.PV as being associated with the fault. However, X1.MV and X6.MV are also identified as abnormal and as more significantly than X1.PV and X6.PV. While those variables are likely affected by the fault, they are operating normally with respect to the control system dynamics and thus should not have been identified. Furthermore, some of these variables continue to be highlighted well after the issue is resolved. The (D)PCA models also only scantily detect the deviations in the later time that are correctly highlighted by the BRNN.





**Fig. 23.** FDI by (D)PCA methods for Fault 2: (a) r-PCA, (b) f-PCA, (c) r-DPCA, and (d) f-DPCA. The red box indicates the period with the foaming issue as recorded by the operator. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

#### 4.2.2. Fault 2 results

The BRNN FDI results for Fault 2 are shown in Fig. 22. The proposed method successfully detects the documented fault and identifying the related variables, as marked by time interval with the red box. For the earlier period around the 6000<sup>th</sup> time point, the BRNN model detects a disturbance and identifies the deviation in X14.PV, X15.PV, and X9.PV. The relative magnitude of the deviation during those periods is not as significant as that during the documented fault period. This assessment was then verified to be fully warranted by inspection of the recorded sensor measurements. Analysis of the period around the 15000<sup>th</sup> data point yield similar results.

As before, the (D)PCA FDI methods are also applied. Their results are shown in Fig. 23. As observed for Fault 1, the r-PCA and r-DPCA models are not as sensitive to the fault and only partially detect the documented fault period, and they did not detect the earlier event highlighted by the BRNN method. The identification

plots by the r-PCA and r-DPCA models in Fig. 23ac also only identify a limited number of faulty variables. The f-PCA and f-DPCA models are again overly sensitive for both FDI, flagging much of the data period. After the foaming issue occurred and the operator intervention, the control system is able to compensate for the disturbance after a while. However, f-PCA and f-DPCA models incorrectly continue to assess the system as in an abnormal state even though the foaming issue has been fully resolved. This can also be observed from the contribution plots in Fig. 23bd, wherein variables X1.PV and X6.MV are identified as problematic during and long after the resolution of the fault.

To summarize, this case study on real data from a chemical process demonstrates the higher accuracy, specificity, and robustness in FDI of the BRNN-based method over (D)PCA-based methods. The proposed method is also shown to provide precise and easily interpretable results for prompt diagnosis and mitigation of fault events in real manufacturing processes.



## 5. Conclusion

This article proposes a novel BRNN-based FDI method for manufacturing processes. The proposed method simultaneously tackles three key challenges in modeling real process data: (1) concurrent spatio-temporal correlations, (2) nonlinearity, and (3) incomplete characterization of the uncertainty in process noise and dynamics. The BRNN model addresses these challenges because of its probabilistic framework built on RNN models. And, for implementation efficiency, the inference is made using variational dropout, which both regularizes the NN during training and efficiently estimates the uncertainty as it evolves through time.

The uncertainty estimates of the BRNN model play a crucial role in FDI. By continuously estimating the uncertainty, the BRNN model provides adaptive confidence intervals that fully characterize the system dynamics based on the current and past information. As demonstrated here, the BRNN framework therefore enables:

- (1) fault detection in processes with nonlinear dynamics, and
- (2) direct fault identification with easily interpreted identification plots and fault propagation path analysis.

The effectiveness of the proposed BRNN method is demonstrated in two case studies: (1) the benchmark TEP dataset and (2) a real chemical manufacturing dataset. The proposed method is compared to the widely applied PCA and DPCA methods, using either full and reduced dimension models. The comparisons show that the BRNN model provides results that are accurate and more specific and directly relevant for fault identification. Furthermore, based on its results, one can distinguish the nature of the faults, between controllable, back to control, or uncontrollable faults.

More broadly, the application of Bayesian methods to fault detection is not a widely explored field. To that end, this paper demonstrates a novel framework involving the systematic application of spatio-temporal models with Bayesian estimation such that the posterior inference results are directly relevant for detection and identification. In this case, a BRNN is used, but the strategy could be adapted for other spatio-temporal models such as dynamic process models.

The proposed BRNN-based FDI framework can be directly applied to any manufacturing process with historical NOC measurements without significant modifications. Moreover, the easy implementation of variational dropout to any model architecture and concurrent online calculating capability make BRNN feasible for large-scale industrial applications.

Some considerations for future work might include:

- (1) The online adaptation of the BRNN model for changing NOC. In real chemical processes, the process conditions evolve and it is unlikely that the training data can cover all of the NOC modes. Thus, online adaptation is crucial for reducing false alarms and maintenance costs.
- (2) While proposed here for FDI, the BRNN model framework also has broad potential applications in industrial manufacturing processes related to time series analysis. The variational dropout can be applied to any deep learning model without modification of the model architecture, which makes it a preferable probabilistic model as compared to other recent advanced techniques.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

## CRediT authorship contribution statement

**Weike Sun:** Methodology, Software, Writing - original draft. **Antonio R.C. Paiva:** Conceptualization, Supervision, Writing - original draft. **Peng Xu:** Conceptualization, Funding acquisition. **Anantha Sundaram:** Validation, Writing - original draft, Writing - review & editing. **Richard D. Braatz:** Writing - review & editing.

