Modeling of Wastewater Treatment Processes Using Dynamic Bayesian Networks Based on Fuzzy PLS

HONGBIN LIU¹,², HAO ZHANG¹, YUCHEN ZHANG¹, FENGSHAN ZHANG², AND MINGZHI HUANG³

¹Co-Innovation Center of Efficient Processing and Utilization of Forest Resources, Nanjing Forestry University, Nanjing 210037, China
²Laboratory for Comprehensive Utilization of Paper Waste of Shandong Province, Shandong Huatai Paper Company Ltd., Dongying 257335, China
³SCNU Environmental Research Institute, Guangdong Provincial Key Laboratory of Chemical Pollution and Environmental Safety and MOE Key Laboratory of Theoretical Chemistry of Environment, School of Environment, South China Normal University, Guangzhou 510006, China

Corresponding authors: Fengshan Zhang (htjszx@163.com) and Mingzhi Huang (mingzhi.huang@m.scnu.edu.cn)

ABSTRACT The complicated characteristics of wastewater treatment plants (WWTPs) significantly hinder the monitoring of industrial processes, and thus much attention has been paid to process modeling and prediction. A fuzzy partial least squares-based dynamic Bayesian networks (FPLS-DBN) is proposed to improve the modeling ability in WWTPs. To adapt the nonlinear process data, fuzzy partial least squares (FPLS) is introduced by using a fuzzy system to extract nonlinear features from process data. In addition, a dynamic extension is included by embedding augmented matrices into Bayesian networks to fit the uncertainty and time-varying characteristics. Regarding the quality indices for effluent suspended solid in the WWTP, the root mean square error of the FPLS-DBN model is decreased by 28.63% and 69.47%, respectively, in comparison with that for partial least squares and Bayesian networks. The results demonstrate the superiority of FPLS-DBN in modeling performance for an actual industrial WWTP application.

INDEX TERMS Fuzzy partial least squares, latent variables, Bayesian networks, dynamic process modeling, wastewater treatment processes.

I. INTRODUCTION

In recent years, more attention is focused on wastewater treatment processes (WWTPs) and stricter requirements for wastewater discharge standards have been prompted to ensure the accurate monitoring of key quality indices. The effluent quality indices are often measured by online instruments or off-line laboratory analysis. However, these methods are difficult to measure the quality indices for real-time monitoring owing to continuous maintenance, insufficient accuracy, measurement delay, and aged deterioration [1], [2]. To improve the monitoring in WWTPs, it is necessary to structure an accurate process model, especially the real-time monitoring for the key process effluent indices [2].

To quickly and accurately obtain key variables, data-driven models have been widely concerned in many engineering fields, which could extract valid information from easy-to-measure auxiliary variables to predict difficult-to-measure key variables online [3]–[5]. Therefore, auxiliary variables based on limited data are more critical for modeling. Too many inputs, however, may lead to overfitting and inefficiency, and thus it is necessary to provide auxiliary variables that have lower dimensions for modeling.

As an essential part of multivariate statistical methods, the latent variable methods (LVMs) are noticed in many fields in recent years [6]. Common LVMs include principal component analysis (PCA) and partial least squares (PLS). The PLS differs from PCA in that PLS is not only searching for the direction of the largest variance in \( \mathbf{X} \), but also finding the direction of the largest covariance in \( \mathbf{X} \) and \( \mathbf{Y} \) space. Therefore, PLS can better describe the relationship between input and output variables.

Although conventional PLS latent variable methods have been widely used in modeling, some of the industrial processes often contain the variable with nonlinear features which could not be extracted easily, leading to the linear model may not perform well when facing complex industrial data. To solve the aforementioned problems, it is therefore...
necessary to try to improve the model performance by nonlinear extensions of PLS. The quadratic PLS (QPLS) was developed by Wold et al. [7] who applied quadratic polynomial to PLS. In addition, the fuzzy PLS (FPLS) model, replacing the linear relation between input and output with a fuzzy system, was proposed and its regression performance was better than the PLS and QPLS [8]. Another method of robust adaptive partial least squares was applied to process monitoring of a complex industrial wastewater treatment process [9]. Also, the kernel method was used to structure the kernel-based PLS model to effectively capture the nonlinear relationship of process variables [10]. And the least squares vector machine was also applied to build least squares vector machine-PLS (LSSVM-PLS) [11]. Recently, the dynamic nonlinear PLS was developed using Gaussian process regression and achieved better performance compared to PLS, QPLS, and LSSVM-PLS [12].

In practice, the time-varying characteristic often exists among the process data collected in the form of a series, which leads to great difficulties for static process modeling [13], [14]. To overcome the problems of time-varying due to the complexity of processes, dynamic modeling techniques including recursive algorithms [15], time differences [16], augmented matrices [12], and auto-regressive [14] are preferred choices. In recent years, the dynamic Bayesian networks was structured by recursive algorithms to predict the bridge condition [17]. Dong and Qin [14] built the dynamic-inner principal component analysis by maximizing auto-covariance. Among these dynamic modeling techniques, the augmented matrix method as a simple but effective method is widely implemented in the dynamic model. Based on conventional modeling methods, the augmented matrix is embedded into PCA and ICA, respectively to structure the dynamic model for process fault detection [17]. Dong and Qin [14] built the dynamic-inner principal component analysis by maximizing auto-covariance. Among these dynamic modeling techniques, the augmented matrix method as a simple but effective method is widely implemented in the dynamic model. Based on conventional modeling methods, the augmented matrix is embedded into PCA and ICA, respectively to structure the dynamic model for process fault detection [17].

