A Gated Recurrent Unit based Echo State Network

Xinjie Wang[1, 2], Yaochu Jin[1, 2, 3] and Kuangrong Hao[1, 2]
[1]Engineering Research Center of Digitized Textile and Fashion Technology, Ministry of Education, Shanghai 201620, P. R. China
[2]College of Information Sciences and Technology, Donghua University, Shanghai 201620, P. R. China
[3]Department of Computing, University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom
Email: wangxinjie0621@foxmail.com, yaochu.jin@surrey.ac.uk, krhao@dhu.edu.cn

Abstract—**Echo State Network (ESN) is a fast and efficient recurrent neural network with a sparsely connected reservoir and a simple linear output layer, which has been widely used for real-world prediction problems. However, the capability of the ESN of handling complex nonlinear problems is limited by the relatively simple neuronal dynamics in the reservoir. Although the gated recurrent unit (GRU) model with multiple nonlinear operators has achieved an excellent performance, gradient-based training algorithms usually require intensive computational resources. In this paper, we present a novel ESN model based on GRUs to tackle complex real-world tasks while reducing the computational costs, taking advantage of the characteristics of both the ESN and the GRU models. In the proposed model, the reservoir unit is replaced by the sparsely connected GRU neurons. Experimental results on three regression problems demonstrate that the proposed method performs better than the original ESN and GRU models.**

Index Terms—**Echo state networks; gated recurrent unit; regression problems.**

I. Introduction

In the past few decades, with continuous improvement of underlying hardware devices and software techniques, a huge amount of streaming data can be generated and collected by using various sensors and actuators [1]. Increasing mature data analysis techniques have made it possible to make use of mass streaming data. Meanwhile, the data-driven modeling approach has been one of the hotspots in current both academy and industry, which has been obtained rapid development and application in the fields of weather forecast [2], air quality prediction [3] as well as prediction and control of industrial process parameters [4].

A popular data-driven modeling technique is artificial neural networks (ANNs), including both feed-forward neural networks (FFNNs) and recurrent neural networks (RNNs) [5], [6]. It has also been shown [7] that RNNs perform better than FFNNs in solving various temporal tasks due to their stronger temporal capabilities and nonlinear properties. Consequently, the RNN was applied to predict the melt-flow-length for mold filling.

However, most gradient-based learning algorithms often suffer from vanishing and exploding gradient problems, deteriorating the performance in training complex ANNs, in particular complex RNNs [8], [9]. The application of gradient-based learning algorithms to real-time industrial processess may also be limited by the heavy computational costs [10]. By contrast, as a biologically plausible and computationally efficient framework of RNNs, echo state networks (ESNs) were proposed to reduce the expensive computational cost in training RNNs [9].

In recent years, several real-life applications based on the ESNs have been reported, such as time series prediction, activity recognition, prediction and control of industrial processes [11], [12]. A canonical ESN can be considered as a three-layer neural network model, a fixed connected input layer, a sparsely connected hidden layer (reservoir), and a readout layer. The original ESN is computationally efficient since only the reservoir-to-readout connection weights need to be trained in while keeping the weights on input-to-reservoir and all weights inside the reservoir fixed. The core of the ESN model is a large-scale sparse connection matrix that transforms the input signals to a high-dimensional feature space. The sparsity in ESNs is loosely inspired by the fact that neurons in the cortex are also sparsely connected, allowing various potential circuits to be generated to encode and process efficiently the internal representations of the external world [13], [14]. In addition, different stimuli generally results in responses from different subsets of neurons [14]. Increasing studies indicated that these sparse representations can improve the system robustness to noise and variability [15], [16].

Theoretically, earlier RNNs methods including the ESNs have the short-term memory property by adopting the internal feedback connections to store information representations about the recent inputs in form of activations [9], [17]. This short-term memory property is also referred to as the fading memory or the echo state property. RNNs with long short-term memory (LSTM) have also been proposed to capture long-term temporal dependences, which significantly improves the prediction performance for problems having large time delays [18], [19]. Later, Cho et al. [20] proposed a gated recurrent unit (GRU) model by means of simplifying the gate units of the LSTM model. Xie et al. [21] presented a two-stream GRU model to predict the melt viscosity index of the real industrial process.

However, training of the LSTM model and its variants is generally based on the gradient descent method such as the backpropagation through time, making the training process computationally very intensive. Moreover, full connections between the neurons in the variants of the LSTM model are biologically implausible.

