Energy Conversion and Management 198 (2019) 111772

Contents lists available at ScienceDirect

Energy Conversion and Management

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

Gaussian mixture model coupled recurrent neural networks for wind speed interval forecast

Shuang Zhu[a], Xiaohui Yuan[b], Zhanya Xu[a], Xiangang Luo[a,*], Hairong Zhang[c]

[a] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China
[b] Department of Computer Science and Engineering, University of North Texas, Denton, TX 76210, USA
[c] Department of Water Resources Management, China Yangtze Power Company Limited, Yichang 443133, China

ARTICLE INFO

Keywords:
Wind speed
Long short-term memory network
Gaussian mixture model
Feature selection

ABSTRACT

The potential of long short-term memory network on ultra-short term wind speed forecast attracted attentions of researchers in recent years. Extending a probabilistic long short-term memory network model to provide an uncertainty estimation than to make a point forecast is more valuable in practice. However, due to complex recurrent structure and feedback algorithm, large scale ensemble forecast based on resampling faces great challenges in reality. Instead, a reliable forecast method needs to be devised. Gaussian process regression is a probabilistic regression model based on Gaussian Process prior. It is reasonable to integrate Gaussian process regression with long short-term memory network for probabilistic wind speed forecast to leverage the superior fitting ability of the deep learning methods and to maintain the probability characteristics of Gaussian process regression. Hence, avoid the repeated training and heavy parameter optimization. The method is evaluated for wind speed forecast using the monitoring dataset provided by the National Wind Energy Technology Center. The results indicated that the proposed method improves the point forecast accuracy by up to 17.2%, and improves the interval forecast accuracy by up to 18.5% compared to state-of-the-art models. This study is of great significance for improving the accuracy and reliability of wind speed prediction and the sustainable development of new energy sources.

1. Introduction

Large-scale wind energy development has become one of the important strategies to solve energy and environmental problems [1]. Physical and statistical models have been used for wind speed prediction [2]. The physical wind speed prediction is to establish a set of real-time meteorological fluid mechanics and thermodynamic equations [3]. It provides in-depth wind speed investigation within the atmospheric cycle. The disadvantage is the time-space complexity and model biases that cannot be ignored. The statistical wind speed prediction includes multiple linear regression, time series analysis, fuzzy clustering, and artificial neural networks [4]. It is well acknowledged that statistical wind speed prediction is convenient, effective, and highly accurate. In recent years, inspired by the rapid development and successful application of deep learning in human perceptions, image classification, and environmental simulation, researchers have begun to introduce the various deep network to wind speed prediction. Wang et al. [5] first used a deep belief network for wind speed prediction using real wind farm data from China and Australia and obtained competitive performance. Zhang et al. [6] presented deep Boltzmann machine technique for short-term and long-term wind speed forecast, proved neural networks with deep architectures having the competitive capability to approximate nonlinear and non-smooth wind speed problem. Yu et al. [7] adopted recurrent neural networks to extract deeper features, it is demonstrated that the accuracy of deep learning prediction outperforms existing methods. Long short-term memory (LSTM) is an improved recurrent neural network architecture capable of solving the vanishing gradient problem. Nowadays, a cluster of LSTMs with diverse input features, hidden layers, and neurons have been introduced to explore many aspects of wind speed prediction, such as LSTM temporal feature extraction [8], original and decomposition wind speed sequence prediction [9], and nonlinear combination of multiple results [10]. The competitive performance and high-stability of LSTM wind speed forecast are proved in these researches.