In this work, the modeling performance can be further improved with the combination of dynamic probability and LVMs method. Firstly, FPLS is proposed to extract the nonlinear characteristic from the lower dimensions of latent variables. Then, the augmented matrices technique is embedded into the BN-based FPLS to obtain an FPLS-dynamic Bayesian networks (FPLS-DBN) model. The goal of this model is to achieve a better prediction performance when facing uncertainty, nonlinearity, and time-varying characteristics of the process. The structure of this paper is as follows. In Section II, we provide an introduction to the FPLS-DBN method in detail. In Section III, the wastewater treatment process and practical industrial datasets from WWTPs are introduced and the proposed model is applied to WWTPs to evaluate the modeling performance of FPLS-DBN. To make a fair comparison, the models of corresponding counterparts are conducted for process modeling. Also, the results are analyzed and discussed. Finally, the main conclusions are presented in Section IV.

II. MATERIALS AND METHODS

In this section, conventional methods are introduced and the details of FPLS-DBN are presented. Based on the PLS model, the FPLS model is structured by introducing the Takagi-Sugeno-Kang (TSK) and fuzzy c-means (FCM) methods. Also, the DBN model makes it possible to use both probabilistic methods and dynamic techniques. Then, the proposed method FPLS-DBN can be obtained with the combination of dynamic probability model and nonlinear method.

A. PARTIAL LEAST SQUARES

The partial least squares aims at searching for suitable latent variables that can exploit variance structures of variables. The input variables $\mathbf{X}$ and the output variables $\mathbf{Y}$ are projected by a PLS model, which takes the maximization of covariance between $\mathbf{X}$ and $\mathbf{Y}$, extracting the orthogonal feature vectors in the space of $\mathbf{X}$ and $\mathbf{Y}$. Therefore, PLS method is widely used in process modeling and dimensionality reduction.

B. FUZZY PARTIAL LEAST SQUARES

The PLS can better extract effective information when solving linear problems, however, it has some limitations for complex nonlinear processes. Therefore, the fuzzy rule and clustering method are used to decompose the complex non-linear system into multiple local linear models, which can effectively overcome the nonlinear problems [8].

1) FUZZY C-MEANS

Clustering is one of the most important methods for data analysis and modeling. Among the clustering algorithms, fuzzy c-means, as a fuzzy clustering method based on the objective function, is the most widely method used in data modeling.

The FCM algorithm can be formulated as follows:

(1) Select the number of clusters $L$ and the fuzzy index $m$ (a larger $m$ indicates a fuzzy membership relationship, and vice versa);

(2) Calculate the membership:

$$
\mu_{ij} = \left( \frac{|x_j - c_i|}{|x_j - c_k|} \right)^{2(m-1)} 
$$

(1)
where \( x_j (j = 1, 2, \cdots, N) \) is sample data; \( c_i (i = 1, 2, \cdots, L) \) is the cluster center.

3) Calculate the cluster center:

\[
c_i = \sum_{j=1}^{N} \mu_{ij}^m x_j / \sum_{j=1}^{N} \mu_{ij}^m
\]  

(2)

4) Structure the objective function of fuzzy clustering \( J_G \):

\[
J_G = \sum_{i=1}^{L} \sum_{j=1}^{N} \mu_{ij}^m \| x_j - c_i \|^2
\]  

(3)

The FCM method is the process of minimizing \( J_G \), using the objective function to optimize the clustering problem.

2) TSK FUZZY INFERENCE SYSTEM

The fuzzy system based on fuzzy reasoning is a method of analysis, reasoning and decision-making process, whose fuzzy language would be transformed into fuzzy rules. Therefore, a mathematical model corresponding to the process can be structured because of its better interpretability and learning ability. Classical fuzzy systems can be divided into three types [28]: Takagi-Sugeno-Kang (TSK), Mamdani-Larsen (ML), and Generalized fuzzy (GF). As an essential part of fuzzy systems with more effectiveness and flexibility, TSK is specifically suitable for dealing with complex nonlinear systems [29]. The IF-THEN rule of TSK whose form is defined as:

If \( x_1 \) is \( A_{1i} \), \( \cdots \), \( x_r \) is \( A_{ri} \),
then \( y_i = b_{i0} + b_{i1} x_1 + \cdots + b_{ir} x_r \) \( (i = 1, 2, \cdots, L) \)  

(4)

where \( L \) is the number of fuzzy rules; \( x \) and \( y \) are input variables and local output variables while \( b_0 \) are input variables, respectively; \( b_{ij} = (b_{i0} b_{i1} \cdots b_{ir})^T \) are the real-valued parameters; \( A_{ij} \) is the fuzzy set.

The overall output of the TSK model is:

\[
y = \frac{\sum_{i=1}^{L} \tau_i \mu(x_1) \times A_1(x_1) \times \cdots \times A_r(x_r)}{\sum_{i=1}^{L} \tau_i}
\]  

(5)

where, \( \tau_i \) is the firing strength of \( i \)-th fuzzy rule, which is computed by \( \tau_i = A_{1i}(x_1) \times A_{2i}(x_2) \times \cdots \times A_{ri}(x_r) \).

