Chemometrics and Intelligent Laboratory Systems 183 (2018) 11–22

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Active learning for modeling and prediction of dynamical fluid processes

Hongying Deng [a,b], Yi Liu [a,*], Ping Li [b], Shengchang Zhang [a]

[a] *Institute of Process Equipment and Control Engineering, Zhejiang University of Technology, Hangzhou, 310023, China*
[b] *State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Zhejiang University, Hangzhou, 310027, China*

ARTICLE INFO

Keywords:
Active learning
Probabilistic modeling
Gaussian process regression
Multiphase pump
Flow rate

ABSTRACT

Accurate prediction of the flow rate curve of a stroke for reciprocating multiphase pumps often encounters several challenges in practice, including process nonlinearity, dynamical characteristics, and changing multiphase transportation conditions. To enhance the prediction performance, an active learning method is proposed to efficiently design informative training data. Some initial training data are first collected from experiments to construct several local Gaussian process regression (GPR) models. Additionally, with the GPR-based probabilistic information, a relative variance-based criterion is proposed to explore which regions the new data should be introduced into the GPR prediction model. Moreover, an evaluation criterion is designed to implement the active learning procedure efficiently. Consequently, without time-consuming experiments, a set of new representative training data are sequentially introduced into the GPR prediction model. Experimental results and comparative studies for dynamical flow rate prediction of a stroke are carried out to demonstrate the effectiveness of the proposed method.

1. Introduction

Reciprocating oil-gas multiphase pumps can efficiently increase oil and gas productions in the crude oil drilling for their good internal compression and anti-gas resistance performance [1–4]. The pump completes the flow configuration of a stroke mainly by the piston, suction and discharge valves. Generally, the instantaneous multiphase migration flows are generated for different proportion of gas, water, and coal tars mixture in the transporting crude oil. They will cause the irregular motions of the suction and discharge valves, and the irregular changes of compression ratios [4,5]. Consequently, with the change of crank angle, the flow rate curve of a pump cavity in a stroke consists of opening lag, suction, closing lag, and discharge stages. These stages exhibit distinct characteristics for different flow patterns. Remarkably, the transitions between two stages, related to the opening and closing moments of the suction valve and discharge valve, exhibit more nonlinear and time-varying properties. These factors will aggravate flow pulsations, vibration, and even reduce pump efficiency [4–6]. To help the designers optimize the structure and improve the reliability of multiphase pumps, it is necessary to predict the flow rates of a stroke in different multiphase transportation conditions.

Traditionally, mechanism models such as computational fluid dynamics (CFD) models can provide useful information and thus have been constructed for the multiphase pumps [4–13]. CFD can be considered as a powerful tool for analyzing fluid flow and transport phenomena in many applications. However, only a few studies were conducted on the new multiphase pump using CFD models [4]. Development of detailed CFD transient models often takes a lot of computer resources and time. As a usual method, the neglect of partial leakages and energy losses and the assumption of homogeneous flows were adopted. Consequently, these CFD models are still insufficient to describe the complex behavior of reciprocating multiphase pumps. Additionally, some CFD modeling procedures from designers' experience, such as the selection of multiphase and turbulence models, dynamic grid technique, user defined functions, etc., have great effects on the accuracy and reliability of calculation results [13]. Therefore, a practical modeling approach should be developed for the description of different multiphase transportation conditions.

Recently, data-driven empirical models have been widely applied in chemical process applications [14–23]. Recent reviews can be referred to the literature [22,23]. Data-driven models can generally be constructed quickly without substantial understanding of the process and requiring much experience of designers. However, they have their own shortcomings. For example, the reliability of training data is an important factor for the accuracy of data-driven empirical models. It is difficult to collect enough historical data of different multiphase transportation

[*] Corresponding author.
E-mail address: yliuzju@zjut.edu.cn (Y. Liu).

https://doi.org/10.1016/j.chemolab.2018.10.005
Received 24 August 2018; Received in revised form 10 October 2018; Accepted 16 October 2018
Available online 17 October 2018

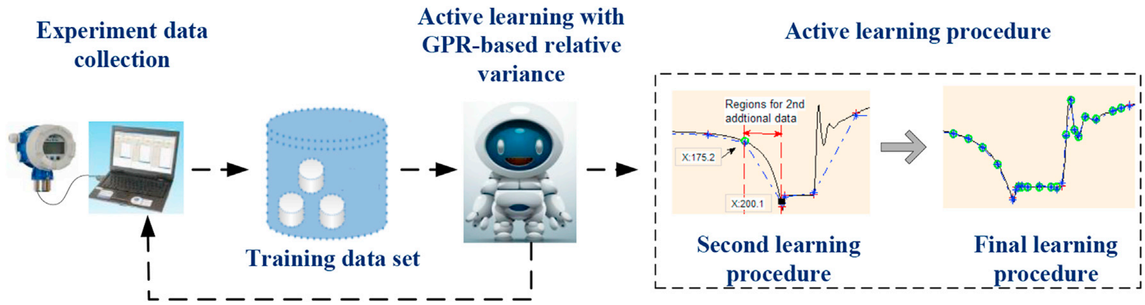
Experiment data collection
Active learning with GPR-based relative variance
Active learning procedure
Training data set
Regions for 2nd addtional data
X:175.2
X:200.1
Second learning procedure
Final learning procedure

Fig. 1. A brief scheme of the active learning method for dynamical flow rate prediction of reciprocating multiphase pumps.

conditions due to expensive instruments and significant human efforts [24,25]. Consequently, conventional training methods may be inefficient to construct a data-driven empirical model for the prediction of flow rates. How to actively design informative training data should be elaborately investigated.

To describe the relationships between flow rates and multiphase conditions, nonlinear data-driven models are more suitable. Gaussian process regression (GPR), as a probabilistic modeling method, can evaluate the uncertainty of predictions [26–33]. However, for complex flow rate curves, a GPR model with limited training data cannot always function well for a long-term utilization. For those regions with poor performance, additional data should be introduced into the current model to improve its prediction accuracy. Recently, some active learning methods were proposed to select significant samples from the unlabeled dataset for labeling, and handle the regression problems under limited labeled data [30–32]. However, they only utilized variance-based criterion which is suitable for steady-state processes rather than dynamic processes. How to actively learn the flow rate curves with dynamic characteristics needs further investigation.

To overcome the problems aforementioned, as briefly shown in Fig. 1, a novel learning method to actively design of informative data is developed to enhance the prediction accuracy of dynamical flow rates of a stroke. The GPR-based relative variance is utilized to find out from which regions the new data should be adopted to enhance the model quality. Moreover, an evaluation criterion is proposed to implement the active learning procedure efficiently. Consequently, a set of informative data are sequentially obtained from the experiments and included into the next learning step.

The remainder parts are so organized. In Section 2, the learning procedure of the traditional GPR model is explored for active design of informative training data. Detailed implementations of the sequential learning strategies are described in Section 3. The prediction results are compared with the experiment data to show its advantages in Section 4. Finally, the work is summarized in Section 5.

2. Experimental system and probabilistic information analysis

In this section, the learning procedure of the trained GPR model is explored. First, the experimental system is described to obtain the training data. Second, the GPR-based probabilistic information is analyzed to discover informative data to be included into the GPR model. It is important to describe the process dynamical characteristics of a flow rate curve, and then introduce new representative data into the current model to enhance its prediction performance.

2.1. Experimental system description

A three-cylinder double-acting reciprocating multiphase pump is adopted as the test pump. The crank angle at the end of a stroke is denoted as θ_e. With the change of crank angle ($\theta = 0^\circ \sim \theta_e$), the modeling data of different multiphase transportation conditions are collected from the experimental system shown in Fig. 2. The rotor pump is driven by the variable-frequency motor to allow the crude oil into the liquid pipeline. Simultaneously, the air supplied by the compressor enters the gas pipeline. The crude oil and air in the gas-liquid mixer are fully mixed and flow into the test pump, and then flow back into the tank for recycling. The gas volume fraction at the pump inlet is controlled by regulating the flow rates of gas and liquid circuits before the mixer. Two limit switches are installed to monitor the opening and closing points of the suction and discharge valves. The flow rates of the oil-gas mixture are measured by a multiphase flowmeter. The crank angles are measured by an angular displacement sensor. The sensors and the industrial control computer are integrated to complete the data acquisition.

It should be mentioned that the experimental system is designed mainly according to the standard “GB/T 7784-2006 Test Methods for Powder Reciprocating Pumps”. In this standard, there are some design requirements and operation instructions to ensure running safety. First, a relief valve must be arranged on the discharge pipeline of the reciprocating pump, to ensure that the discharge pressure is not higher than its rating. Second, before the pump starts, all discharge valves of the pipelines must be opened. Third, the joints of the suction pipelines have no leak to prevent the outside air from entering the pipelines. Consequently, in the following parts, the experimental data under designed would not be from dangerous regimes.