However, climate variables, including wind speed, are plagued by uncertainties [11], thus making probabilistic wind speed prediction an appealing option. Generalized likelihood uncertainty estimation (GLUE) [12], bootstrap [13] and multi-objective optimization [14] are

[*] Corresponding author.
E-mail address: billlxg@126.com (X. Luo).

https://doi.org/10.1016/j.enconman.2019.06.083
Received 11 April 2019; Received in revised form 28 June 2019; Accepted 28 June 2019
Available online 15 July 2019

Nomenclature

LSTM	long short-term memory network
NWTC	national wind energy technology center
GLUE	generalized likelihood uncertainty estimation
CNNs	conventional neural networks
TDRF	top-down relevant feature search
TGPLSTM	hybrid model of TDRG, GPR and LSTM
RMSE	root mean squared error
RSE	relative squared error
SS_{CRPS}	forecast skill score based on CRPS
CRPS	continuous ranked probability score
T	temperature
RH	relative humidity
SP	station pressure
AP	accumulated precipitation
MLP	multilayer perceptron
GPR	Gaussian process regression
LSTMs	long short-term memory networks
GP	Gaussian process
MI	mutual information
RNN	recurrent neural network
ANN	artificial neural network
MAE	mean absolute error
PIT	probability in truth
AWS	average wind speed
PWS	peak wind speed
DPT	dew point temperature
SH	specific humidity
SLP	sea-level pressure
PCC	Pearson correlation coefficient
GLM	generalized linear mode

main techniques for traditional probabilistic prediction. These methods generate a large number of initial parameter sets by random sampling algorithm and then construct corresponding models based on different initializations to obtain predicted intervals [15]. Currently, only very few studies made probabilistic wind speed forecast within a deep learning framework. Wang et al. [16] successfully estimated forecast uncertainty via an ensemble of twenty-four individual deep conventional neural networks (CNNs) that have different numbers of hidden layers and neurons. To overcome the defects of linear representation of combined models, Chen et al. [17] used support vector regression machine to integrate six diverse LSTMs forecasts and obtained a probabilistic interval of wind speed forecasts. Their researches provide a meaningful reference for probabilistic deep learning wind speed prediction. It is worth mentioning that the ensemble forecast has better reliability when the number of individuals is large enough, usually, thousands of times [18]. However, large scale ensemble forecast based on LSTM individuals is hard, due to the fact that LSTM handles sequence dependent through complex recurrent structure and feedback algorithm, which is complicated and time-consuming. Instead of the ensemble method, Wu et al. [19] and Zhang et al. [20] provided another solution to obtain probabilistic LSTM wind speed forecast by analyzing the distribution characteristics of point forecast errors. Essentially, it is a post-processing analysis technique implemented on the deterministic forecast. Thus, a more direct, reliable and probabilistic LSTM wind speed prediction research needs to be extended.

Gaussian Process Regression (GPR) [21] is a probabilistic regression model based on Gaussian Process (GP) prior. A posterior of the predicted result is derived through the joint conditional probability distribution of target variables and predictors with historical data. The kernel function is introduced to solve the covariance matrix required for the prediction distribution in high-dimensional space and avoid the complicated computation. With the convenience properties of GPR, it is reasonable to couple GPR into the internal structure of the LSTM for probabilistic wind speed forecast. The hybrid model is probabilistic in nature. It has the strong fitting ability of deep learning, and maintains the probability characteristics of GPR, and avoids the repeated training and heavy parameter optimization tasks in ensemble technique. In addition, since the input variables and training samples have a great

Algorithm 1: Boruta algorithm

```
Input: OriginalData - input dataset; RFruns - the number of random forest runs.
for each RFruns do
    $shadowAttr \leftarrow permute(originalPredictors)$
    $extendedData \leftarrow cbind(originalPredictors, shadowAttr)$
    $zScoreSet \leftarrow randomForest(extendedData)$
    $MZSA \leftarrow max(zScoreSet(shadowAttr))$
    for each $a \in originalPredictors$ do
        if $zScoreSet(a) > MZSA$ then
            $hit(a)++$
        end
    end
end
for each $a \in originalPredictors$ do
    $significance(a) \leftarrow twoSidedEqualityTest(a)$
    if $significance(a) \gg MZSA$ then
        $confirmedSet \leftarrow finalSet \cup a$
    end
    else if $significance(a) \ll MZSA$ then
        $rejectedSet \leftarrow rejectedSet \cup a$
    end
end
return $finalSet \leftarrow rejectedSet \cup confirmedSet$
```