The Gaussian membership function is defined as:

\[
A_{ri}(x_r) = \exp(-\frac{(x_r - c_{ri})^2}{2\sigma_i^2}), \quad i = 1, 2, 3, \cdots, L
\]  

(6)

where, \( c_{ri} \) is the clustering center of Gaussian membership function, \( \sigma_i \) is the center width of the membership function.

C. DYNAMIC BAYESIAN NETWORKS

The Bayesian networks is a probabilistic model based on a directed acyclic graph (DAG), illustrating the relationship among process variables and each variable represents a node in the networks. The details of the Bayesian networks model can be found in the literature [30]. The target of Bayesian networks is to fit the nonlinear and uncertain characteristics in complex modeling. Assume that \( X = \{x_1, x_2, \cdots, x_m\} \), the Bayesian formula could be given by:

\[
P(\xi | x_1, x_2, x_3, \cdots, x_m) = \frac{\pi(\xi) P(x_1, x_2, x_3, \cdots, x_m | \xi)}{P(x_1, x_2, x_3, \cdots, x_m)}
\]  

(7)

where \( \pi(\xi) \) is the prior distribution of \( \xi \), \( P(x_1, x_2, x_3, \cdots, x_m | \xi) \) is a conditional probability.

The Bayesian networks model as a DAG model fully describes the relationship between different variables. However, there still may be limitations for the BN model with time-varying characteristics of the data. Therefore, an augmented matrix technique is embedded in the BN to fit the dynamic characteristics. The construction of the dynamic model is as follows:

The original input matrix \( X \) (the parent nodes in the Bayesian network structure) is defined as:

\[
X = \begin{bmatrix}
x_1(1) & x_2(1) & \cdots & x_m(1) \\
x_1(2) & x_2(2) & \cdots & x_m(2) \\
\vdots & \vdots & \ddots & \vdots \\
x_1(n) & x_2(n) & \cdots & x_m(n)
\end{bmatrix}_{n \times m}
\]  

(8)

Based on the input matrix \( X \), the augmented matrix \( X_i \) is obtained as:

\[
X_i = \begin{bmatrix}
x_i(t) & x_i(t-1) & \cdots & x_i(t-d+1) \\
x_i(t+1) & x_i(t) & \cdots & x_i(t-d+2) \\
\vdots & \vdots & \ddots & \vdots \\
x_i(t+n-1) & x_i(t+n-2) & \cdots & x_i(t-d+n)
\end{bmatrix}_{(i = 1, 2, \cdots, m)}
\]  

(9)

The dynamic technique more clearly presents the structure of data with time-varying by expanding the matrix and introducing a time lag coefficient. Therefore, dynamic characteristics are also more easily extracted in modeling. The DBN model obtains a further development of the dynamic probability model, which improves the ability of the DBN to interpret the data from dynamic processes.

D. STRUCTURE OF FPLS-DBN MODEL

Based on the PLS model, the FPLS model is structured by introducing the TSK and FCM methods to preferably extract the nonlinear characteristics of process data. Moreover, the DBN model makes it possible to use both probabilistic methods and dynamic techniques, which could solve the problems of uncertainty and dynamic characteristics. Then, the latent variables (LVs) of FPLS update the original inputs used for the DBN model. Through combining the FPLS with the DBN, the FPLS-DBN is more conducive to practical applications and achieves a further development of dynamic probability models based on LVMs. As shown in Fig. 1,
the FPLS-DBN model is structured according to the research above to provide an accurate modeling method for WWTPs. The PLS is used to decompose the input \(X\) and output \(Y\) as follows:

\[
\begin{align*}
X &= T^TP + E = t_1p_1^T + t_2p_2^T + \cdots + t_hp_h^T \\
Y &= UQ + F = u_1q_1^T + u_2q_2^T + \cdots + u_qq_q^T
\end{align*}
\] (10)

where \(T\) and \(U\) are score matrices for \(X\) and \(Y\); \(P\) and \(Q\) are loading matrices, and \(E\) and \(F\) are residual matrices, respectively; and \(t\) and \(u\) are score vectors for \(X\) and \(Y\), and \(p\) and \(q\) are loading matrices, respectively.

Calculate eigenvectors \(t_h\) and \(u_h\) for each factor \(h\):

\[
\begin{align*}
& w_h^T = u_h^T E_h^{-1}/u_h^T u_h \\
& w_h = w_h/\|w_h\| \\
& t_h = E_h^{-1} w_h \\
& c_h^T = t_h^T F_h^{-1}/t_h^T t_h \\
& c_h = c_h/\|c_h\| \\
& u_h = F_h^{-1} c_h
\end{align*}
\] (11-16)

Obtain the clustering center of Gaussian membership function:

\[
\begin{align*}
& c_i = \sum_{j=1}^{N} \mu_{ji}^2 / \sum_{j=1}^{N} \mu_{ji}^2 \quad (17) \\
& \mu_{ij} = 1/\sum_{k=1}^{L} \left( \frac{|t_j - c_i|}{|t_j - c_k|} \right)^m \quad (18)
\end{align*}
\]

where \(c_i (i = 1, 2, \cdots, L)\) is the cluster center, \(\mu_{ij}\) is membership grade; the fuzzy index \(m\) is taken as 2 in this study. The dataset is clustered into \(L\) categories, and sub-models are built for each category.