2.2. Probabilistic information analysis for trained GPR models

Generally, GPR learns a model f approximating a training set $\mathbf{S} = \{\mathbf{X}, \mathbf{y}\}$. Due to different dynamical characteristics of flow rate curves under different multiphase transportation conditions, the samples of a multiphase condition are considered as one subclass for simpler and more effective learning procedures. Therefore, the initial dataset $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_m)^{\mathrm{T}}, m = 1, \ldots, M$ are collected from M multiphase transportation conditions. The mth subclass with N_m samples can be represented as $\mathbf{S}_m = \{\mathbf{X}_m, \mathbf{y}_m\} = \{\mathbf{x}_{m,i}, y_{m,i}\}_{i=1}^{N_m}$.

As important factors, the suction pressure P_s, the discharge pressure P_d, the gas volume fraction β, and the crank angle θ, are selected as the input variables, i.e., $\mathbf{x}_{m,i} = [P_{sm,i}, P_{dm,i}, \beta_{m,i}, \theta_{m,i}]^{\mathrm{T}}$ [3,6]. Without loss of generality, the flow rate of a pump cavity is the output variable, i.e., $y_{m,i} = Q_{m,i}$. For an output variable $\mathbf{y}_m$, the GPR model is the regression function with a Gaussian prior distribution and zero mean or, in a discrete form [26].

$$\mathbf{y}_m = \left(y_{m,1}, \ldots, y_{m,N_m}\right)^{\mathrm{T}} \sim G(0, \mathbf{C}_m) \tag{1}$$

where $\mathbf{C}_m$ is the $N_m \times N_m$ covariance matrix with the ij-th element $C_m(\mathbf{x}_{m,i}, \mathbf{x}_{m,j})$. Using the Bayesian method to train the mth GPR model, the matrix $\mathbf{C}_m$ can be calculated [26].

For the tth test set with N_t samples $\mathbf{X}_t = \{\mathbf{x}_{t,i}\}_{i=1}^{N_t}, t = 1, \cdots, T$, the predicted output of $y_{t,i}$ (i.e., $\widehat{y}_{m,ti}$) with its variance (i.e., $\sigma^2_{\widehat{y}_{m,ti}}$) can be obtained below [26].

$$\widehat{y}_{m,ti} = \mathbf{k}^{\mathrm{T}}_{m,ti}\mathbf{C}_m^{-1}\mathbf{y}_m \tag{2}$$

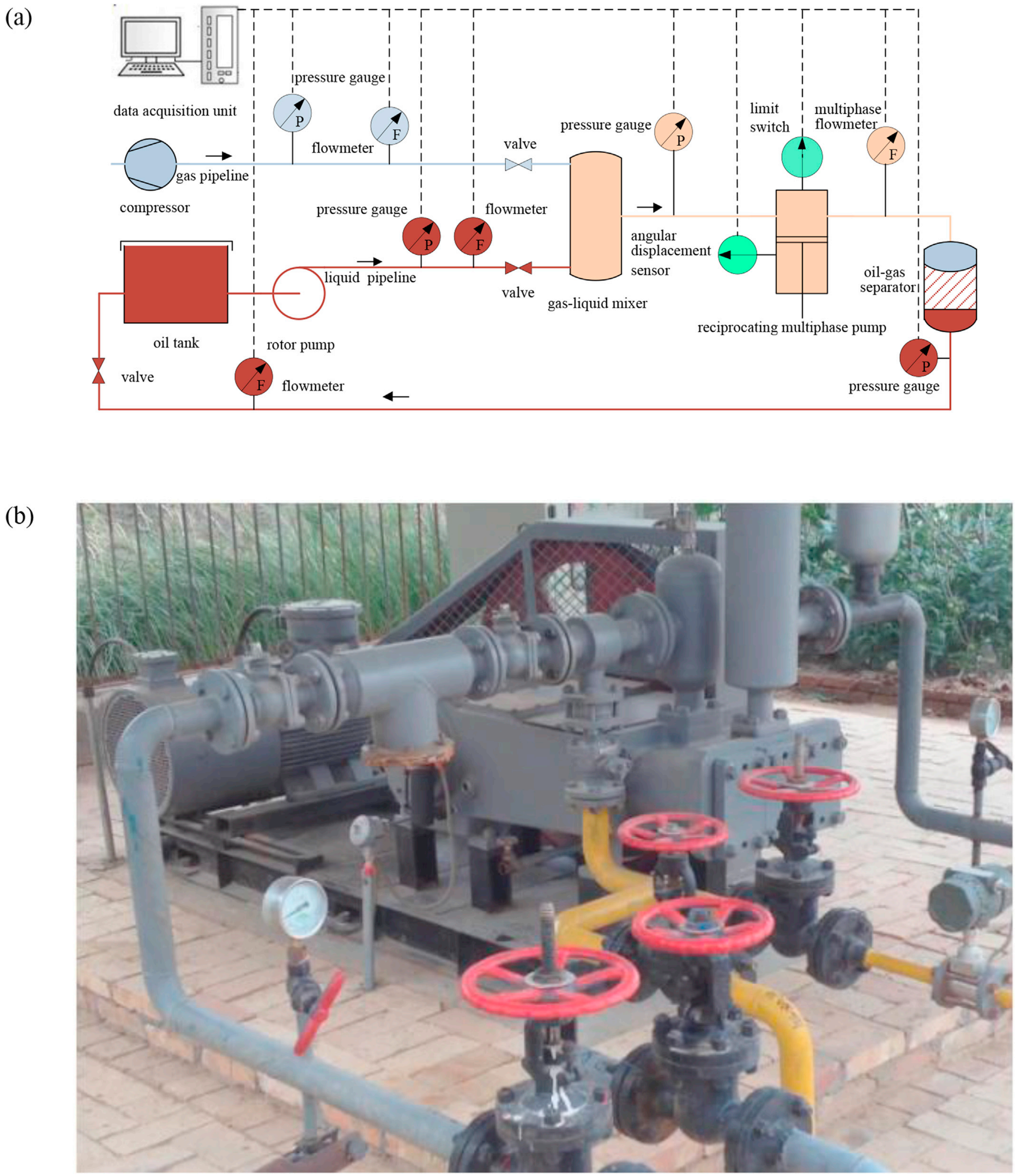

Fig. 2. **(a)** The flowchart of the experimental system for measuring the flow rate of the three-cylinder double-acting reciprocating multiphase pump **(b)** Experimental set-up for measuring the flow rate of the three-cylinder double-acting reciprocating multiphase pump.

$$\sigma^2_{\hat{y}_{m,ti}} = k_{m,ti} - \mathbf{k}^{\mathrm{T}}_{m,ti}\mathbf{C}_m^{-1}\mathbf{k}_{m,ti} \tag{3}$$

where $\mathbf{k}_{m,ti} = [C_m(\mathbf{x}_{t,i}, \mathbf{x}_{m,1}), C_m(\mathbf{x}_{t,i}, \mathbf{x}_{m,2}), \ldots, C_m(\mathbf{x}_{t,i}, \mathbf{x}_{m,N_m})]^{\mathrm{T}}$ is the covariance vector between the test input and the training data, and $k_{m,ti} = C(\mathbf{x}_{t,i}, \mathbf{x}_{t,i})$ is the covariance of the test input sample. The detailed algorithmic implementations are described in the literature [26]. Consequently, M GPR models, denoted as GPR$_m$, $m = 1, \ldots, M$, can be constructed for M subclasses.

As an illustrated case, a GPR model is offline constructed with 12

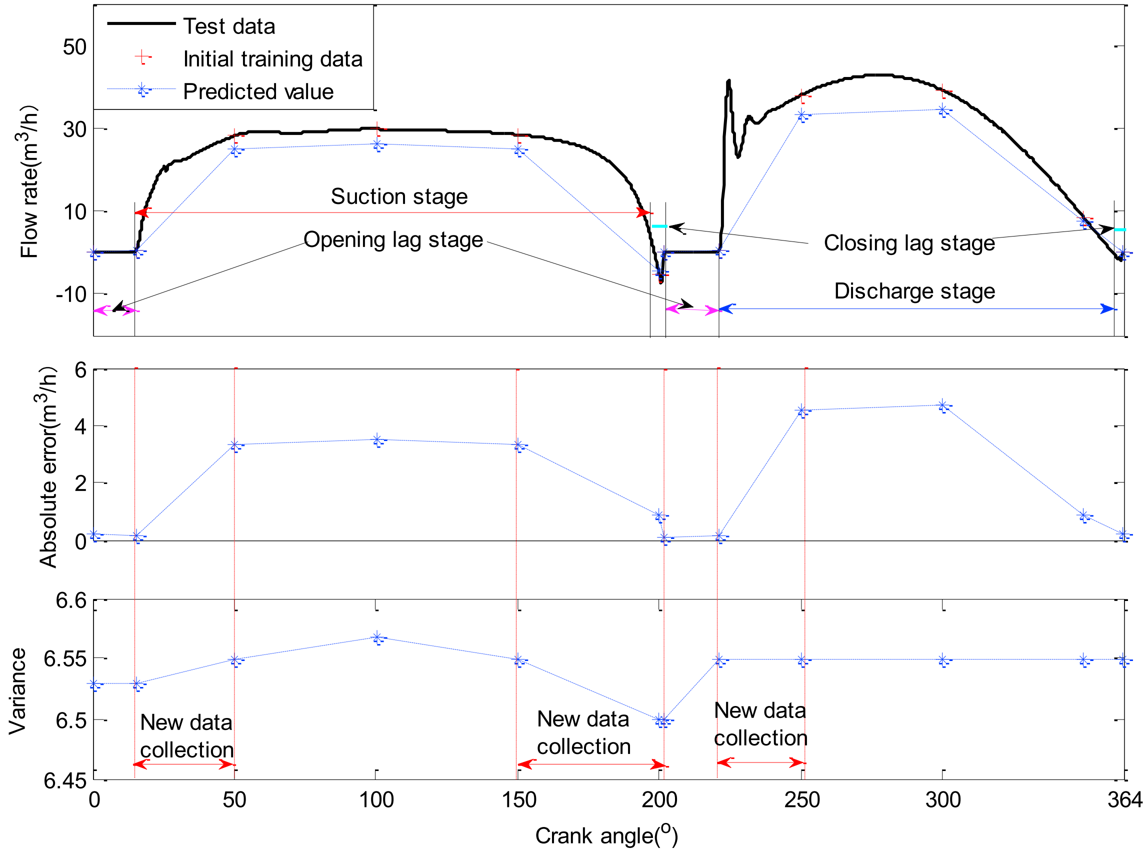

Fig. 3. Learning results of the traditional GPR model for a dynamical flow rate curve in a multiphase transport condition.