## References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
- Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5 (2), 157–166. doi:10.1109/72.279181.
- Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. LOF: Identifying density-based local outliers. In: *Proc. 2000 ACM SIGMOD. Association for Computing Machinery*, New York, NY, p. 93104. doi:10.1145/335191.335388.
- Cakir, E., Heittola, T., Huttunen, H., Virtanen, T., 2015. Polyphonic sound event detection using multi label deep neural networks. In: 2015 IJCNN, pp. 1–7. doi:10.1109/IJCNN.2015.7280624.
- Chiang, L.H., Braatz, R.D., 2003. Process monitoring using causal map and multivariate statistics: fault detection and identification. *Chemom. Intell. Lab. Syst.* 65 (2), 159–178. doi:10.1016/S0169-7439(02)00140-5.
- Chiang, L.H., Kotanchek, M.E., Kordon, A.K., 2004. Fault diagnosis based on Fisher discriminant analysis and support vector machines. *Comput. Chem. Eng.* 28 (8), 1389–1401. doi:10.1016/j.compchemeng.2003.10.002.
- Chiang, L.H., Russell, E.L., Braatz, R.D., 2000. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemom. Intell. Lab. Syst.* 50 (2), 243–252. doi:10.1016/S0169-7439(99)00061-1.
- Chiang, L. H., Russell, E. L., Braatz, R. D., 2000b. Tennessee Eastman problem simulation data. <https://web.mit.edu/braatzgroup/links.html>.
- Chine, W., Mellit, A., Lughii, V., Malek, A., Sulligoi, G., Pavan, A.M., 2016. A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks. *Renew. Energy* 90, 501–512. doi:10.1016/j.renene.2016.01.036.
- Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: encoder-decoder approaches. In: *Proc. SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Association for Computational Linguistics*, Doha, Qatar, pp. 103–111. doi:10.3115/v1/W14-4012.
- De Ketelaere, B., Hubert, M., Schmitt, E., 2015. Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. *J. Qual. Technol.* 47 (4), 318–335.
- Dong, D., McAvoy, T., 1996. Nonlinear principal component analysis based on principal curves and neural networks. *Comput. Chem. Eng.* 20 (1), 65–78. doi:10.1016/0098-1354(95)00003-K.
- Downs, J., Vogel, E., 1993. A plant-wide industrial process control problem. *Comput. Chem. Eng.* 17 (3), 245–255. doi:10.1016/0098-1354(93)80018-1.
- Duda, R., Hart, P., Stork, D., 2001. *Pattern Classification*. Wiley, New York, NY.
- Elman, J.L., 1990. Finding structure in time. *Cogn. Sci.* 14 (2), 179–211. doi:10.1016/0364-0213(90)90002-E.
- Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7 (2), 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.
- Fortunato, M., Blundell, C., Vinyals, O., 2017. Bayesian recurrent neural networks. arXiv preprint: 1704.02798.
- Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: *Proc. 33rd ICML*, pp. 1050–1059.
- Gal, Y., Ghahramani, Z., 2016. A theoretically grounded application of dropout in recurrent neural networks. In: *Proc. 30th NeurIPS*, pp. 1027–1035.
- Ge, Z., Song, Z., Gao, F., 2013. Review of recent research on data-based process monitoring. *Ind. Eng. Chem. Res.* 52 (10), 3543–3562. doi:10.1021/ie302069q.
- Graves, A., Mohamed, A., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: *ICASSP 2013*, pp. 6645–6649. doi:10.1109/ICASSP.2013.6638947.
- Hernández-Lobato, J.M., Adams, R.P., 2015. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In: *Proc. 32nd ICML*, 37, p. 18611869.
- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.
- Hsieh, W.W., 2006. Nonlinear principal component analysis of noisy data. In: *Proc. IJCNN 2006*, pp. 4582–4586. doi:10.1109/IJCNN.2006.247086.