Define the input variable as \(x = [x_1 x_2 \cdots x_r]^T\), model parameters are \(b_i = [b_{i0} b_{i1} \cdots b_{ir}]^T\). The TSK fuzzy function is defined as:

\[
f_h(t) = \sum_{i=1}^{L} G_i (b_{i0} + b_{i1} t) \quad (19)
\]

where \(G_i\) is the normalized firing strength.
The normalized firing strength $G_i$ and Gaussian-type firing strength for $i$-th rule $\tau_i$ are calculated as:

$$G_i = \frac{\tau_i}{\sum_{i=1}^{L} \tau_i}$$  \hspace{1cm} (20)

$$\tau_i = \exp\left(-\frac{(t - c_i)^2}{2\sigma_i^2}\right)$$ \hspace{1cm} (21)

where $\sigma_i$ is the width of the membership function; $i = 1, 2, \ldots, L$.

The width of the membership function is calculated using the nearest neighbor rule:

$$\sigma_i = \left[\frac{1}{n} \sum_{i=1}^{n} (c_i - c)^2\right]^{\frac{1}{2}}$$ \hspace{1cm} (22)

where $c_i$ and $c$ are the $n$ closest cluster centers, respectively, and $n$ is often taken as 2; $l = 1, 2, \ldots, n$.

The overall output of TSK model is calculated by:

$$\hat{u} = \frac{\sum_{i=1}^{L} \tau_i y_i}{\sum_{i=1}^{L} \tau_i} = \frac{\sum_{i=1}^{L} (b_{i0} + b_{i1} x_1 + \cdots + b_{in} x_n)}{\sum_{i=1}^{L} \tau_i}$$ \hspace{1cm} (23)

Minimize the objective function $J_G$:

$$J_G = \sum_{k=1}^{N} [u(k) - \hat{u}(k)]^2$$ \hspace{1cm} (24)

Calculate the loading vectors of $X$ and $Y$:

$$p_{h}^{T} = t_{h}^{T} E_{h-1}^{-1} / \left(t_{h}^{T} t_{h}\right)$$ \hspace{1cm} (25)

$$q_{h}^{T} = \hat{u}_{h}^{T} F_{h-1}^{-1} / \left(\hat{u}_{h}^{T} \hat{u}_{h}\right)$$ \hspace{1cm} (26)

Calculate the residual matrices $E_{h}$ and $F_{h}$:

$$E_{h} = E_{h-1} - t_{h} p_{h}^{T}$$ \hspace{1cm} (27)

$$F_{h} = F_{h-1} - \hat{u}_{h} q_{h}^{T}$$ \hspace{1cm} (28)

Let $h = h + 1$ and perform the iterative calculation until the valid information is extracted from residual matrices $E_{h}$ and $F_{h}$, and then the calculation is terminated.

The number of latent variables in the FPLS model is selected according to the cumulative variance, and a dynamic Bayesian network is structured by augmented matrices techniques. Then the score matrices of FPLS update the original inputs for the dynamic probability method of DBN.

Assume that the score matrix extracted using the FPLS latent variable model is $X_{T}$:

$$X_{T} = \begin{bmatrix} x_{1}(1) & x_{2}(1) & \cdots & x_{m}(1) \\ x_{1}(2) & x_{2}(2) & \cdots & x_{m}(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_{1}(n) & x_{2}(n) & \cdots & x_{m}(n) \end{bmatrix}_{n \times m}$$ \hspace{1cm} (29)

Expand the scoring matrix $X_{T}$ and introduce the time lag $d$ to obtain the matrix $\tilde{X}_{T}$:

$$\tilde{X}_{T} = \begin{bmatrix} x_{1}(t) & x_{2}(t) & \cdots & x_{m}(t) \\ x_{1}(t+1) & x_{2}(t+1) & \cdots & x_{m}(t+1) \\ \vdots & \vdots & \ddots & \vdots \\ x_{1}(t+n) & x_{2}(t+n) & \cdots & x_{m}(t+n) \end{bmatrix}_{(i = 1, 2, \ldots, m)}$$ \hspace{1cm} (30)

where $x(t)$ is sample data; $d$ is the time lag.

Use $\tilde{X}_{T}$ as nodes for BN, and calculate the conditional density of the sample $x_1, x_2, x_3, \ldots, x_m$ for a random variable $\xi$; then the prior distribution $\pi(\xi)$ and conditional density $P(x_1, x_2, x_3, \ldots, x_m | \xi)$ are used to calculate the posterior probability $P(\xi | x_1, x_2, x_3, \ldots, x_m)$ by the Bayes formula, which is defined as follows:

$$P(\xi | x_1, x_2, x_3, \ldots, x_m) = \frac{\pi(\xi)P(x_1, x_2, x_3, \ldots, x_m | \xi)}{P(x_1, x_2, x_3, \ldots, x_m)}$$ \hspace{1cm} (31)

The testing set variable $\xi$ is inferred from the posterior probability density. Besides, root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination ($R^2$) are adopted to evaluate model performance. The statistical indices are defined by:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{N}}$$ \hspace{1cm} (32)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$ \hspace{1cm} (33)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$ \hspace{1cm} (34)

where $y_i$ is real value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean value, and $N$ is the number of samples.

III. RESULTS AND DISCUSSION

Given the fact that the discharge of hazardous pollutants from WWTPs poses a significant challenge for the environment, it is necessary to structure an accurate and reliable model for process monitoring. In this section, the data for testing the modeling performance were collected from WWTPs. To illustrate the performance of the FPLS-DBN, the key quality index SS is used for prediction. Besides, five counterparts, consisting of PLS, BN, PLS-BN, PLS-DBN, and FPLS-BN are considered for comparison.