initial training data of a typical multiphase transportation condition using Eq. (1). Four training data directly come from two limit switches mentioned in Section 2.1. Their flow rates are near zero, corresponding to the opening and closing points of the suction and discharge valves. The remaining data are sampled at an interval of 50° with the change of crank angle ($\theta = 0° \sim 364°$). The predicted and variance values of 12 initial samples are calculated using Eq. (2) and Eq. (3), respectively. The detailed learning results are shown in Fig. 3.

It can be seen that the learning performance of the opening and closing lag stages is good with relatively small errors. However, the learning results of the suction and discharge stages are inaccurate. In particular, it is shown that the trained GPR model cannot capture the process characteristics during three transitions (i.e., (15°∼50°), (150°∼202°), and (221°∼250°)) mainly because of insufficient training data and more complex flow fields. Consequently, an effective and practical method should be developed to introduce new informative data into these regions.

As also shown in Fig. 3, the trained data far away from the experiment ones generally have a relatively large prediction variance. From the modeling point of view, a larger value of prediction variance generally means a larger uncertainty when the current GPR model is adopted for its prediction [32]. Thus, the prediction variance is an important factor for the estimation of regression performance. Intuitively, a variance-based criterion of adding new data into the current model can be considered. That is to say, the new data should be first introduced into the regions (located nearby $\theta = 100°$) mainly because the mode with $\theta = 100°$ has the largest prediction variance. Actually, the mentioned transitions should be improved and learned first for their strong nonlinearity and fast time-varying characteristics. Therefore, only using the variance-based method may be inefficient to capture the transient behavior because the process characteristics are not explored.

Fortunately, as also shown in Fig. 3, there is a big change in variance values for two adjacent samples in different stages. It is mainly because different stages of a flow rate curve show distinguished characteristics, especially for those transitions. The first transient region corresponds to the lag opening moment of the suction valve, and the mixture flows out of the pump cavity at a slightly larger speed for the gas expansion in the pump. Additionally, the second transient region is the lag closing moment of the suction valve, and the backflow phenomenon appears for the excess of cavity pressure. Moreover, the third transient region relates to the lag opening moment of the discharge valve. The mixture contains continually and highly compressed gas, and it flows out of the pump cavity rapidly to reach the maximum flow rate. Finally, the flow rate reduces suddenly and vibrates to coincide with the volume change rate for the decompression of the gas. Consequently, compared with four different stages, all three transitions correspond to more complex flow fields. In the following section, how to use this interesting information to add significant data and efficiently improve the quality of the GPR model will be developed.

3. Relative variance-based active learning method

As analyzed in Section 2.2, the transitions with a relatively large change in variance should be enhanced to achieve good regression performance. The new data in these regions should be introduced into the current GPR model. For the description of the change in variance, the absolute relative variance (ARV) between two adjacent samples is proposed and defined as

$$\text{ARV}^k_{m,i} = \left|\sigma^2_{\hat{y}_{m,i+1}} - \sigma^2_{\hat{y}_{m,i}}\right|, i = 1, ..., N_m + k - 2, k = 1, ..., K_m \quad (4)$$

where $\text{ARV}^k_{m,i}$ is the ith absolute relative variance of the mth training subclass $\mathbf{S}_m$ in the kth learning procedure, K_m is the learning number of $\mathbf{S}_m$.

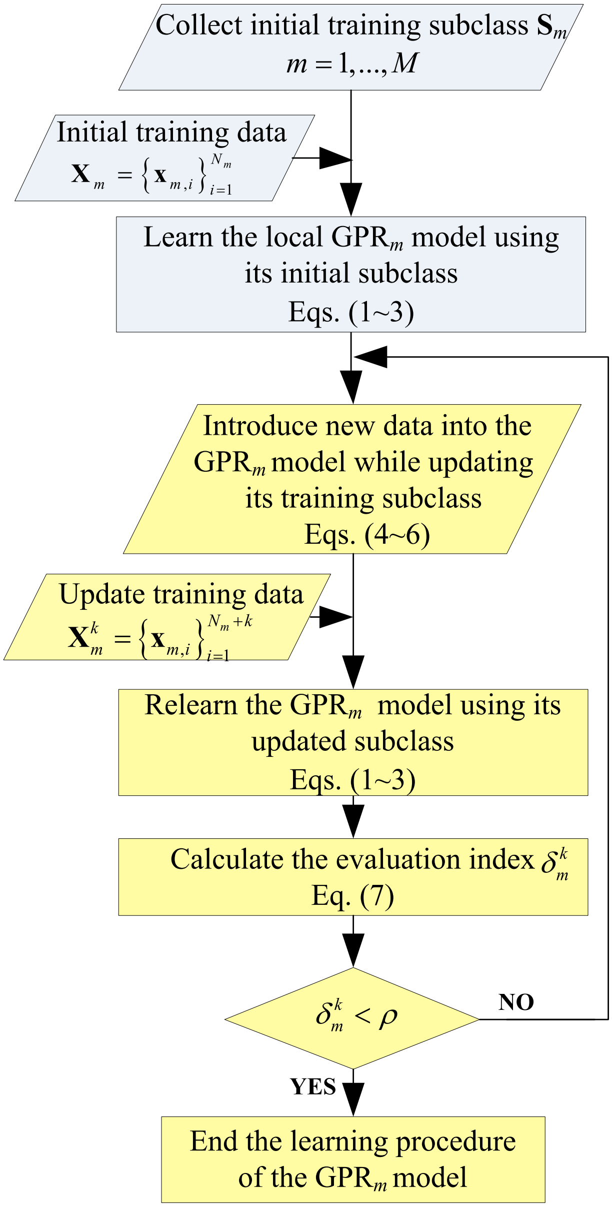
Collect initial training subclass $\mathbf{S}_m$
$m = 1,\dots,M$
Initial training data
$\mathbf{X}_m = \{\mathbf{x}_{m,i}\}_{i=1}^{N_m}$
Learn the local GPR$_m$ model using its initial subclass
Eqs. (1~3)
Introduce new data into the GPR$_m$ model while updating its training subclass
Eqs. (4~6)
Update training data
$\mathbf{X}_m^k = \{\mathbf{x}_{m,i}\}_{i=1}^{N_m+k}$
Relearn the GPR$_m$ model using its updated subclass
Eqs. (1~3)
Calculate the evaluation index δ_m^k
Eq. (7)
$\delta_m^k < \rho$
NO
YES
End the learning procedure of the GPR$_m$ model

Fig. 4. The flowchart of the active learning method for dynamical flow rate prediction of reciprocating multiphase pumps.

Additionally, as mentioned in Section 2.2, those transitions where new data should be first introduced have generally larger values of ARV than other stages. Consequently, the input regions of the data with the maximum ARV (MARV) should be first improved. To describe this, a relative variance-based criterion of adding new data into the model can be defined as follows

$$\mathrm{MARV}_m^k = \max \mathrm{ARV}_{m,i}^k, i = 1,\dots,N_m + k - 2, k = 1,\dots,K_m \tag{5}$$

where MARV$_m^k$ is the value of the maximum absolute relative variance of $\mathbf{S}_m$ in the kth learning procedure. Without loss of generality, suppose the ith training data has the MARV value. Its crank angle $\theta_{m,i}^k$ can be obtained from the input variables. Accordingly, the crank angle next to $\theta_{m,i}^k$ is $\theta_{m,i+1}^k$. The new data $\mathbf{x}_{m,N_m+k}$ between $\theta_{m,i}^k$ and $\theta_{m,i+1}^k$ should be introduced into the GPR model. And its crank angle is described as

$$\theta_{m,N_m+k} = \theta_{m,i}^k + \frac{\theta_{m,i+1}^k - \theta_{m,i}^k}{2}, k = 1,\dots,K_m \tag{6}$$

Consequently, with the relative variance-based criterion in Eqs. (4)–(6), new data will be sequentially introduced into the existing GPR model.