Based on above discussions, we can see that the reservoir

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

computing framework is able to significantly reduce the computational cost in comparison with some traditional RNNs. On the other hand, ESNs model are not well suited to learning long-term dependencies and tackling complex problems [9], [22]. Inspired by the characteristics of both the ESN and the GRU models, here we present a novel ESN model by replacing the neurons in the reservoir with GRUs to deal with multiple-output regression tasks while reducing the computational costs of LSTM models.

The proposed ESN model based on GRUs is examined on two time-series prediction problems and a parameter estimation problem in an esterification process. Compared with existing the ESN and the GUR models, the proposed method has been demonstrated to exhibit much better prediction performance on all three tasks considered in this work. In addition, the proposed ESN based on GRUs is computationally more efficient than the original GRU model.

The rest of this paper is organized as follows. In Section II, the ESN model and the GRU model are briefly introduced. The GRU-based ESN model is proposed in Section III. Experimental settings and results are provided in Sections IV and V, respectively. Conclusions and future work are provided in Section VI.

II. Related work

A. Echo State Network

The conventional ESN model without output feedback contains an input layer, a reservoir and an output layer, as illustrated in Fig. 1. The dynamic reservoir is used to map the high-dimensional dynamical representations of input signals [23]. The output units implement the linear readout of input representations by using a simple regression algorithm [10].

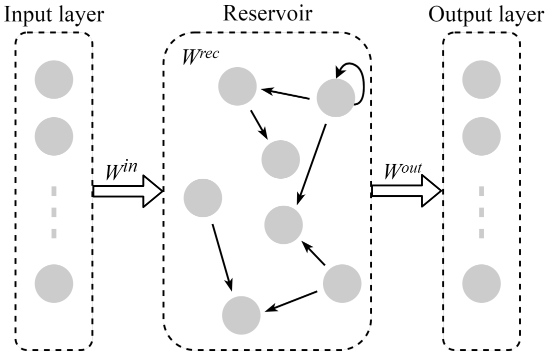

Fig. 1. The basic structure of ESN model.

Let us consider an ESN model with K input units, M reservoir units and L output units. Input connection weights W^{in} and internal recurrent connection weights W^{res} are randomly generated and fixed beforehand. Only connection weights of the readout layer W^{out} need to be trained. The reservoir state equation $x(t)$ and output equation $y(t)$ of the ESN are updated as follows:

$$x(t) = f(W^{in}u(t) + W^{res}x(t-1)), \tag{1}$$

$$y(t) = W^{out}x(t) \tag{2}$$

where, $u(t+1)$ is the external input signal, and the sigmoid function f is used as activation function of reservoir neurons. In this work, the following normalized root mean square error ($NRMSE$) is adopted to be the evaluation criterion of the network:

$$NRMSE(W^{out}) = \sqrt{\frac{\sum_{t=1}^{T}(y^{desired}(t) - y(t))^2}{T\sigma^2}}, \tag{3}$$

where, σ^2 denotes the variance of the desired outputs. T is the number of training samples. The internal state and desired output vectors of the network are stored in X and Y over time t=1, 2, ..., T, respectively. The calculation formula of the readout weights is given as follows:

$$W^{out} = (X^{\mathrm{T}}X)^{-1}X \cdot Y, \tag{4}$$

B. Gated Recurrent Unit

The GRU network, a simplified variant of the LSTM architecture, can also represent context information by storing previous inputs. Fig. 2 shows the architecture of GRU model.

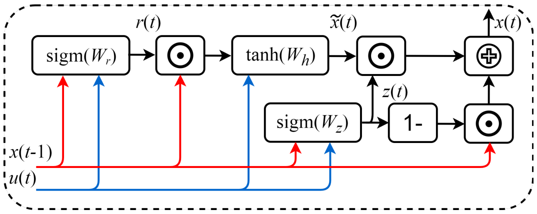

Fig. 2. The architecture of GRU model.

Indeed, the GRU couples the input gate and the forget gate of the initial LSTM into an update gate z, which makes the output nonlinearity less important [19].