A. WWTP DATA

The process data were collected from Daewoo nutrient removal wastewater treatment process from March 9, 2007, to February 29, 2008 [31]. As shown in Fig. 2, the wastewater treatment process mainly includes biological reactors, primary and secondary clarifiers, concentration tank, and dewatering system. The bioreactors include denitrification, anaerobic, anoxic, and aerobic reactions. As shown in Fig. 3,
the influent variables include flow rate ($Q$), suspended solid ($SS_{in}$), biochemical oxygen demand ($BOD_{in}$), chemical oxygen demand ($COD_{in}$), total nitrogen ($TN_{in}$), and total phosphorus ($TP_{in}$). The effluent variable is suspended solid ($SS_{eff}$). The data consist of 358 groups, of which the first 238 groups were used as the training set and the last 120 groups were used as the testing set. Table 1 provides a statistical description of WWTP data.

### B. RESULTS AND DISCUSSION

The PLS model as a linear method uses linear inner relation in latent space, whose internal relationship model of the latent variable is shown in Fig. 4 ($t(1)$ and $u(1)$ are input and output score vectors of the latent variable, respectively, and regression models are represented by the solid lines). As shown in Tables 2 to 4, “This LV” represents the variance (%), “Total” represents the cumulative variance (%), and the LVs is selected by cumulative variance. According

| Variable | Mean | Standard Deviation | Unit |
|----------|------|--------------------|------|
| $Q$      | 8819.59 | 1006.37           | m$^3$/d |
| $SS_{in}$ | 98.81   | 24.06             | g SS/m$^3$ |
| $BOD_{in}$ | 103.96 | 30.38           | g O$_2$/m$^3$ |
| $COD_{in}$ | 64.84  | 18.93            | g COD/m$^3$ |
| $TN_{in}$ | 25.68   | 6.29               | g N/m$^3$ |
| $TP_{in}$ | 2.68    | 0.75              | g P/m$^3$ |
| $SS_{eff}$ | 1.87    | 0.91               | g SS/m$^3$ |
To Table 2, there is no obvious increase in the cumulative variance of output after the second LVs is reached, and thus we selected second LVs for PLS modeling.

The scattered points of the latent variable, as shown in Fig. 4, indicate that the process data has strong nonlinearity, it is therefore difficult to effectively extract data information. More specifically, in terms of Table 2, the variance by the second LVs is 1.88% for the output data, whereas the variance by the third LVs is just 0.53%. It follows that PLS modeling has large limitations and it is difficult to structure an accurate model for nonlinear processes.

To overcome the shortage of PLS in modeling, the FPLS model is structured based on the TSK and FCM methods to provide a nonlinear model. The fuzzy rules are critical to the model performance of the FPLS model. The different fuzzy rules thus are used for modeling to explore the impact of fuzzy rules on different models.

The FPLS regression model of the latent variable is shown in Fig. 5, which is composed of two subplots. The upper score plot consists of \( t(1) \) and \( u(1) \) which are the input and output score vectors of the latent variable, and the lower plot is composed of \( u(1) \) and firing strength (the dashed lines show the normalized firing strength \( G_i \), and solid lines represent the firing strength \( \tau_i \)).

In addition, Figs. 5(a) - (d) correspond to the score plot with fuzzy rules from 2 to 5, respectively. As shown in Table 3, the variance captured by different FPLS model is represented. And according to the cumulative variance of models, the LVs of FPLS_1 is 2, and the LVs of FPLS_2 to FPLS_4 is 3 in Table 3.

It can be seen from Fig. 5 that obvious nonlinearity is shown in industrial data, and regression lines with different fuzzy rules have different tracking for score points. In Fig. 5 (a), when the fuzzy rule is taken 2, the regression line of FPLS_1 can better fit the score points compared with the internal regression line of PLS (shown in Fig. 4). Similarly, as the number of fuzzy rules increases, the tracking for score points by regression lines becomes more accurate and complicated. Besides, it is concluded from Table 3 that the cumulative variance of the FPLS model is significantly higher than PLS, and the cumulative variance of FPLS is also raised with the increase of fuzzy rules.

However, it is worth noting that in the FPLS_3 model, the variance of the third LVs (1.67%) is lower than the third LVs of the FPLS_2 model (2.15%). It can be understood that effective information extraction is sufficient when the fuzzy rule is 4. Although more variance can be extracted in the FPLS_4 model, there may be redundancy in the extracted information, which is not conducive to modeling.

As shown in Table 3, the variance captured by different models is given (dashed line and solid line represent the static models and dynamic models, respectively). The fuzzy rule on the abscissa is 1 representing the PLS-BN or PLS-DBN models, and 2-5 representing the FPLS-BN or FPLS-DBN models with fuzzy rules of 2, 3, 4, and 5, respectively.

According to Fig. 6, the models based on the fuzzy method (FPLS-BN and FPLS-DBN) are better than the models based on the linear method (PLS-BN and PLS-DBN). Moreover, FPLS-BN and FPLS-DBN models achieve optimal model performance with a fuzzy rule of 4 (the number of the fuzzy rule is 3). However, there is a significant difference in RMSE value between FPLS-BN and FPLS-DBN. The reason is that time-varying characteristic usually exists in actual industrial processes, which greatly hinders modeling performance.