Moreover, to judge whether any new data should be introduced into the model, an evaluation index of the active learning procedure is proposed as follows

$$\delta_m^k = \left|\hat{\bar{y}}_m^k - \hat{\bar{y}}_m^{k-1}\right|, m = 1,\dots,M \tag{7}$$

where δ_m^k is the evaluation index of the kth learning procedure of $\mathbf{S}_m$. The items $\hat{\bar{y}}_m^{k-1}$ and $\hat{\bar{y}}_m^k$ represent the average prediction values in the (k−1)th and the kth learning procedures of $\mathbf{S}_m$, respectively. It means that, if the predictions of the kth and the (k−1)th learning procedures of $\mathbf{S}_m$ are almost the same, the active learning procedure can be finished. Thus, a small positive value ρ is chosen. If $\delta_m^k > \rho$, the active learning procedure should be continued. Otherwise, it should be stopped because no further information can be introduced into the model to enhance its prediction performance.

In summary, taking the GPR$_m$, $m = 1,\dots,M$ model as an example, the main implementations of the active learning method with the relative variance-based criterion are illustrated in Fig. 4. Each local model completes its entire learning process respectively. For this case, the training dataset is not large. The local models are trained from scratch after a set of new informative data adding into the current models. The step-by-step procedures are described as follows.

Step 1: Collect the initial training subclass $\mathbf{S}_m$ of the mth multiphase transportation condition.
Step 2: Learn the local GPR$_m$ model using its initial training subclass and Eqs. (1)–(3).
Step 3: Introduce new data into the GPR$_m$ model while updating its training subclass using Eqs. (4)–(6).
Step 4: Relearn the GPR$_m$ model using its updated training subclass and Eqs. (1)–(3).
Step 5: Obtain the evaluation index δ_m^k to judge whether the active learning procedure of the GPR$_m$ model should be continued using Eq. (7). If $\delta_m^k > \rho$, go to **Step 3.** Otherwise, the active learning procedure should be stopped.

In the above learning steps, the proposed method can actively design informative data to enhance the model. Compared with only using the variance-based criterion [32], the complex dynamical characteristics of a flow rate curve can be better described in this way. From an engineering standpoint, the active learning method can be implemented straightforward.

4. Experimental results and discussion

The effects of input variables, such as suction pressure P_s, discharge pressure P_d, gas volume fraction β, and crank angle θ, on the flow rates of a reciprocating multiphase pump are shown in Fig. 5. It can be seen that, compared with other input variables, the gas volume fraction β and crank angle θ have more influences on the flow rates. Simultaneously, considering the operation conditions of the test pump, the experiments are conducted in the same angular velocity of crank ($\omega = 8\pi$ rad/s), clearance volume ($V_0 = 1.9 \times 10^5\,\mathrm{mm}^3$), suction pressures ($P_s = 0.4$ MPa), discharge pressures ($P_d = 3.0$ MPa), and different gas volume fractions (β = 0.1, 0.3, 0.5, 0.7, 0.8, 0.9).

As shown in Fig. 6, with the change of crank angle ($\theta = 0° \sim \theta_e$), the modeling samples of six operating conditions are collected. They are denoted as $\mathbf{S} = (\mathbf{S}_1,\dots,\mathbf{S}_6)$. Altogether 48 initial samples of four sets (i.e., $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_5$ and $\mathbf{S}_6$) are used for training, and 32 samples of $\mathbf{S}_3$ and 38 samples of $\mathbf{S}_4$ are for test. For each sample set, four training data directly come from two limit switches corresponding to the opening and closing points of the suction and discharge valves. Additionally, for $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_5$ and

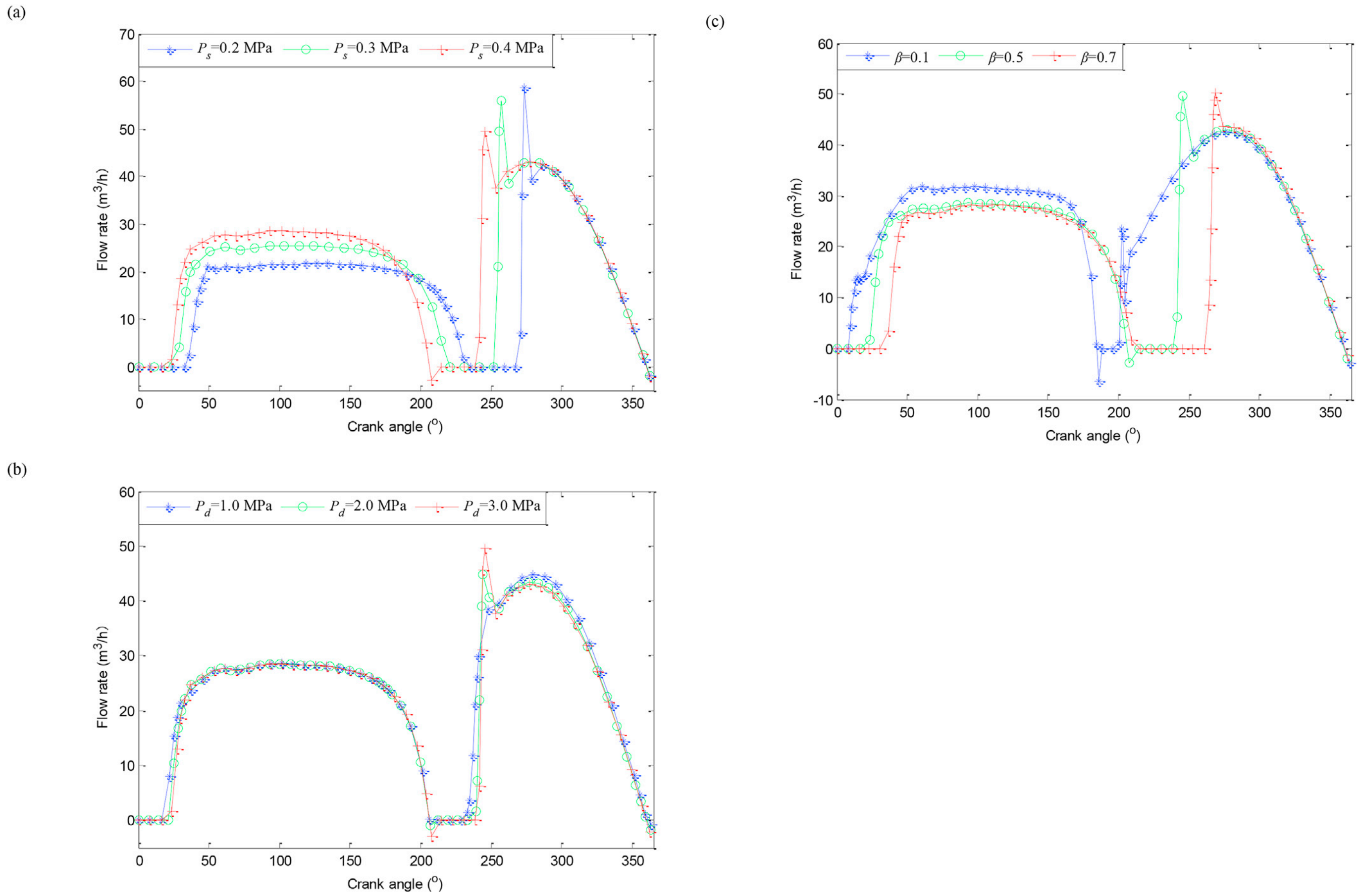

Fig. 5. (a) The flow rates of a reciprocating multiphase pump with different suction pressures **(b)** The flow rates of a reciprocating multiphase pump with different discharge pressures **(c)** The flow rates of a reciprocating multiphase pump with different gas volume fractions.

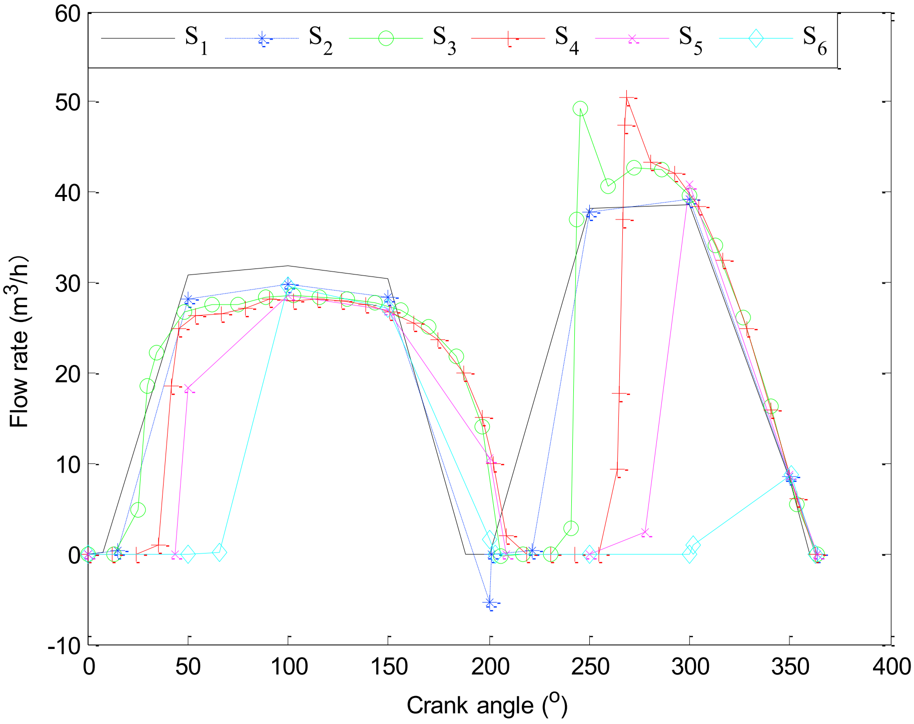

Fig. 6. The initial training data of six operation conditions.