- Hu, Z., Jiang, P., 2019. An imbalance modified deep neural network with dynamical incremental learning for chemical fault diagnosis. *IEEE Trans. Ind. Electron.* 66 (1), 540–550. doi:[10.1109/TIE.2018.2798633](https://doi.org/10.1109/TIE.2018.2798633).
- Ince, T., Kiranyaz, S., Eren, L., Askar, M., Gabbouj, M., 2016. Real-time motor fault detection by 1-D convolutional neural networks. *IEEE Trans. Ind. Electron.* 63 (11), 7067–7075. doi:[10.1109/TIE.2016.2582729](https://doi.org/10.1109/TIE.2016.2582729).
- Jackson, J.E., Mudholkar, G.S., 1979. Control procedures for residuals associated with principal component analysis. *Technometrics* 21 (3), 341–349. doi:[10.2307/1267757](https://doi.org/10.2307/1267757).
- Jia, F., Lei, Y., Lin, J., Zhou, X., Lu, N., 2016. Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data. *Mech. Syst. Signal Process.* 72–73, 303–315. doi:[10.1016/j.ymssp.2015.10.025](https://doi.org/10.1016/j.ymssp.2015.10.025).
- Jolliffe, I., 2011. *Principal Component Analysis*. Springer, New York, NY.
- Jordan, M.I., 1997. Chapter 25 - Serial Order: A Parallel Distributed Processing Approach. In: Donahoe, J.W., Dorsel, V.P. (Eds.), *Neural-Network Models of Cognition*, 121. North-Holland, pp. 471–495. doi:[10.1016/S0166-4115\(97\)80111-2](https://doi.org/10.1016/S0166-4115(97)80111-2).
- Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y., 2016. Exploring the limits of language modeling. arXiv preprint: [1602.02410](https://arxiv.org/abs/1602.02410).
- Kourti, T., MacGregor, J.F., 1996. Multivariate SPC methods for process and product monitoring. *J. Qual. Technol.* 28 (4), 409–428. doi:[10.1080/00224065.1996.11979699](https://doi.org/10.1080/00224065.1996.11979699).
- Kramer, M.A., 1992. Autoassociative neural networks. *Comput. Chem. Eng.* 16 (4), 313–328.
- Ku, W., Storer, R.H., Georgakis, C., 1995. Disturbance detection and isolation by dynamic principal component analysis. *Chemom. Intell. Lab. Syst.* 30 (1), 179–196. doi:[10.1016/0169-7439\(95\)00076-3](https://doi.org/10.1016/0169-7439(95)00076-3).
- Lee, J.M., Yoo, C., Choi, S.W., Vanrolleghem, P.A., Lee, I.B., 2004. Nonlinear process monitoring using kernel principal component analysis. *Chem. Eng. Sci.* 59 (1), 223–234. doi:[10.1016/j.ces.2003.09.012](https://doi.org/10.1016/j.ces.2003.09.012).
- Lee, K.B., Cheon, S., Kim, C.O., 2017. A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes. *IEEE Trans. Semicond. Manuf.* 30 (2), 135–142. doi:[10.1109/TSM.2017.2676245](https://doi.org/10.1109/TSM.2017.2676245).
- Li, W., Monti, A., Ponci, F., 2014. Fault detection and classification in medium voltage DC shipboard power systems with wavelets and artificial neural networks. *IEEE Instrum. Meas. Mag.* 63 (11), 2651–2665. doi:[10.1109/TIM.2014.2313035](https://doi.org/10.1109/TIM.2014.2313035).
- Louizos, C., Welling, M., 2017. Multiplicative normalizing flows for variational Bayesian neural networks. In: *Proc. 34th ICML*, 70, pp. 2218–2227.
- Malhotra, P., Vig, L., Shroff, G., Agarwal, P., 2015. Long short term memory networks for anomaly detection in time series. In: *Proc. ESANN*, pp. 89–94.
- Merity, S., Xiong, C., Bradbury, J., Socher, R., 2016. Pointer sentinel mixture models. arXiv preprint: [1609.07843](https://arxiv.org/abs/1609.07843).
- Miller, P., Swanson, R.E., Heckler, C.E., 1998. Contribution plots: a missing link in multivariate quality control. *AMCS* 8 (4), 775–792.
- Moustapha, A.I., Selmic, R.R., 2007. Wireless Sensor Network Modeling Using Modified Recurrent Neural Networks: Application to Fault Detection. 2007 IEEE International Conference on Networking, Sensing and Control. IEEE.
- Nie, L., Li, Y., Kong, X., 2018. Spatio-temporal network traffic estimation and anomaly detection based on convolutional neural network in vehicular ad-hoc networks. *IEEE Access* 6, 40168–40176. doi:[10.1109/ACCESS.2018.2854842](https://doi.org/10.1109/ACCESS.2018.2854842).
- Pachitariu, M., Sahani, M., 2013. Regularization and nonlinearities for neural language models: when are they needed? arXiv preprint: [1301.5650](https://arxiv.org/abs/1301.5650).
- Patan, K., Witczak, M., Korbicz, J., 2008. Towards robustness in neural network based fault diagnosis. *Int. J. Appl. Math. Comput.* 18 (4), 443–454.