The update gate z controls how much previous state information will be stored in the current hidden state x. The reset gate r controls what information will be thrown away from the state information of the previous moment, and the update rules of the GRU model are described in (5) to (8):

$$r(t) = sigm(W_{ru}u(t) + W_{rx}x(t-1) + b_r), \tag{5}$$

$$z(t) = sigm(W_{zu}u(t) + W_{zx}x(t-1) + b_z), \tag{6}$$

$$\tilde{x}(t) = tanh(W_{xu}u(t) + W_{xx}(r(t) \odot x(t-1)) + b_x), \tag{7}$$

$$x(t) = (1 - z(t)) \odot x(t-1) + z(t) \odot \tilde{x}(t), \tag{8}$$

where $\tilde{x}$ is the candidate state. The feed-forward weights W_{*u} and the recurrent weights W_{*x} are the connection matrices for the current input and the hidden state of the previous moment, respectively. In the above equations, b_* is the bias vectors, $sigm$ and $tanh$ represent the logistic sigmoid function and the hyperbolic tangent function, respectively, and $\odot$ denotes element-wise multiplication.

III. GRU-based ESN model

Inspired by the advantages of both the ESN and the GRU models, a novel ESN model based on GRU is presented to process multiple-output regression tasks and reduce the computational training costs. Fig. 3 shows the architecture of the proposed GRU-based ESN model.

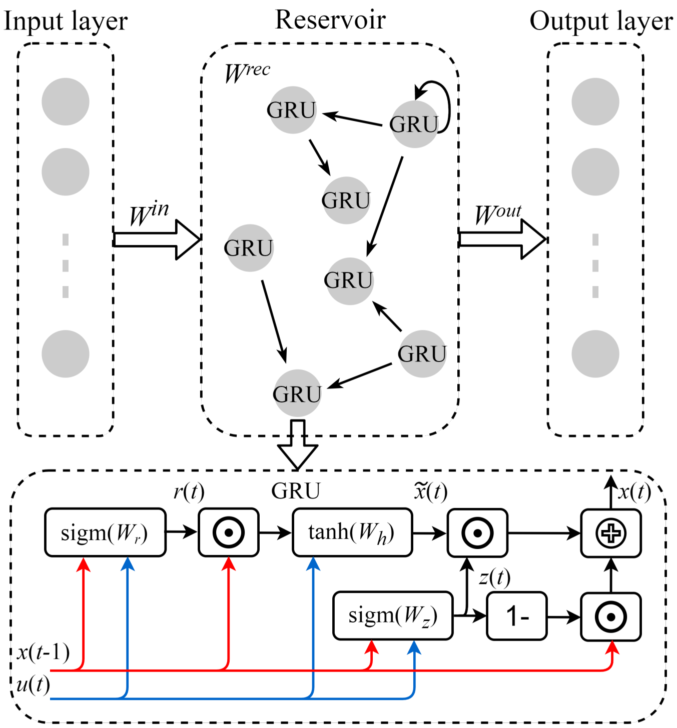

Fig. 3. The architecture of the proposed GRU-based ESN model.

In the ESN model based on GRU, the reservoir is replaced by a large-scale sparsely connected GRUs. Note that the input-reservoir fully connected weights W_{*u} and internal reservoir sparsely connected weights W_{*x} are randomly generated and fixed beforehand. Similarly, only connection weights of read-out layer W_{out} need to be trained by using the least square estimation (LSE) method. The initial state of the model is all set to $r(0)$=$z(0)$=$x(0)$=0.

IV. Experimental settings

A. Datasets

1) Weather forecast: The real-time weather data of the south coast of England is collected from the local sensor network, which can be downloaded from http://www.chimet.co.uk/ [2]. In this work, 52755 samples from August 2018 to January 2019 are used in the following experiments. More concretely, each sample consists of abundant weather information such as wind speed and direction, maximum gust, air temperature, barometric pressure, water depth and wave height. The weather data is used to evaluate the proposed model, the sizes of training dataset, validation dataset, testing data are 30100, 10000, and 12655 respectively, and the first 100 training samples are used to washout.

2) Prediction of Beijing air-quality: This dataset includes hourly air quality data from environmental monitoring sites in Beijing. Removing missing data, 31876 samples from March 2013 to February 2017 are used to evaluate the performance of the proposed method. In this task, the initial 100 training samples are only used to washout, the sizes of training dataset, validation dataset, testing data are set to 10100, 10000 and 11776, respectively [24].

3) Prediction of the esterification process: Polymerization process is the first step in polyester fiber production process, which includes the following three stages: esterification, pre-polycondensation and the final polycondensation. In the esterification stage, purified terephthalic acid (PTA) and excess moles of ethylene glycol (EG) are often used as industrial raw materials. Then, the bis-hydroxyethyl terphthalate (BHET) is produced by the esterification reaction using the mixture of two raw materials with certain proportion. Fig. 4 shows the flowchart of esterification process.