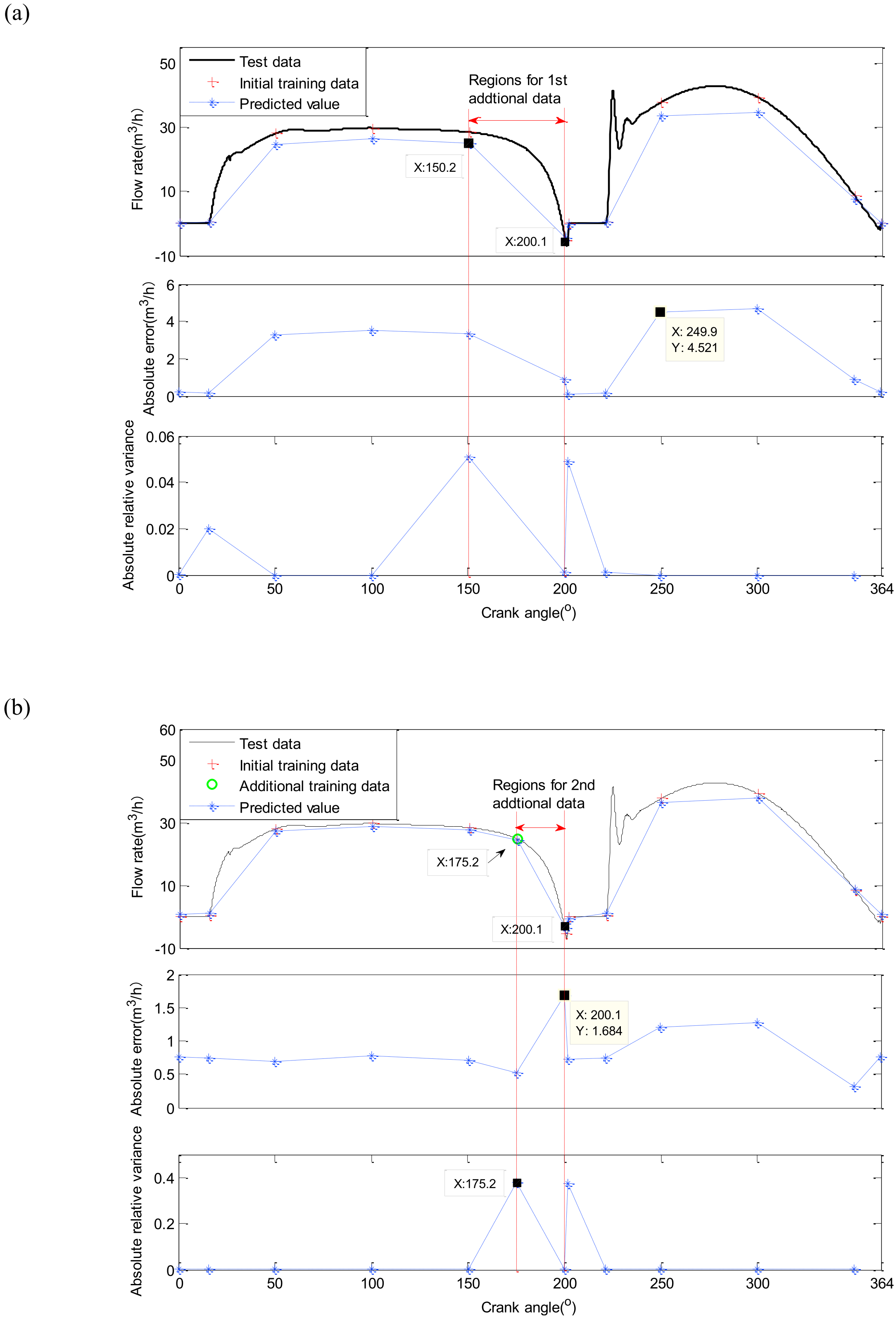

Fig. 7. **(a)** The first learning results with the relative variance-based criterion for $\mathbf{S}_2$ **(b)** The second learning results with the relative variance-based criterion for $\mathbf{S}_2$.

$\mathbf{S}_6$, the remaining data are sampled at an interval of 50° with the change of crank angle ($\theta = 0^\circ \sim \theta_e$). Moreover, based on four training data from two limit switches, the remaining data of $\mathbf{S}_3$ and $\mathbf{S}_4$ are sampled at relatively small intervals in different stages to validate the proposed method.

To compare the prediction performance of different models, two common indices, the root-mean-square error (RMSE) and maximum error (ME) are adopted. As an example, for the tth ($t = 1, \ldots, T$) test subclass, its RMSE$_t$ and ME$_t$ indices are defined as follows

Table 1
Learning result comparisons of the variance-based and relative variance-based methods (24 learning iterations) for four training sets.

Method	Index	S_2	S_1	S_5	S_6
Initial	ME	17.55	4.69	16.11	27.30
Variance-based [32]	(m^3/h)	17.16	2.70	15.75	29.65
Relative variance-based		<u>3.57</u>	<u>1.65</u>	<u>5.66</u>	<u>7.71</u>
Initial	RMSE	7.92	2.56	8.03	12.94
Variance-based [32]	(m^3/h)	6.93	1.52	6.62	9.78
Relative variance-based		<u>0.94</u>	<u>0.80</u>	<u>1.61</u>	<u>1.76</u>

$$\mathrm{RMSE}_t = \sqrt{\sum_{i=1}^{N_t} \left(\widehat{y}_{t,i} - y_{t,i}\right)^2 / N_t} \tag{8}$$

$$\mathrm{ME}_t = \max\left|\widehat{y}_{t,i} \ - \ y_{t,i}\right|, i = 1, \dots, N_t \tag{9}$$

where $\widehat{y}_{t,i}$ denotes the prediction of $y_{t,i}$, and N_t is the sample number of the tth test set.

4.1. Active learning results and discussion

Taking $\mathbf{S}_2$ as an illustrated case, the first and second iterations of active learning are exhibited to show how the proposed approach enhances the GPR model in a sequential and efficient way. As shown in Fig. 7(a), the data point with $\theta = 150.2°$ has the largest relative variance. Using Eq. (6), the first new data point to be introduced locates in the region of $150.2° \sim 200.1°$. As a result, the data point with $\theta = 175.2°$ is collected from the experiment and then added to the training set $\mathbf{S}_2$. With the same method, the second new data point (located nearby $\theta = 188°$) in the region $(175.2° \sim 200.1°)$ is introduced into $\mathbf{S}_2$ after the second learning iteration shown in Fig. 7(b). Moreover, compared with the first learning procedure, the GPR_1 model obtains better regression performance with the smaller ME and RMSE values in the second learning procedure. The ME value is reduced from 4.52 m^3/h to 1.68 m^3/h, and the RMSE value is reduced from 2.56 m^3/h to 0.91 m^3/h.

Consequently, k new data points are adopted into the training set $\mathbf{S}_2$ in a sequential step until the evaluation index $\delta_1^k < \rho$. Here, a small positive value is chosen, i.e., $\rho \ = 0.2$. Simultaneously, the prediction performance of the GPR_1 model can be improved gradually with the introduction of the informative data. Using the same method, the other training sets can complete the learning procedures, respectively.

Table 2
Learning result comparisons of the traditional, variance-based, and relative variance-based methods for $\mathbf{S}_2$.