influence on the prediction accuracy, the choice of features is another concerned problem in this paper. Pearson correlation coefficient (PCC), mutual information (MI) are widely used for feature selection and extraction [9]. PCC is suitable for the linear correlation measurement of two independent variables with a normal distribution. MI refers to the amount of information that one random variable contains about the other. However, the above methods are not able to deal with the issue of redundant features. To eliminate irrelevant as well as redundant features, top-down relevant feature search (TDRF), which is a ranking algorithm based on the random forest to estimate the importance of a feature, is used to determine the final inputs of the prediction model. Therefore, the innovations of this paper mainly include the following two points: 1) an improved deep learning network for directly probabilistic wind speed prediction; 2) seeking an intelligent search strategy to determine the appropriate input factors for the wind speed prediction model.

The remainder of this paper is organized as follows. The proposed model for probabilistic wind speed forecast is explained in Section 2. The evaluation indicators are presented in Section 3. Section 4 presents the results and performance analysis. Section 5 concludes the paper with a summary.

2. Probabilistic long short-term memory network

A probabilistic long short-term memory network for wind speed forecast is developed in this paper. The top-down relevant feature search algorithm (TDRF) is applied to determine the model inputs, GPR is introduced into the internal structure of the LSTM network to construct a hybrid model, a semi-stochastic alternating gradient descent optimization procedure is applied to carry out weight updates and fully joint training. Within the proposed strategy, predictor sets are mapped into a single vector in the hidden space, then continued to be projected into a high dimensional feature space with the kernel function, the output layer of the hybrid model naturally produces the posterior distribution over target. Detailed description of TDRF algorithm, LSTM, GPR for probabilistic forecast, probabilistic LSTM model and parameter optimization are introduced.

2.1. Top-down relevant feature selection

Determining appropriate feature factors plays an important role in training neural network models. In general, the more features select, the longer it takes to train the model, and the redundancy dependencies between features can easily lead to a reduction in generalization capabilities. Top-down related feature selection algorithms can be used to build multiple models with different subsets and identify the attributes needed to build an accurate prediction model [22]. It is a kind of ranking algorithm based on the random forest to estimate the importance of the feature. The pseudocode of top-down relevant feature selection is shown as follows [23].

2.2. Long short-term memory network

Recurrent neural network (RNN) is a circular artificial neural network (ANN) where additional input is added to represent the updated state of the hidden neuron [24]. It takes into account the current information as well as other adjoining information in the data.

Fig. 1 shows the repeated module in a recurrent network with a single layer. It contains an input of the state of the neuron in the hidden layer at the previous time steps [24]. The state of the neuron at the current time step h_t is computed as follows:

$$h_t = tanh(W_h \cdot [h_{t-1}, X_t] + b_h), \tag{1}$$

where h_t, h_{t-1} represent the hidden neuron states at the time step t and $t-1$ respectively, W_h means weights of input and hidden neurons, $tanh(\cdot)$ is element-wise hyperbolic tangent function.

The RNN model learns the target by using linear parametric maps followed by nonlinear activations

$$y = \psi(W_{hy}^T h_{t-1}) \tag{2}$$

where ψ is fixed element-wise function, W_{hy}^T is weights of output layer.

In a standard RNN trained on long sequences (e.g. 100 time-steps), the gradients can easily explode or vanish, since the error of partial derivative accumulates through time steps. Long short-term memory (LSTM) has been proposed and refined to solve the vanishing/exploding gradient problem [25]. The repeating module of LSTM is different from RNN, it has four neural network layers interacting in a very special way. With a memory cell and three gates, the LSTM has the ability to update the information to the cell state based on the new input and forget irrelevant content. Suppose p_t and q_t are two control gates, C_{t-1} is the old state of the network, C_t' is the updated information of the network, p_t is old messages removed, and q_t indicates new messages added. The neuron state at the current time step h_t is updated based on the Eqs. (3)–(8).