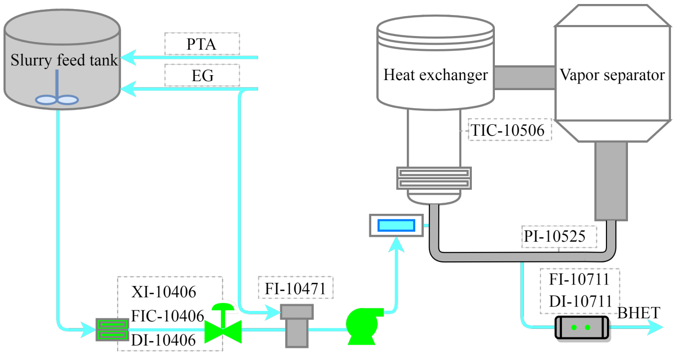

Fig. 4. The flowchart of esterification process.

The esterification stage involves a large number of parameters which can be acquired from different sensors installed in the polyester plant. In this work, eight important process parameters of the esterification process are to be predicted. Table I lists the parameters in esterification stage considered in this work.

TABLE I
Process parameters in esterification stage

No.	Tag Name	Description
1	FIC-10406	Injection flow of esterification
2	DI-10406	Injection density of esterification
3	FI-14701	Injection flow of EG
4	XI-10406	Mole ratio of PTA to EG
5	PI-10525	Pressure of siphon
6	TIC-10506	Temperature of esterification process
7	FI-10711	Flow of oligomer
8	DI-10711	Density of oligomer

Experimental data of esterification stage is collected from the distributed control system with tangible and hardware sensors of the polyester fiber plant in China. Sampling frequency of the sensors is 1 Hz. According to the expert experience and mechanism analysis, eight important parameters of the process are collected, including pressures, temperatures and injection flows. In this dataset, 40900 samples are collected to single-step prediction, of which the sizes of training dataset, validation dataset, testing data are 20100, 10000, and 10800, respectively, and the first 100 training samples are used to washout the initial transient of the network.

B. Experimental Setup

For the setting of connection weights, input-reservoir fully connected weights W_{*u} and internal reservoir sparsely connected weights W_{*x} are randomly generated within the interval of [-1, 1]. Typically, input weights also need to be scaled resulting the activation function works in the linear region, which is between -0.1 and 0.1. The sparsity is set to 0.01.

For the setting of reservoir size, the performance of networks with different reservoir sizes is compared before building the model. Fig. 5 shows the total training and testing errors of eight process parameters on the esterification task by using the ESN with different reservoir sizes. In this work, reservoir size is set to 100. In addition, the size of the hidden layer of original GRU model is set to 60, resulting in the best performance compared to other sizes. We trained all networks with Nvidia GeForce RTX2060 GPU. In addition, we compare the performance of the reservoir with different values in sparsity to determine the sparsity parameter. From Fig. 6 we can see that the ESN can obtain the minimum testing error, when sparsity is set to 0.01.

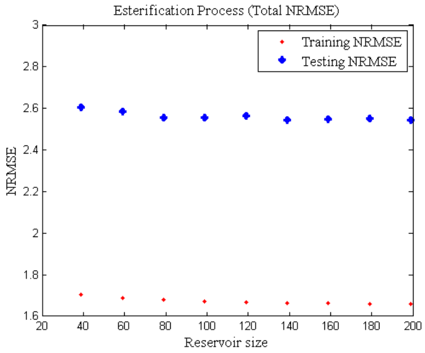

Fig. 5. Effects of reservoir sizes on the prediction performance on process parameters in esterification stage.

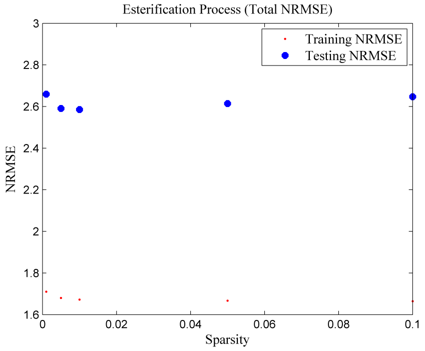

Fig. 6. Effects of sparsity on the prediction performance on process parameters in esterification stage.

V. Experimental results

In this section, the ESN model with GRU is designed to predict the real-world multiple-output benchmark regression tasks.