Method	No. of training data	ME (m^3/h)	RMSE (m^3/h)
Traditional	123	20.52	2.44
Variance-based [32]	78	8.08	1.49
Relative variance-based	<u>36</u>	<u>1.65</u>	<u>0.80</u>

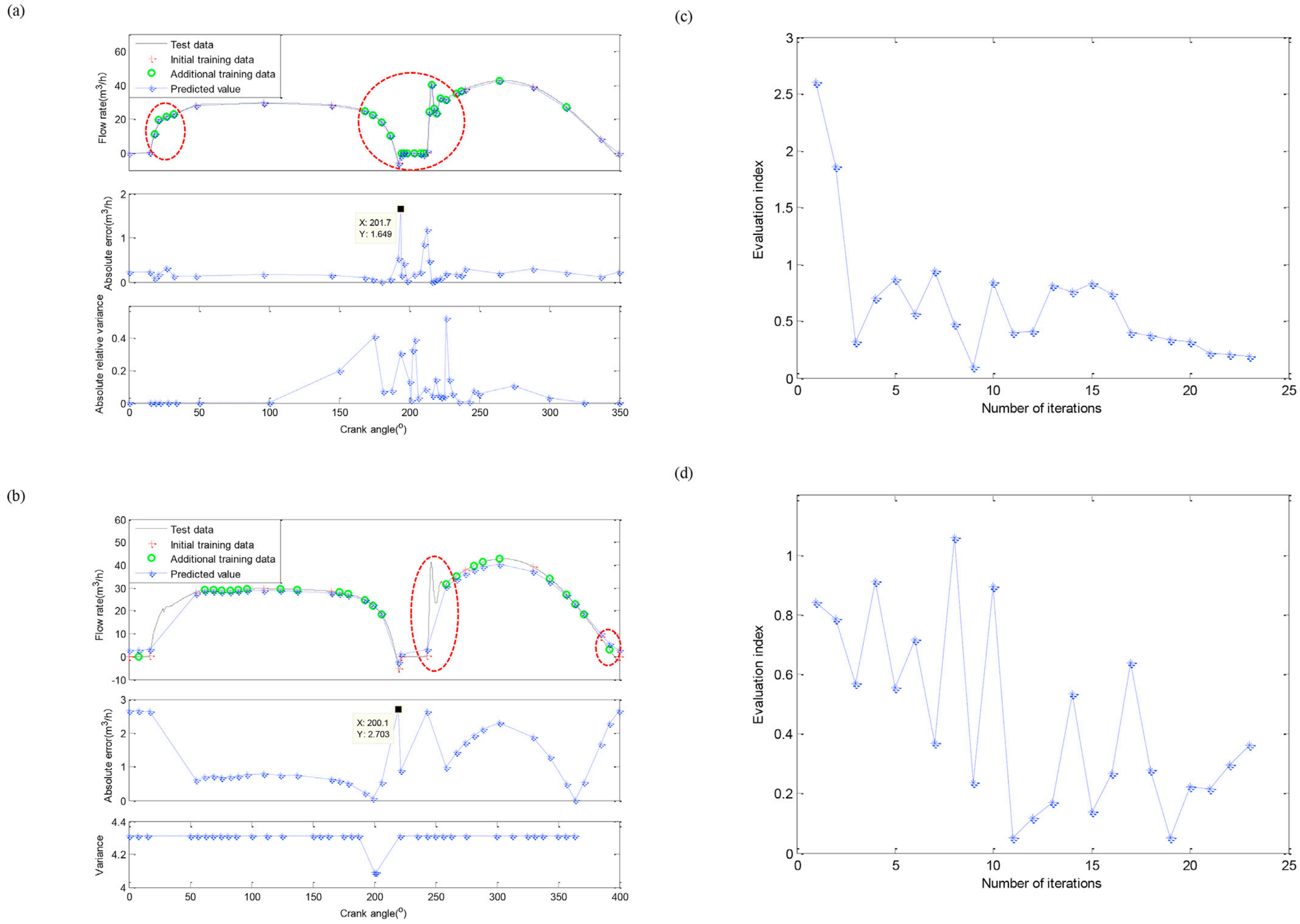

Fig. 8. **(a)** Final learning results using the relative variance-based criterion for $\mathbf{S}_2$ (24 learning iterations) **(b)** Learning results using the variance-based criterion for $\mathbf{S}_2$ (24 learning iterations) **(c)** Evaluation indices using the relative variance-based criterion for $\mathbf{S}_2$ (24 learning iterations) **(d)** Evaluation indices using the variance-based criterion for $\mathbf{S}_2$ (24 learning iterations).

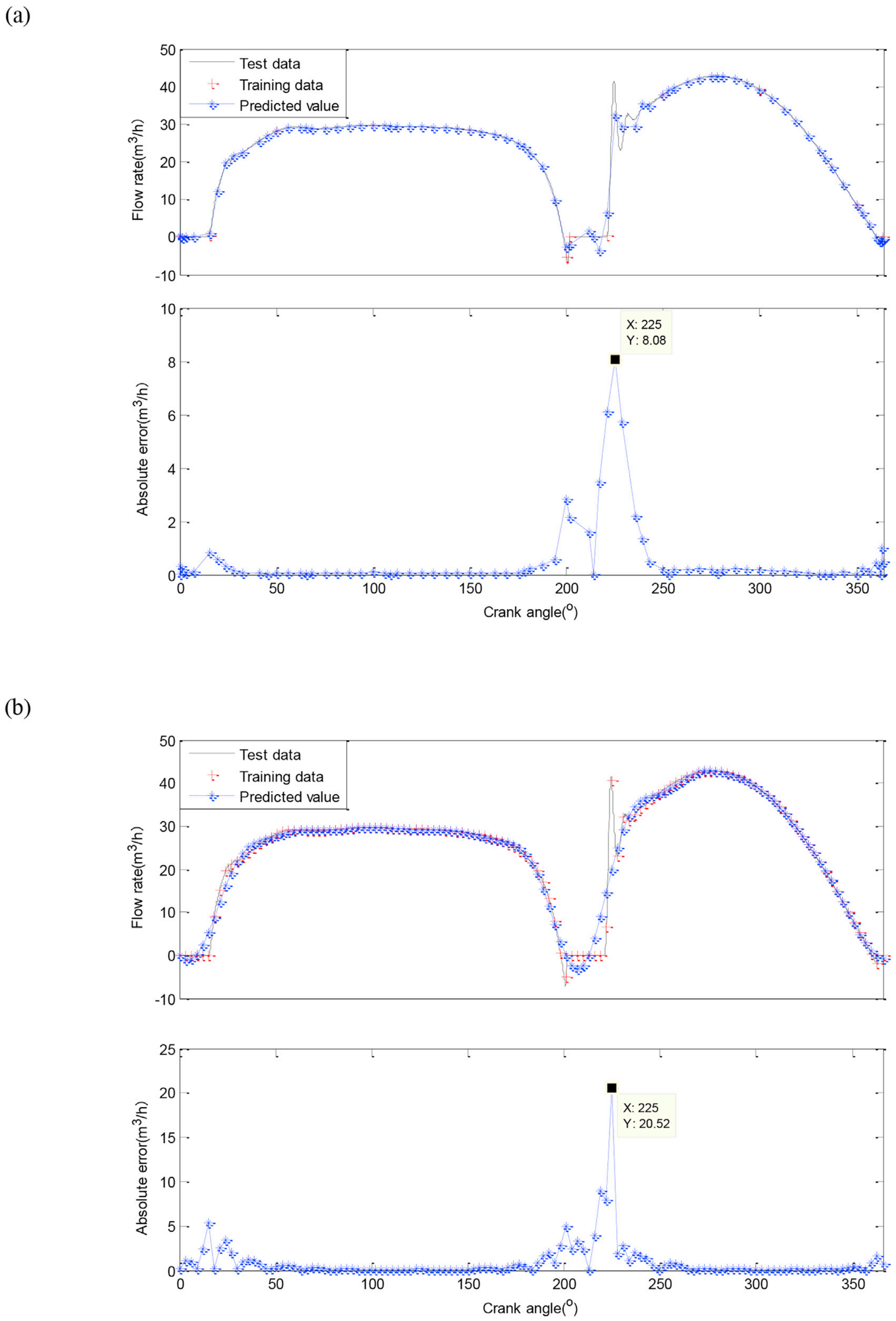

Fig. 9. **(a)** Learning results using the variance-based criterion for $\mathbf{S}_2$ (78 training data) **(b)** Learning results of the traditional method for $\mathbf{S}_2$ (123 training data).

To show the effectiveness of the proposed method, the compared learning results of the 24th iterations using the relative variance-based and variance-based criteria are listed in Table 1. It can be found that all trained GPR models with the relative variance-based criterion obtain better regression performance with the smaller ME and RMSE indices. Additionally, the prediction performance of four models has been improved greatly after 24 learning iterations.

However, in comparison with the initial learning results, four trained GPR models with the variance-based criterion have better prediction performance. However, the regression performance for data in some

Table 3
Prediction performance comparisons of the initial, traditional, variance-based, and relative variance-based trained GPR models for the test sets of $\mathbf{S}_3$ and $\mathbf{S}_4$.

Model	No. of training data	S_3		S_4	
		ME (m^3/h)	RMSE (m^3/h)	ME (m^3/h)	RMSE (m^3/h)
Initial	43	44.18	16.00	45.57	14.96
Traditional	505	29.51	8.51	29.60	10.47
Variance-based trained [32]	163	30.44	8.31	40.38	12.03
Relative variance-based trained	163	9.99	5.27	27.78	8.99

areas is still poor for relatively large ME indices. It is mainly because the variance-based criterion only considers the prediction variance. The dynamical flow rate curve of a stroke is not explored.

As an example, detailed comparisons of $\mathbf{S}_2$ between the proposed method and variance-based method are shown in Fig. 8. As shown in Figs. 8(a), 24 new data have been sequentially introduced into the regions with the proposed method using Eqs. (4)–(6). It should be mentioned that 20 new data are added into the transitions. It indicates that representative data are explored and then quickly introduced into the current model. However, as shown in Fig. 8(b), only 5 out of 24 new data are introduced into the transitions using the variance-based criterion. Consequently, some important dynamical information in the opening moment of the discharge valve, including the maximum flow rate, the characteristics of flow fluctuation, etc., cannot be captured by the current model.