- Pawlowski, N., Brock, A., Lee, M. C., Rajchl, M., Glocker, B., 2017. Implicit weight uncertainty in neural networks. arXiv preprint: [1711.01297](https://arxiv.org/abs/1711.01297).
- Pham, V., Bluche, T., Kermorvant, C., Louradour, J., 2014. Dropout improves recurrent neural networks for handwriting recognition. In: *Proc. ICFHR*, pp. 285–290.
- Qin, S.J., 2009. Data-driven fault detection and diagnosis for complex industrial processes. *Proc. IFAC* 42 (8), 1115–1125. doi:[10.3182/20090630-4-ES-2003.00184](https://doi.org/10.3182/20090630-4-ES-2003.00184).
- Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. *Nature* 323 (6088), 533–536. doi:[10.1038/323533a0](https://doi.org/10.1038/323533a0).
- Russell, E.L., Chiang, L.H., Braatz, R.D., 2000. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. *Chemom. Intell. Lab. Syst.* 51 (1), 81–93. doi:[10.1016/S0169-7439\(00\)00058-7](https://doi.org/10.1016/S0169-7439(00)00058-7).
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15 (1), 1929–1958.
- Sun, W., 2020. *Advanced Process Data Analytics*. Massachusetts Institute of Technology, Cambridge, MA.
- Talebi, H.A., Khorasani, K., Tafazoli, S., 2009. A recurrent neural-network-based sensor and actuator fault detection and isolation for nonlinear systems with application to the satellite's attitude control subsystem. *IEEE Trans. Neural Netw.* 20 (1), 45–60. doi:[10.1109/TNN.2008.2004373](https://doi.org/10.1109/TNN.2008.2004373).
- Valle, S., Li, W., Qin, S.J., 1999. Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Ind. Eng. Chem. Res.* 38 (11), 4389–4401. doi:[10.1021/ie990110i](https://doi.org/10.1021/ie990110i).
- Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.N., 2003. A review of process fault detection and diagnosis: part i: quantitative model-based methods. *Comput. Chem. Eng.* 27 (3), 293–311. doi:[10.1016/S0098-1354\(02\)00160-6](https://doi.org/10.1016/S0098-1354(02)00160-6).
- Wang, L., Zhang, Z., Long, H., Xu, J., Liu, R., 2017. Wind turbine gearbox failure identification with deep neural networks. *IEEE Trans. Ind. Informat.* 13 (3), 1360–1368. doi:[10.1109/TII.2016.2607179](https://doi.org/10.1109/TII.2016.2607179).
- Werbos, P.J., 1974. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Harvard University, Boston, MA.
- Werbos, P.J., 1990. Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78 (10), 1550–1560. doi:[10.1109/5.58337](https://doi.org/10.1109/5.58337).
- Westerhuis, J.A., Gurden, S.P., Smilde, A.K., 2000. Generalized contribution plots in multivariate statistical process monitoring. *Chemom. Intell. Lab. Syst.* 51 (1), 95–114. doi:[10.1016/S0169-7439\(00\)00062-9](https://doi.org/10.1016/S0169-7439(00)00062-9).
- Wise, B.M., Ricker, N.L., Veltkamp, D.J., 1989. Upset and sensor failure detection in multivariate processes. *AIChE Annual Meeting*.
- Wu, H., Zhao, J., 2018. Deep convolutional neural network model based chemical process fault diagnosis. *Comput. Chem. Eng.* 115, 185–197. doi:[10.1016/j.compchemeng.2018.04.009](https://doi.org/10.1016/j.compchemeng.2018.04.009).
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., 2016. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint: [1609.08144](https://arxiv.org/abs/1609.08144).
- Yin, S., Ding, S.X., Xie, X., Luo, H., 2014. A review on basic data-driven approaches for industrial process monitoring. *IEEE Trans. Ind. Electron.* 61 (11), 6418–6428. doi:[10.1109/TIE.2014.2301773](https://doi.org/10.1109/TIE.2014.2301773).
- Zarei, J., Tajeddini, M.A., Karimi, H.R., 2014. Vibration analysis for bearing fault detection and classification using an intelligent filter. *Mechatronics* 24 (2), 151–157. doi:[10.1016/j.mechatronics.2014.01.003](https://doi.org/10.1016/j.mechatronics.2014.01.003).
- Zhang, Z., Zhao, J., 2017. A deep belief network based fault diagnosis model for complex chemical processes. *Comput. Chem. Eng.* 107, 395–407. doi:[10.1016/j.compchemeng.2017.02.041](https://doi.org/10.1016/j.compchemeng.2017.02.041).
- Zhu, X., Braatz, R.D., 2014. Two-dimensional contribution map for fault identification [focus on education]. *IEEE Contr. Syst. Mag.* 34 (5), 72–77. doi:[10.1109/MCS.2014.2333295](https://doi.org/10.1109/MCS.2014.2333295).