A. Prediction of Esterification Process

Fig. 6 shows the prediction results of eight process parameters in esterification stage. As Fig. 6 illustrates, the proposed method can effectively forecast some important process parameters. For parameters prediction of the esterification process, the proposed method is compared with original ESN and GRU models. The mean values of the testing NRMSE are listed in Table II. From Table II we can see that the GRU-based ESN model has better prediction performance compared to the ESN and GRU. Note that, although the gradient-based GRU model can obtain the best performance in a few of those process parameters, the total testing error is still large, and above methods are run independently 30 times.

On the other hand, the training time of the GRU-based ESN model is significantly lower than the gradient-based GRU model. It also should be noted that the training time of the GRU-based ESN model is slightly higher than the original ESN model because of its multiple gate units.

TABLE II
THE PERFORMANCE OF DIFFERENT METHODS ON THE ESTERIFICATION PROCESS

NRMSE,Time Model Index	ESN	GRU	GRU-based ESN
FIC-10406	0.4284	**0.2658**	0.4183
DI-10406	0.2140	**0.1607**	0.1991
FI-14701	0.3842	**0.1663**	0.3799
XI-10406	0.3901	0.7545	**0.3457**
PI-10525	0.3168	0.6052	**0.2994**
TIC-10506	0.2158	**0.1479**	0.2022
FI-10711	0.2013	0.2569	**0.1959**
DI-10711	0.4448	0.5087	**0.4202**
Total error	2.5854	2.8663	**2.4607**
Training time	**1.317**s	104.92s	2.679s

B. Weather Forecast Dataset

Two real-world datasets, England weather dataset and Beijing air-quality dataset, are used to further substantiate the effectiveness of the GRU-based ESN model. For weather forecast task, the prediction results of ten dimensional weather data are presented in Fig. 7. The actual output of GRU-based ESN model could fit well with the real meteorological data. In addition, the GRU-based ESN model is compared with the ESN and GRU models on the weather dataset. 30 independent simulations have been conducted and the averaged testing NRMSEs of different methods are listed in Table III. Similarly, the GRU-based ESN model can obtain the minimum total prediction error on testing set compared to the ESN and GRU.

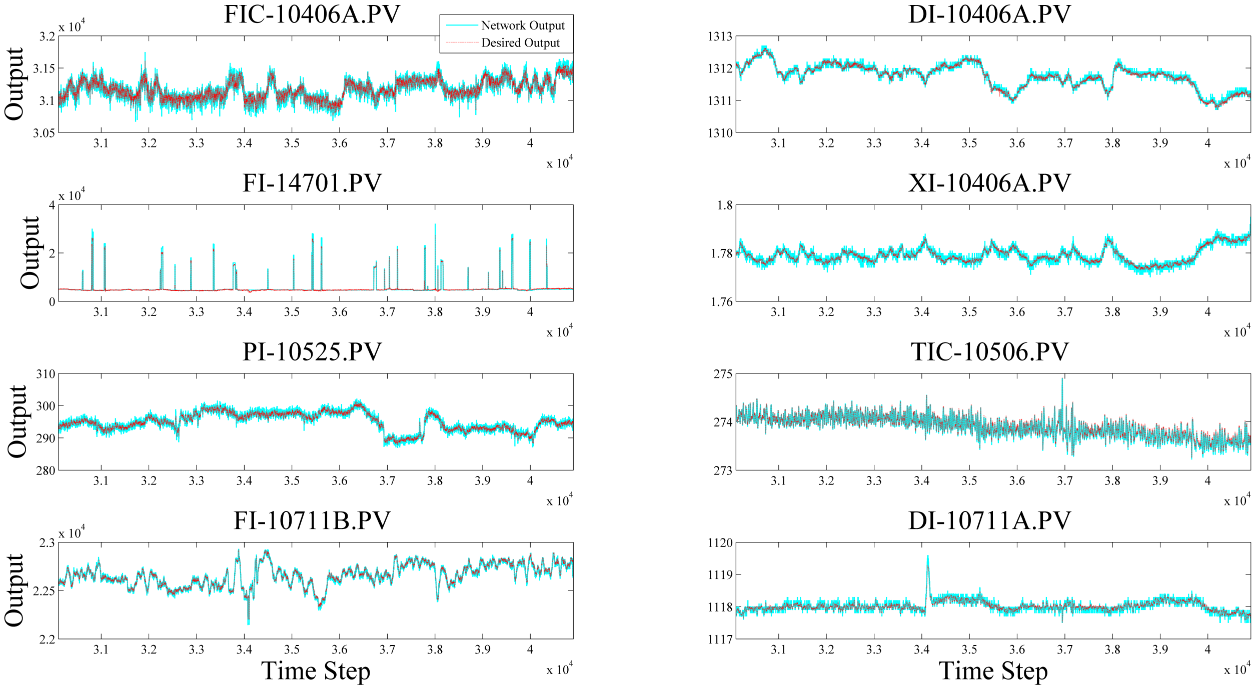

Fig. 7. The prediction results produced by the GRU-based ESN model for eight process parameters in the esterification stage.