Forget gate

$$p_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \tag{3}$$

Input gate

$$q_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \tag{4}$$

Memory cell

$$C_t' = tanh(W_C \cdot [h_{t-1}, X_t] + b_C) \tag{5}$$

$$C_t = p_t \cdot C_{t-1} + q_t \cdot C_t' \tag{6}$$

Output gate

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \tag{7}$$

$$h_t = o_t * tanh(C_t) \tag{8}$$

where $\sigma(\cdot)$ is a element-wise sigmoid function.

LSTM extends RNN by including an oblivion gate to decide if the

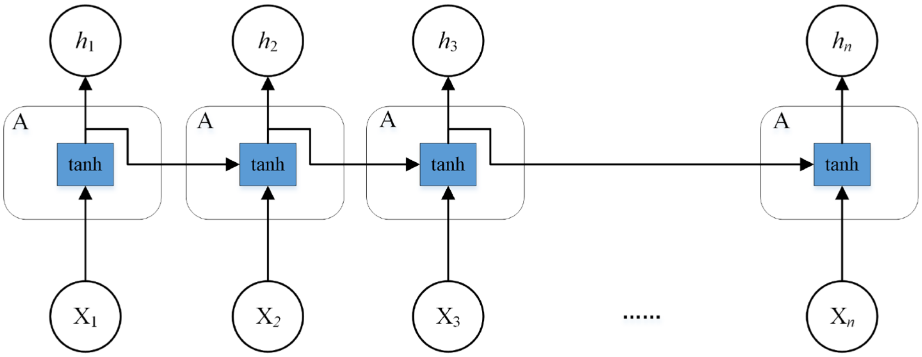

Fig. 1. The repeated module in a RNN with a single layer.

Algorithm 2: Semi-stochastic alternating gradient descent

```
Input: Data - (X,y),kernel-$k_{\theta}(\cdot,\cdot)$,recurrent transformation-$\phi_{\mathbf{w}}(\cdot)$.
Initialize $\boldsymbol{\theta}$ and $\mathbf{w}$;compute initial $K$.
repeat
    for all mini-batches $X_b$ in $X$ do
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + update_{\boldsymbol{\theta}}(X, \boldsymbol{\theta}, K)$.and
        $\mathbf{w} \leftarrow \mathbf{w} + update_{\mathbf{w}}(X_b, \mathbf{w}, K)$.
        Update the kernel matrix,K.
    end
until Convergence;
Output: Optimal $\boldsymbol{\theta}^*$ and $\mathbf{w}^*$
```

information is consistent and useful. The advantage of exhibiting temporal dynamic behavior is very suitable for drought recurrence mode in drought-prone areas.

2.3. Gaussian process regression for probabilistic forecast

Gaussian process regression (GPR) is firstly proposed by Gibbs and Mark [26] and lately extended by Kersting et al. [21] and Tolvanen et al. [27]. The key idea of GPR is to assume that the learning sample follows the prior probabilities of the Gaussian process and then calculate the corresponding posterior probability. It is developed based on the Bayesian regression model. Due to the advantage of handling uncertainty, it has begun to be applied for probabilistic analysis in the last few years [28].

Simply, given an input vector X_* , the Gaussian posterior distribution of the predicted value y_* is

$$p(y_*|X_*, X, y) = N\left(\frac{1}{\sigma^2}X_*^T A^{-1} X y, X_*^T A^{-1} X_*\right) \tag{9}$$

where $A = \sigma^{-2}XX^T + \sum_p^{-1}$, $\sum_p$ is the covariance matrix.