The evaluation indices of 24 iterations with the proposed method are shown in Fig. 8(c). The trend is decreasing with iterations. Actually, after three iterations, the quality of the GPR_1 model has been enhanced greatly. And after 24 iterations, the GPR_1 model has finished its learning procedures for the evaluation index $\delta_1^k < \rho = 0.2$. Unlike Fig. 8(c), as shown in Fig. 8(d), the decreasing trend is relatively slow with the variance-based criterion. This means that the proposed method is more efficient for dynamical processes than only using the variance-based criterion.

Moreover, detailed comparisons among the proposed, the variance-based, and traditional methods are listed in Table 2. To achieve comparable RMSE indices, the proposed method only adopts 36 new data while the variance-based method introduces 78 data and the traditional method collects 123 data at intervals of 3°. It indicates that the proposed method with much less training data obtains better regression performance. The detailed learning results shown in Fig. 9 also illustrate the variance-based and traditional learning methods cannot capture the process characteristics of the opening moment of the discharge valve. It is mainly because the regions are more complex and some significant data are not included into the GPR model using the variance-based and traditional methods.

Therefore, all learning results shown in Figs. 7–9 and listed in Tables 1 and 2 validate that the relative variance-based method can sequentially enhance the quality of a prediction model. By elaborately designing of information data, this active learning method can be implemented in a simple and efficient way.

4.2. Prediction results and discussion

To further validate the proposed method, four trained sets using the relative variance-based criterion are merged into a training set for construction of a global GPR model. Simultaneously, with the same iterations, four trained sets using the variance-based criterion also are used to establish another global GPR model. Their prediction results for $\mathbf{S}_3$ and $\mathbf{S}_4$ are compared with the traditional learning method.

First, the prediction results for $\mathbf{S}_3$ are listed in Table 3 and shown in

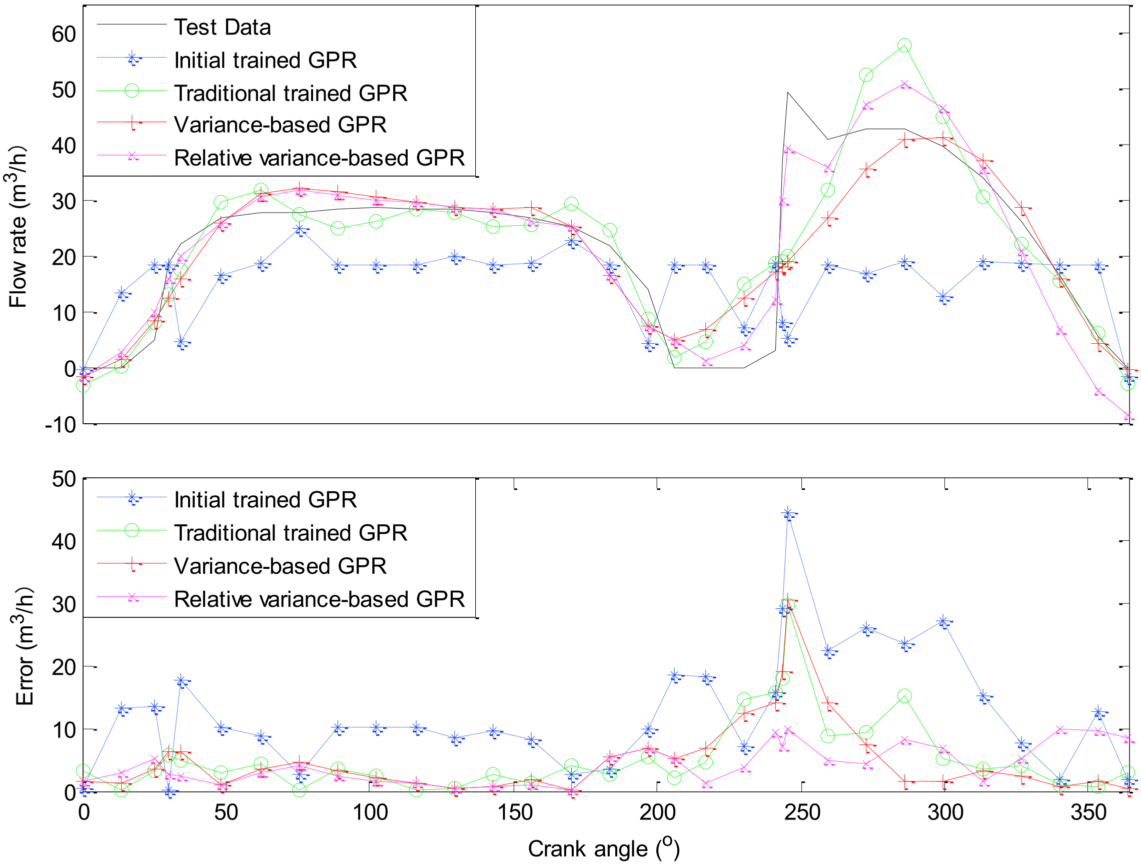

Fig. 10. Prediction performance comparisons of the initial, traditional trained, variance-based, and relative variance-based GPR models for the test set $\mathbf{S}_3$.

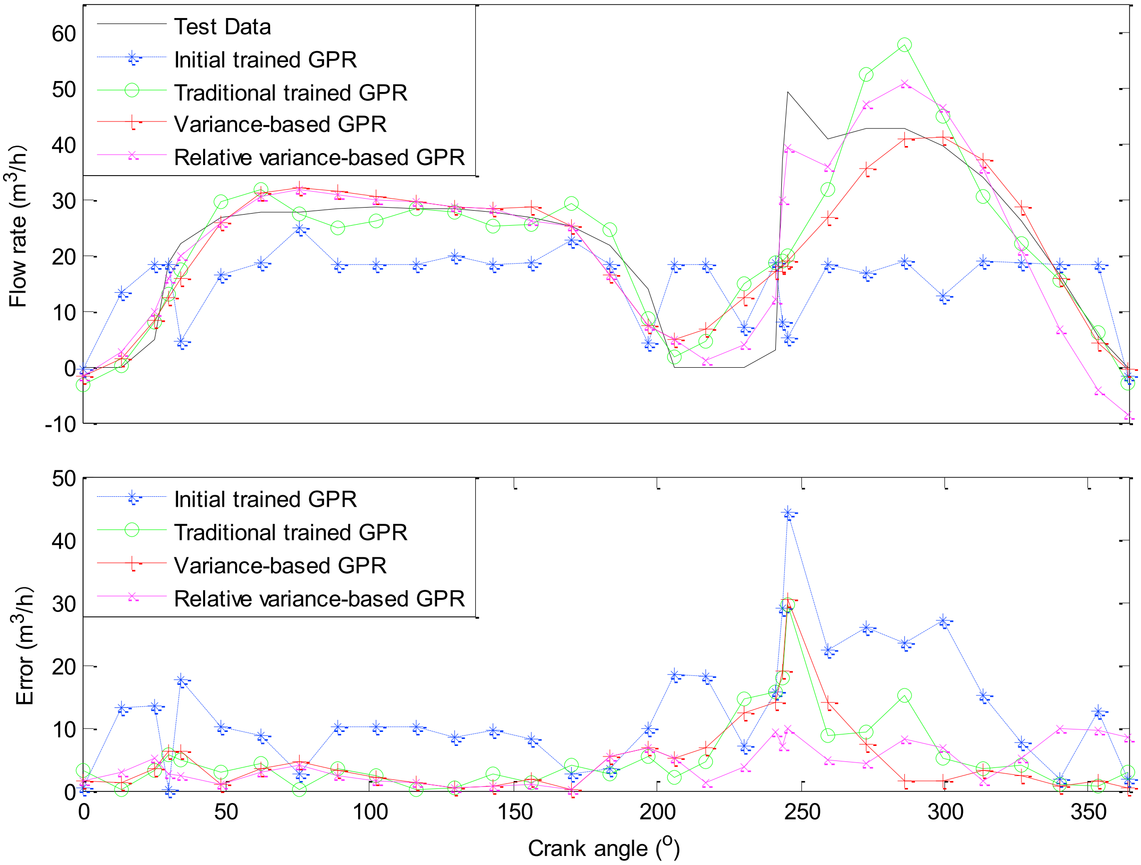

Fig. 11. Prediction performance comparisons of the initial, traditional trained, variance-based, and relative variance-based GPR models for the test set $\mathbf{S}_4$.

Fig. 10. The comparison results show that the initial prediction performance is poor mainly because there are not enough training data to capture the process characteristics. The traditional GPR model trained by 505 data has relatively large prediction errors for most of the data in the transitions (210°∼280°). It is mainly because these regions have complex dynamic characteristics, and some representative data cannot be collected with the equal-interval sampling method. The model with the variance-based criterion achieves almost the same performance as the traditional model only using 163 training data, including the initial 43 and the additional 120 data. However, the prediction results of the mentioned transitions are still inaccurate because the variance-based criterion only considers the prediction variance. Only the proposed active learning approach improves the quality of the model remarkably, and then obtains the best prediction performance using the same number data of the variance-based criterion.