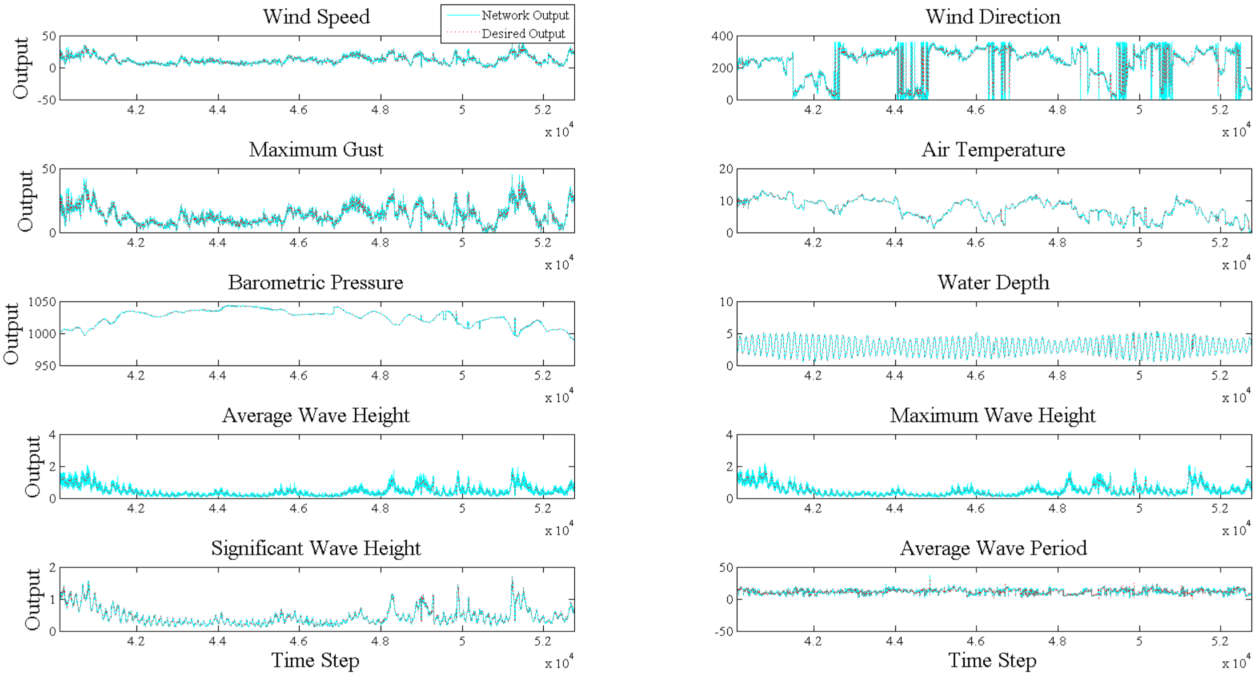

Fig. 8. The prediction results produced by the GRU-based ESN model for the weather forecast dataset.

TABLE III
THE PERFORMANCE OF DIFFERENT METHODS ON THE WEATHER FORECAST DATASET

NRMSE,Time Model Index	ESN	GRU	GRU-based ESN
Wind speed	0.2268	**0.2164**	0.2187
Wind direction	**0.4192**	0.5403	0.4201
Maximum gust	0.2332	0.2526	**0.2234**
Air temperature	0.0591	0.2298	**0.0589**
Barometric pressure	0.0487	0.2636	**0.0484**
Water depth	0.0676	0.2154	**0.0571**
Average wave height	0.3595	**0.2654**	0.3427
Maximum wave height	0.3166	**0.2858**	0.2966
Significant wave height	0.1018	0.1295	**0.1011**
Average wave period	0.5124	**0.1381**	0.4969
Total error	2.3452	2.5374	**2.2459**
Training time	**4.4733**s	207.04s	11.9244s

Obviously, the GRU-based ESN model owns much cheaper computational complexity compared with the gradient-based GRU model. In addition, the proposed model with multiple gate units results in a slightly higher time cost compared with the original ESN model.