Considering the fact that the FPLS-BN has not taken the time-varying characteristic into consideration, it is necessary to try to improve the model performance by using the dynamic technique. Therefore, the FPLS-DBN shows an obvious improvement of model accuracy after the dynamic technique is applied to the static model. The prediction results of RMSE are shown in Fig. 6. In terms of FPLS-DBN model, the RMSE value of FPLS-DBN (RMSE = 0.7175) is decreased by 56.47% compared with FPLS-BN.

According to the results, the proposed dynamic model can obtain a better model performance when the fuzzy rule is 4. Moreover, the number of LVs for fuzzy rules may also lead to different (even worse) model performance. To explore the effectiveness of models with different number of fuzzy rules (the fuzzy rule is taken as 4), a different number of fuzzy rules are used to obtain the FPLS-DBN and then the 92135
The cumulative variance of FPLS is shown in Table 4 (the LVs of FPLS_5 to FPLS_9 are 2, 3, 4, 5, and 5, respectively). Also, the cumulative variance and the number of LVs for FPLS models are shown in Fig. 7 (the bar graph displays the cumulative variance (%), and the complexity of the model is assessed through the number of LVs which is represented by...
As shown in Table 4, there is no significant increase in terms of the cumulative variance after the second LVs for FPLS_5. Compared with the other FPLS models with the second LVs (FPLS_6 to FPLS_9), the cumulative variance values of FPLS_5 (17.22% for output data) seem to be similar to those of other FPLS models. However, there are some differences in variance between FPLS models with third LVs. The cumulative variance of FPLS_5 is lower than other FPLS models (FPLS_6 to FPLS_9) which have shown a better ability to extract the data information.

It is worth noting that a greater amount of fuzzy rules has a higher cumulative variance for the model, but inevitably leads to higher dimensions of latent variables, which is not conducive to modeling. The reason is that the increase in the number of LVs leads to more complex models, which may reduce model accuracy and raise the costs of modeling. This explains why more fuzzy rules result in the worse performance of the FPLS-DBN models. Table 5 provides the RMSE value of FPLS-DBN with different numbers of fuzzy rules, which clearly shows that the FPLS-DBN achieves a better model performance when the number of fuzzy rules is 3. The main reason for this result is that the FPLS-DBN model can effectively extract nonlinear data information by the appropriate fuzzy parameters (the fuzzy rule is 4, and the number of fuzzy rules is 3), and avoid the decrease in model accuracy caused by high data dimensions. Moreover, the dynamic probability method and Bayesian networks is combined with the FPLS latent variable model, leading to an improvement in the model performance.
To evaluate the effectiveness of the proposed method, several conventional methods are used to compare modeling performance. For the detailed performance of different models, the prediction results are shown in Table 6 and the detailed prediction variations are shown in Fig. 8. In terms of the conventional models, PLS shows a better model performance than BN. This is probably because the BN model requires a significant amount of data in modeling. Besides, part of the issue is that the BN model may not meet the complexity of industrial processes. After using the latent-variables method of PLS, the PLS-BN shows an obvious improvement than BN. Nevertheless, the prediction results of PLS-BN are even worse compared with PLS. The reason is that in combination with the probabilistic method, process variables are more fully described by the model. When used in this way, however, the problem of data nonlinearity will become more complex, which is not conducive to modeling. For the nonlinear characteristic of process data, a fuzzy-based method is used to extract the nonlinear information. In terms of the fuzzy-based method, the RMSE value of FPLS-BN is decreased by 6.98% compared with PLS-BN.

Considering the fact that the static model unavoidably has some limitations in confronting the dynamic process. It, therefore, is necessary to fit the time-varying characteristic by dynamic technique. Compared with the static methods of BN, PLS, PLS-BN, and FPLS-BN, the RMSE value of FPLS-DBN is decreased by 69.47%, 28.63%, 59.50%, and 56.47%, respectively, and the MAE value of FPLS-DBN is decreased by 45.19%, 30.24%, 36.36%, and 33.88%, respectively. In addition, compared with the prediction results of PLS-DBN, the RMSE of FPLS-DBN is decreased by 3.52% and MAE is reduced by 5.21%, which is mainly due to the improvement of nonlinear modeling capacity. In recent years, Liu et al. [27] developed a locally weighted regression method based on BN (LW-BN) to predict butane content, and its RMSE is reduced by 21.35% compared with LW-PLS. Liu et al. [12] applied the dynamic Gaussian process regression based partial least squares method to predict the effluent ammonia concentration in BSM1 and the results showed that the RMSE of D-GPR-PLS is reduced by 16.80% compared to PLS method. The RMSE value of the proposed method FPLS-DBN in this study is decreased by 28.63% compared with the PLS, which illustrates that the fuzzy LVMs based on the dynamic probability model is beneficial to improve the modeling accuracy in industrial applications.