The detailed results for $\mathbf{S}_4$ are shown in Fig. 11. It can be seen that the proposed method also achieves the best prediction performance. Additionally, from the ME index listed in Table 3, the proposed method is not good for some samples in the transitions. It is mainly because the gas volume fractions $\beta = 0.7$ of $\mathbf{S}_4$ is relatively high, and the internal flow fields are more complex for more important information.

From all the prediction results listed in Table 3 and shown in Figs. 10 and 11, the proposed active learning approach with the relative variance-based criterion has better prediction performance with much less training data. The main dynamical characteristics during transitions can be learned by introducing new informative data. Additionally, it can be simply and efficiently implemented for sequential training of a GPR model in practice.

5. Conclusion

This paper aims to develop a reliable model for dynamical flow rate prediction of a stroke of reciprocating multiphase pumps. In view of the difficulty in collecting lots of training data, a relative variance-based method is proposed to actively and sequentially introduce informative training data into the GPR model. Two main distinguished characteristics can be summarized. First, without any prior knowledge of the process, the proposed method can quickly introduce representative data into the transitions to capture the process nonlinearity and dynamical characteristics. Second, the probabilistic information can be effectively integrated into the modeling procedure to improve the quality of a GPR model and reduce the human efforts. Consequently, the trained GPR model can better capture the dynamical characteristics of a flow rate curve with much less training data. The experimental results show that, compared with traditional trained and variance-based methods, the proposed relative variance-based approach explores the dynamical characteristics between the samples and has significant advantages in the feasibility and simplicity.

The proposed active learning method for other chemical processes with dynamical and nonlinear characteristics is an interesting direction. Especially for those multi-stage or multi-phase processes with transitions, an active learning and modeling method is efficient in practice. Another future work is to develop an ensemble active model for multimode processes. Additionally, one interesting research direction is to integrate the process nonlinearity and output values to further enhance the proposed method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61873241) and Zhejiang Provincial Natural Science Foundation of China (Grant No. LY18F030024).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2018.10.005.

List of abbreviations

ARV	absolute relative variance
CFD	computational fluid dynamics
GPR	Gaussian process regression
MARV	maximum absolute relative variance
ME	maximum error
RMSE	root-mean-square error

References

[1] J. Falcimaigne, S. Decarre, Multiphase Production: Pipeline Transport, Pumping and Metering, Editions Technip, Paris, 2008.

[2] G. Hua, G. Falcone, C. Teodoriu, G.L. Morrison, Comparison of multiphase pumping technologies for subsea and downhole applications, Oil Gas Facil. 1 (2012) 36–46.

[3] H.Y. Deng, Y. Liu, P. Li, S.C. Zhang, Hybrid model for discharge flow rate prediction of reciprocating multiphase pumps, Adv. Eng. Software 124 (2018) 53–65.

[4] H.Y. Deng, Y. Liu, P. Li, Y. Ma, S.C. Zhang, Integrated probabilistic modeling method for transient opening height prediction of check valves in oil-gas multiphase pumps, Adv. Eng. Software 118 (2018) 18–26.

[5] M. Pietrzak, S. Witczak, Experimental study of air-oil-water flow in a balancing valve, J. Petrol. Sci. Eng. 133 (2015) 12–17.

[6] I.J. Karassik, J.P. Messina, P. Cooper, C.C. Heald, Pump Handbook, fourth ed., McGraw-Hill Education, New York, 2007.

[7] C.Y. Nakashima, S.D.O. Junior, E.F. Caetano, Thermodynamic Model of a Twin-screw Multiphase Pump, ASME 2002 Engineering Technology Conference on Energy, Houston, USA, 2002.

[8] K.K. Singh, S.M. Mahajani, K.T. Shenoy, S. Ghosh, CFD modeling of pump-mix action in continuous flow stirred tank, AIChE J. (54) (2008) 42–55.

[9] K. Räbiger, T.M.A. Maksoud, J. Ward, G. Hausmann, Theoretical and experimental analysis of a multiphase screw pump, handling gas-liquid mixtures with very high gas volume fractions, Exp. Therm. Fluid Sci. 32 (2008) 1694–1701.

[10] X. Yang, C.C. Hu, Y. Hu, Z.C. Qu, Theoretical and experimental study of a synchronal rotary multiphase pump at very high inlet gas volume fractions, Appl. Therm. Eng. 110 (2017) 710–719.

[11] J. Wang, H.B. Zha, J.M. Mcdonough, D.H. Zhang, Analysis and numerical simulation of a novel gas-liquid multiphase scroll pump, Int. J. Heat Mass Tran. 91 (2015) 27–36.

[12] J.H. Kim, H.C. Lee, J.H. Kim, Y.K. Lee, Y.S. Choi, Reliability verification of the performance evaluation of multiphase pump, J. Clim. Appl. Meteorol. (24) (2014) 1782–1786.

[13] Rainald Löhner, Applied Computational Fluid Dynamics Techniques: an Introduction Based on Finite Element Methods, John Wiley & Sons, New York, 2008.

[14] Y. Liu, J.H. Chen, Integrated soft sensor using just-in-time support vector regression and probabilistic analysis for quality prediction of multi-grade processes, J. Process Contr. (23) (2013) 793–804.

[15] W. Zheng, Y. Liu, Z. Gao, J. Yang, Just-in-time semi-supervised soft sensor for quality prediction in industrial rubber mixers, Chemometr. Intell. Lab. Syst. 180 (2018) 36–41.

[16] Y. Liu, Y. Liang, Z.L. Gao, Y. Yao, Online flooding supervision in packed towers: an integrated data-driven statistical monitoring method, Chem. Eng. Technol. 41 (2018) 436–446.

[17] O.T. Kajero, T. Chen, Y. Yao, Y.C. Chuang, D.S.H. Wong, Meta-modeling in chemical process system engineering, J. Taiwan Inst. Chem. Eng. 73 (2017) 135–145.

[18] Y. Liu, Y. Fan, J.H. Chen, Flame images for oxygen content prediction of combustion systems using DBN, Energy Fuel. (31) (2017) 8776–8783.

[19] Q. Xuan, B. Fang, Y. Liu, J. Wang, J. Zhang, Y. Zheng, G. Bao, Automatic pearl classification machine based on a multistream convolutional neural network, IEEE Trans. Ind. Electron. 65 (2018) 6538–6547.

[20] Q. Xuan, H. Xiao, C. Fu, Y. Liu, Evolving convolutional neural network and its application in fine-grained visual categorization, IEEE Access (6) (2018) 31110–31116.

[21] Y. Liu, C. Yang, Z.L. Gao, Y. Yao, Ensemble deep kernel learning with application to quality prediction in industrial polymerization processes, Chemometr. Intell. Lab. Syst. 174 (2018) 15–21.

[22] Z. Ge, Z. Song, S. Ding, Data mining and analytics in the process industry: the role of machine learning, IEEE Access (5) (2017) 20590–20616.

[23] Z. Ge, Process data analytics via probabilistic latent variable models: a tutorial review, Ind. Eng. Chem. Res. (2018), https://doi.org/10.1021/acs.iecr.8b02913.

[24] G. Falcone, G.F. Hewitt, C. Alimonti, B. Harrison, Multiphase flow metering: current trends and future developments, J. Petrol. Technol. 54 (2002) 77–84.

[25] R. Hanus, Application of the Hilbert transform to measurements of liquid-gas flow using gamma ray densitometry, Int. J. Multiphas. Flow (72) (2015) 210–217.

[26] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, 2006.

[27] Y. Liu, T. Chen, J.H. Chen, Auto-switch Gaussian process regression-based probabilistic soft sensors for industrial multigrade processes with transitions, Ind. Eng. Chem. Res. 54 (2015) 5037–5047.

[28] H.P. Jin, X.G. Chen, L. Wang, K. Yang, L. Wu, Adaptive soft sensor development based on online ensemble Gaussian process regression for nonlinear time-varying batch processes, Ind. Eng. Chem. Res. 54 (2015) 7320–7345.

[29] Y.J. He, J.N. Shen, J.F. Shen, Z.F. Ma, State of health estimation of lithium-ion batteries: a multiscale Gaussian process regression modeling approach, AIChE J. 61 (2015) 1589–1600.

[30] Z. Ge, Active learning strategy for smart soft sensor development under a small number of labeled data samples, J. Process Contr. (24) (2014) 1454–1461.

[31] Z. Ge, Active probabilistic sample selection for intelligent soft sensing of industrial processes, Chemometr. Intell. Lab. Syst. 151 (2016) 181–189.

[32] Y. Liu, Q.Y. Wu, J.H. Chen, Active selection of informative data for sequential quality enhancement of soft sensor models with latent variables, Ind. Eng. Chem. Res. 56 (2017) 4804–4817.

[33] L. Zhou, J.H. Chen, Z.H. Song, Recursive Gaussian process regression model for adaptive quality monitoring in batch processes, Math. Probl Eng. (2015) 1–9.