C. Prediction of Beijing Air-Quality

For prediction problem of Beijing air-quality, Fig. 8 shows the prediction results of ten air indicators. As Fig. 8 illustrates, the air-quality data could be effectively fitted by the proposed method.

In experiments on prediction task of Beijing air-quality, the proposed method is compared with existing ESN and GRU models, which also demonstrate that the GRU-based ESN model owns minimum total testing error and acceptable training time.

TABLE IV
THE PERFORMANCE OF DIFFERENT METHODS ON PREDICTION TASK OF THE AIR-QUALITY

NRMSE,Time Model Index	ESN	GRU	GRU-based ESN
PM2.5 concentration	0.2330	**0.1951**	0.2327
PM10 concentration	0.3144	0.5426	**0.3124**
SO_2 concentration	0.3370	**0.2130**	0.3298
NO_2 concentration	0.3026	0.3848	**0.2902**
CO concentration	0.2898	**0.2653**	0.2849
O_3 concentration	0.2424	0.2438	**0.2367**
Temperature	0.0879	0.2306	**0.0833**
Pressure	0.0465	0.2146	**0.0463**
Dew point temperature	0.0849	0.1661	**0.0846**
Wind speed	0.5674	**0.2120**	0.5669
Total error	2.5059	2.6683	**2.4678**
Training time	**3.4269**s	142.76s	8.9969s

VI. CONCLUSION

In this paper, an ESN model with GRU is presented to deal with several real-world multiple-output regression problems. The effectiveness of the GRU-based ESN model is verified on three real-world regression tasks and our experimental results demonstrate that the proposed method can predict effectively multiple output indexes compared with the original ESN and GRU models. More important, the GRU-based ESN model can significantly reduce the time complexity compared with the gradient-based method, which is essential to real-time online prediction of streaming data.

In this work, connection weights within the reservoir are randomly generated and fixed, which also limit the network performance. For future work, some unsupervised learning rules can be used to optimize the connection weights and intrinsic excitability of neurons in this ESN model with GRU such as synaptic plasticity and intrinsic plasticity learning rules.

ACKNOWLEDGMENT

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant (nos. CUSF-DH-D-2018101, 2232017D-08), in part by the National Natural Science Foundation of China (nos. 61806051, 61903078), Shanghai Sailing Program (no. 17YF1426100), and Agricultural Project of the Shanghai Committee of Science and Technology (no. 16391902800). The authors would like to thank R. Xie, D. Zhao, Z. Li, W. Hu and X. Zhu for their helpful discussions and providing the experimental data.

REFERENCES

[1] Shen Yin, Xianwei Li, Huijun Gao, and Okyay Kaynak. Data-based techniques focused on modern industry: An overview. *IEEE Transactions on Industrial Electronics*, 62(1):657–667, 2014.

[2] Changsheng Li, Fan Wei, Weishan Dong, Xiangfeng Wang, Qingshan Liu, and Xin Zhang. Dynamic structure embedded online multiple-output regression for streaming data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):323–336, 2018.

[3] Ke Gu, Junfei Qiao, and Weisi Lin. Recurrent air quality predictor based on meteorology-and pollution-related factors. *IEEE Transactions on Industrial Informatics*, 14(9):3946–3955, 2018.

[4] Zhiqiang Ge. Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemometrics and Intelligent Laboratory Systems*, 171:16–25, 2017.

[5] Le Yao and Zhiqiang Ge. Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application. *IEEE Transactions on Industrial Electronics*, 65(2):1490–1498, 2017.

[6] In-Su Han, Chonghun Han, and Chang-Bock Chung. Melt index modeling with support vector machines, partial least squares, and artificial neural networks. *Journal of Applied Polymer Science*, 95(4):967–974, 2005.

[7] Xi Chen, Furong Gao, and Guohua Chen. A soft-sensor development for melt-flow-length measurement during injection mold filling. *Materials Science and Engineering: A*, 384(1-2):245–254, 2004.

[8] Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[9] Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

[10] Benjamin Schrauwen, Marion Wardermann, David Verstraeten, Jochen J Steil, and Dirk Stroobandt. Improving reservoirs using intrinsic plasticity. *Neurocomputing*, 71(7-9):1159–1171, 2008.

[11] Ying Liu, Quanli Liu, Wei Wang, Jun Zhao, and Henry Leung. Data-driven based model for flow prediction of steam system in steel industry. *Information Sciences*, 193:104–114, 2012.

[12] Michael Buehner and Peter Young. A tighter bound for the echo state property. *IEEE Transactions on Neural Networks*, 17(3):820–824, 2006.