Function ϕ is introduced to project the input X into a high dimensional feature space to implement polynomial regression. The Gaussian posterior is as follows:

$$p(y_*|X_*, X, y) = N\left(\frac{1}{\sigma^2}\phi(X_*^T)A^{-1}\phi(X)y, \phi(X_*^T)A^{-1}\phi(X_*)\right) \tag{10}$$

where $A = \sigma^{-2}XX^T + \sum_p^{-1}$, $\sum_p$ is covariance matrix.

To make predictions, A matrix needs to be inverted with a kernel function. The probability distribution is

$$\begin{aligned} &p(y_*|X_*, X, y) \\ &= N(K(X_*, X)(K + \sigma^2 I_n)^{-1}y, K(X_*, X_*) - K(X_*, X)(K + \sigma^2 I_n)^{-1}K \\ &\quad (X, X_*)) \end{aligned} \tag{11}$$

where $A = \phi^T \sum_p \phi$, K is kernel function.

Assume that the training samples follow a normal distribution with a zero mean, a unique GPR model is obtained with the kernel function K.

2.4. Probabilistic long short-term memory network

For the probabilistic long short-term memory network framework, GPR is embedded as an internal unit of the LSTM network. The input time series are mapped into a single vector in the recurrently updated hidden space, H, which is projected into a high dimensional feature space using the kernel function K and modeled with the GPR layer. The output layer of LSTM produces the posterior distribution over target.

Combining Eq. (2) and Eq. (11), the Gaussian posterior distribution of the hybrid model over target becomes

$$\begin{aligned} &p(y_*|X_*, X, y) \\ &= N(K(h(X_*), h(X))(K + \sigma^2 I_n)^{-1}y, K(h(X_*), h(X_*)) - K(h(X_*), h(X)) \\ &\quad (K + \sigma^2 I_n)^{-1}K(h(X), h(X_*))) \end{aligned} \tag{12}$$

where h is the state of hidden neuron.

In LSTM model, the hidden layer of the deep network is directly followed by the output layer. In the proposed model, the output of the hidden layer is used as the input of the GPR layer followed by the probabilistic kernel function regression prediction.

The expectation is point result and the uncertainty interval is derived with mean and standard deviation.

Point forecast:

$$\mu = K(h(X_*), h(X))(K + \sigma^2 I_n)^{-1}y \tag{13}$$

95% confidence interval:

$$Range = [\mu - z_{\alpha/2}\sigma, \mu + z_{\alpha/2}\sigma] \tag{14}$$

where $\alpha = 0.05$, $\sigma = K(h(X_*), h(X_*)) - K(h(X_*), h(X))(K + \sigma^2 I_n)^{-1}K.$ $(h(X), h(X_*))$

The conditional probability distribution function given in Eq. (11) is the full forecast probability density of the hybrid model. The probability forecast score and the reliability are calculated according to the conditional probability distribution function.

2.5. Parameter optimization

The parameters of the hybrid model include weight W of the LSTM

Table 1
The statistics of minute data from April 1 to April 30, 2015.

Variable	Mean	Min	p1	p5	p25	p50	p75	p95	p99	Max
AWS (m/s)	4.04	0.34	0.34	0.78	2.25	3.48	5.17	9.35	12.94	25.03
PWS (m/s)	4.74	0.34	0.34	1.06	2.73	4.11	5.92	10.88	15.25	28.07
T (°C)	8.25	−4.42	−3.98	−0.44	3.43	8.74	12.8	17.49	19.8	22.05
DPT (°C)	−1.1	−14.64	−12.37	−8.09	−3.61	−0.72	2.02	4.48	5.91	9.3
RH (%)	58.74	10.24	13.27	17.41	34.2	55.13	85.37	100	100	100
SH (%)	4.42	1.32	1.61	2.37	3.46	4.42	5.41	6.44	7.16	8.94
SP (mBar)	811.31	799.86	802.52	804.06	807.28	811.1	815.01	819.96	822.09	823.14
SLP (mBar)	1015.6	1001.9	1005.1	1006.9	1010.8	1015.4	1020.1	1026	1028.5	1029.8
Prec (mm)	1.45	0	0	0	0	0	0	13.21	25.15	25.65

Table 2
The statistics of minute data from October 13 to November 13, 2015.