IV. CONCLUSIONS

In this paper, a dynamic probability model based on the nonlinear method denoted FPLS-DBN has been developed for industrial process modeling, which aims to improve the prediction accuracy of the key quality indices in WWTPs. The proposed method is structured based on the FPLS latent model framework by TSK and FCM fuzzy techniques, which effectively extracts the non-linear features and significantly decreases the data dimension. Then, the dynamic method of augmented matrices is embedded into the probabilistic method of Bayesian networks. Based on the case study of WWTPs, it was shown that the proposed FPLS-DBN model improves the accuracy of the model and effectively overcomes the nonlinearity, uncertainty, and time-variability in the complex wastewater treatment processes. The results show that the RMSE of FPLS-DBN is 0.7175, its RMSE is reduced by 28.63% and 69.47%, respectively, compared with the PLS and BN models, which effectively improves the prediction accuracy of the conventional model and enhances the modeling ability of wastewater treatment processes.
The modeling method from this study can be used to structure the process model online, providing an accurate and reliable modeling method. Besides, the simple but effective dynamic technique and nonlinear method are beneficial to practical application. However, there is still much space for improvement in terms of accuracy. Future studies will pay more attention to maximize modeling accuracy.

ACKNOWLEDGMENT
(Hongbin Liu and Hao Zhang are contributed equally to this work.)

REFERENCES
[1] H. Liu, C. Yang, M. Huang, and C. Yoo, “Soft sensor modeling of industrial process data using kernel latent variables-based relevance vector machine,” Appl. Soft Comput., vol. 90, May 2020, Art. no. 106149.
[2] H. Haimi, M. Mulas, F. Corona, and R. Vahala, “Data-derived soft-sensors for biological wastewater treatment plants: An overview,” Environ. Model. Softw., vol. 47, pp. 88–107, Sep. 2013.
[3] X. Yuan, Y. Wang, C. Yang, Z. Ge, Z. Song, and W. Gui, “Weighted linear dynamic system for feature representation and soft sensor application in nonlinear dynamic industrial processes,” IEEE Trans. Ind. Electron., vol. 65, no. 2, pp. 1508–1517, Feb. 2018.
[4] C. Shang, F. Yang, D. Huang, and W. Lyu, “Data-driven soft sensor development based on deep learning technique,” J. Process Control, vol. 24, no. 3, pp. 223–233, Mar. 2014.
[5] W. Yan, P. Guo, Y. Tian, and J. Gao, “A framework and modeling method of data-driven soft sensors based on semisupervised Gaussian regression,” Ind. Eng. Chem. Res., vol. 55, no. 27, pp. 7394–7410, Jul. 2016.
[6] H. Yu and F. Khan, “Improved latent variable models for nonlinear and dynamic process monitoring,” Chem. Eng. Sci., vol. 168, pp. 325–338, Aug. 2017.
[7] S. Wold, N. Kettaneh-Wold, and B. Skagerberg, “Nonlinear PLS modeling,” Chemometric Intell. Lab. Syst., vol. 7, nos. 1–2, pp. 53–65, Dec. 1989.
[8] Y. H. Bang, C. K. Yoo, and I.-B. Lee, “Nonlinear PLS modeling with fuzzy inference system,” Chemometric Intell. Lab. Syst., vol. 64, no. 2, pp. 137–155, Nov. 2002.
[9] H. W. Lee, M. W. Lee, and J. M. Park, “Robust adaptive partial least squares modeling of a full-scale industrial wastewater treatment process,” Ind. Eng. Chem. Res., vol. 46, no. 3, pp. 955–964, Jan. 2007.
[10] S. H. Woo, C. O. Jeon, Y.-S. Yun, H. Choi, C.-S. Lee, and D. S. Lee, “Online estimation of key process variables based on kernel partial least squares in an industrial cokes wastewater treatment plant,” J. Hazardous Mater., vol. 161, no. 1, pp. 538–544, Jan. 2009.
[11] Y. Lv, J. Liu, and T. Yang, “Nonlinear PLS integrated with error-based LSSVM and its application to NOx modeling,” Ind. Eng. Chem. Res., vol. 51, no. 49, pp. 16902–16100, Dec. 2012.
[12] H. Liu, C. Yang, B. Carlsson, S. J. Qin, and C. Yoo, “Dynamic nonlinear partial least squares modeling using Gaussian process regression,” Ind. Eng. Chem. Res., vol. 58, no. 36, pp. 16676–16686, Sep. 2019.
[13] W. Ku, R. H. Storer, and C. Georgakis, “Disturbance detection and isolation by dynamic principal component analysis,” Chemometric Intell. Lab. Syst., vol. 30, no. 1, pp. 179–196, Nov. 1995.
[14] Y. Dong and S. J. Qin, “Dynamic latent variable analytics for process operations and control,” Comput. Chem. Eng., vol. 114, pp. 69–90, Jun. 2018.
[15] B. R. Noack, W. Stankiewicz, M. Morzyński, and P. J. Schmid, “Recursive dynamic mode decomposition of transient and post-transient wake flows,” J. Fluid Mech., vol. 809, pp. 843–872, Dec. 2016.
[16] H. Kaneko and K. Funatsu, “Discussion on time difference models and intervals of time difference for application of soft sensors,” Ind. Eng. Chem. Res., vol. 52, no. 3, pp. 1322–1334, Jan. 2013.
[17] M. I. Rafiq, M. K. Chrysanthopoulos, and S. Sathanathan, “Bridge condition modelling and prediction using dynamic Bayesian belief networks,” Struct. Infrastruct. Eng., vol. 11, no. 1, pp. 38–50, Jan. 2015.
[18] J.-M. Lee, C. Yoo, and I.-B. Lee, “Statistical modeling of dynamic processes based on dynamic independent component analysis,” Chem. Eng. Sci., vol. 59, no. 14, pp. 2995–3006, Jul. 2004.
[19] Q. Liu, Q. Zhu, S. J. Qin, and T. Chai, “Dynamic concurrent kernel CCA for strip-thickness relevant fault diagnosis of continuous annealing processes,” J. Process. Control, vol. 67, pp. 12–22, Jul. 2018.
[20] H. Liu, C. Yang, M. Huang, and C. Yoo, “Multivariate statistical monitoring of subway indoor air quality using dynamic concurrent partial least squares,” Environ. Sci. Pollut. Res., vol. 27, no. 4, pp. 4159–4169, Feb. 2020.
[21] D. Li, H. Wang, and X. Liang, “Bayesian network based approach for diagnosis of modified sequencing batch reactor,” J. Shanghai Jiaotong Univ., Sci., vol. 24, no. 4, pp. 417–429, Aug. 2019.
[22] S. Kathibisepehr, B. Huang, and S. Khare, “Design of inferential sensors in the process industry: A review of Bayesian methods,” J. Process Control, vol. 23, no. 10, pp. 1575–1596, Nov. 2013.
[23] J. Mori, V. Mahalec, and J. Yu, “Identification of probabilistic graphical network model for root-cause diagnosis in industrial processes,” Comput. Chem. Eng., vol. 71, pp. 171–209, Dec. 2014.
[24] B. Cai, H. Liu, and M. Xie, “A real-time fault diagnosis methodology of complex systems using object-oriented Bayesian networks,” Mech. Syst. Signal Process., vol. 80, pp. 31–44, Dec. 2016.
[25] D. Li, H. Z. Yang, and X. F. Liang, “Prediction analysis of a wastewater treatment system using a Bayesian network,” Environ. Model. Softw., vol. 40, pp. 140–150, Feb. 2013.
[26] Y. Liu, Y. Pan, Z. Sun, and D. Huang, “Statistical monitoring of wastewater treatment plants using variational Bayesian PCA,” Ind. Eng. Chem. Res., vol. 53, no. 8, pp. 3272–3282, Feb. 2014.
[27] Z. Liu, Z. Ge, G. Chen, and Z. Song, “Adaptive soft sensors for quality prediction under the framework of Bayesian network,” Control Eng. Pract., vol. 72, pp. 19–28, Mar. 2018.
[28] Z. Deng, L. Cao, Y. Jiang, and S. Wang, “Minimax probability TSK fuzzy system classifier: A more transparent and highly interpretable classification model,” IEEE Trans. Fuzzy Syst., vol. 23, no. 4, pp. 813–826, Aug. 2015.
[29] Y.-Y. Lin, J.-Y. Chang, and C.-T. Lin, “A TSK-type based self-evolving compensatory interval type-2 fuzzy neural network (TSCIT2FNN) and its applications,” IEEE Trans. Ind. Electron., vol. 61, no. 1, pp. 447–459, Jan. 2014.
[30] I. Chaturvedi, E. Ragu, P. Gasitardo, R. Zanino, and E. Cambria, “Bayesian network based extreme learning machine for subjectivity detection,” J. Franklin Inst., vol. 52, no. 3, pp. 1322–1334, Jan. 2015.
[31] H. Liu, M. Huang, and C. Yoo, “A fuzzy neural network-based soft sensor for modeling nutrient removal mechanism in a full-scale wastewater treatment system,” Desalination Water Treatment, vol. 51, nos. 31–33, pp. 6184–6193, Sep. 2013.