[13] Dmitri B Chklovskii, BW Mel, and K Svoboda. Cortical rewiring and information storage. *Nature*, 431(7010):782, 2004.

[14] Friedemann Zenke and Wulfram Gerstner. Hebbian plasticity requires compensatory processes on multiple timescales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1715):20160259, 2017.

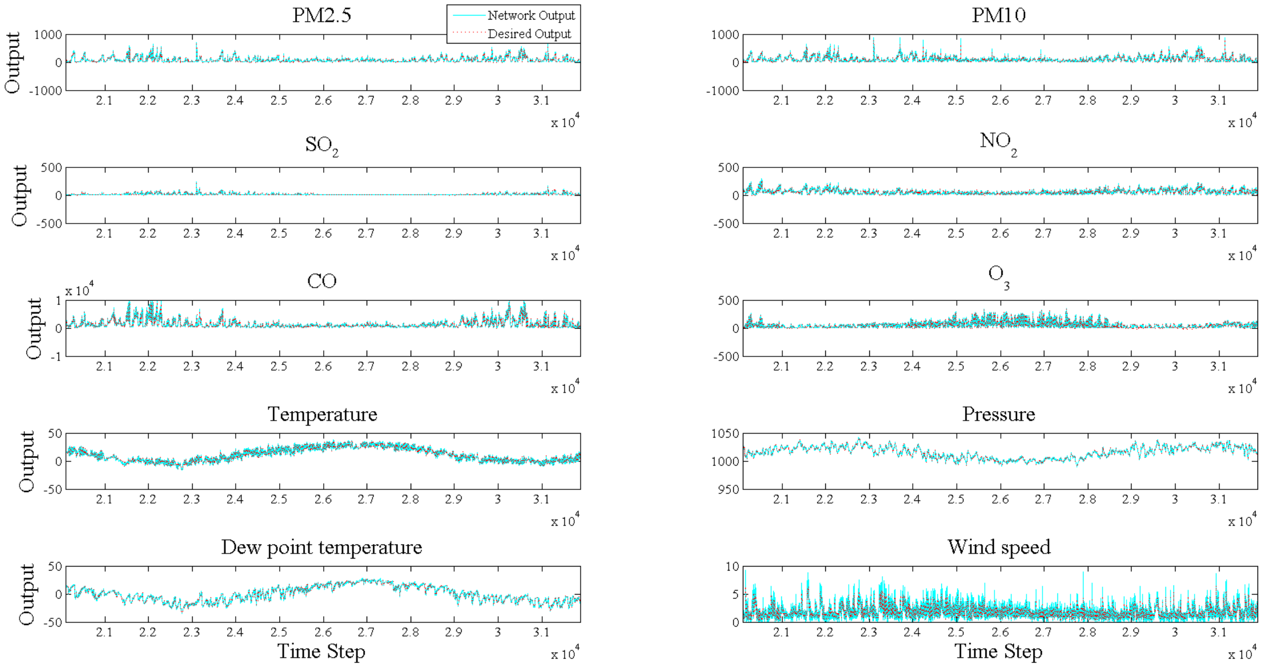

Fig. 9. The prediction results produced by the GRU-based ESN model for the prediction of the air-quality.

[15] Yuwei Cui, Subutai Ahmad, and Jeff Hawkins. The HTM spatial poolera neocortical algorithm for online sparse distributed coding. *Frontiers in Computational Neuroscience*, 11:111, 2017.

[16] Jeff Hawkins and Subutai Ahmad. Why neurons have thousands of synapses, a theory of sequence memory in neocortex. *Frontiers in Neural Circuits*, 10:23, 2016.

[17] Michiel Hermans and Benjamin Schrauwen. Memory in reservoirs for high dimensional input. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2010.

[18] Felix A Gers and E Schmidhuber. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6):1333–1340, 2001.

[19] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2016.

[20] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[21] Ruimin Xie, Kuangrong Hao, Biao Huang, Lei Chen, and Xin Cai. Data-driven modeling based on two-stream λ gated recurrent unit network with soft sensor application. *IEEE Transactions on Industrial Electronics*, 2019.

[22] Haibin Duan and Xiaohua Wang. Echo state networks with orthogonal pigeon-inspired optimization for image restoration. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2413–2425, 2015.

[23] Herbert Jaeger. *Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach*, volume 5. GMD-Forschungszentrum Informationstechnik Bonn, 2002.

[24] Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.