HONGBIN LIU received the M.S. degree from the South China University of Technology, Guangzhou, China, in 2009, and the Ph.D. degree from Kyung Hee University, Seoul, South Korea, in 2013. He has been an Associate Professor with Nanjing Forestry University, Nanjing, China, since 2015. He is the author of more than 60 journal articles. His research interests include process modeling, process monitoring, fault detection and diagnosis, and wastewater treatment.

HAO ZHANG received the B.S. degree from Nanjing Forestry University, Nanjing, China, in 2018, where he is currently pursuing M.S. degree in process modeling and monitoring for wastewater treatment processes.
YUCHEN ZHANG received the B.S. degree from Nanjing Forestry University, Nanjing, China, in 2019, where he is currently pursuing M.S. degree. His research interests include process modeling and monitoring.

FENGSHAN ZHANG received the Ph.D. degree from Nanjing Forestry University, Nanjing, China, in 2010. He is currently the Chief Engineer with Shandong Huatai Paper Company Ltd., Dongying, China. His research interests include process modeling and engineering.

MINGZHI HUANG received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2010. He has been a Professor with South China Normal University, China, since 2017. He is the author of more than 100 journal articles. His research interests include process modeling, process monitoring, fault detection and diagnosis, and wastewater treatment.

* * *

YUCHEN ZHANG received the B.S. degree from Nanjing Forestry University, Nanjing, China, in 2019, where he is currently pursuing M.S. degree. His research interests include process modeling and monitoring.

FENGSHAN ZHANG received the Ph.D. degree from Nanjing Forestry University, Nanjing, China, in 2010. He is currently the Chief Engineer with Shandong Huatai Paper Company Ltd., Dongying, China. His research interests include process modeling and engineering.

MINGZHI HUANG received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2010. He has been a Professor with South China Normal University, China, since 2017. He is the author of more than 100 journal articles. His research interests include process modeling, process monitoring, fault detection and diagnosis, and wastewater treatment.

* * *