Variable	Mean	Min	p1	p5	p25	p50	p75	p95	p99	Max
AWS (m/s)	4.57	0.31	0.31	0.54	1.86	3.31	5.98	12.47	21.18	31.45
PWS (m/s)	5.43	0.31	0.31	0.73	2.25	3.89	6.93	15.04	25.69	36.45
T(°C)	10.33	−2.09	−1.13	0.79	5.49	10.64	14.65	19.72	23.43	24.46
DPT (°C)	−1.71	−12.95	−11.51	−9.37	−5.2	−1.62	1.82	6.09	7.29	8.67
RH (%)	49.65	11.73	13.73	17.17	32.13	43.03	61.44	100	100	100
SH (%)	4.26	1.51	1.74	2.1	3.01	4.08	5.32	7.15	7.78	8.57
SP (mBar)	813.79	803.07	803.56	805.41	809.83	813.52	818.14	821.5	824.98	825.98
SLP (mBar)	1018.58	1005.7	1006.3	1008.5	1013.8	1018.3	1023.8	1027.8	1032	1033.2
Prec (mm)	1.09	0	0	0	0	0	0	12.7	14.99	15.24

Table 3
The original historical features and their feature numbers.

Feature type	Historical features	Numbers
AWS	$AWS_{t-1}, AWS_{t-2}, AWS_{t-3}, \cdots, AWS_{t-15}$	1–15
T	$T_{t-1}, T_{t-2}, T_{t-3}, \cdots, T_{t-15}$	16–30
DPT	$DPT_{t-1}, DPT_{t-2}, DPT_{t-3}, \cdots, DPT_{t-15}$	31–45
RH (%)	$RH_{t-1}, RH_{t-2}, RH_{t-3}, \cdots, RH_{t-15}$	46–60
SH (%)	$SH_{t-1}, SH_{t-2}, SH_{t-3}, \cdots, SH_{t-15}$	61–75
SP	$SP_{t-1}, SP_{t-2}, SP_{t-3}, \cdots, SP_{t-15}$	76–90
SLP	$SLP_{t-1}, SLP_{t-2}, SLP_{t-3}, \cdots, SLP_{t-15}$	91–105
AP	$AP_{t-1}, AP_{t-2}, AP_{t-3}, \cdots, AP_{t-15}$	106–120
PWS	$PWS_{t-1}, PWS_{t-2}, PWS_{t-3}, \cdots, PWS_{t-15}$	121–135

Table 4
The thirty most important features.

Methods	1–10 Dimensional Features	11–20 Dimensional Features	21–30 Dimensional Features
TDRF	121, 1, 122, 2, 3 123 124, 125, 94, 4	23, 92, 27, 78, 106 24, 41, 80, 25, 126	91, 60, 72, 74, 28 67,89, 54, 56, 45
PCC	1, 121, 122, 2, 123 3, 124, 4, 125, 5	126, 6, 127,7, 128 8, 129, 9, 130, 10	131, 11, 132, 12, 133 13,105, 90, 89, 104

Table 5
Statistics of the forecast skill involving point forecast and interval forecast, dataset one.

Station	Model	RMSE	RSE	MAE	SS_{CRPS}
One step ahead	TGPLSTM	0.90	0.24	0.58	0.82
	MLP	0.91	0.25	0.57	–
	GLM	1.07	0.29	0.63	0.81
Two steps ahead	TGPLSTM	1.25	0.58	0.82	0.78
	MLP	1.27	0.61	0.86	–
	GLM	1.33	0.64	0.91	0.71
Three steps ahead	TGPLSTM	1.49	0.95	1.01	0.77
	MLP	1.51	1.05	1.07	–
	GLM	1.56	1.09	1.12	0.65
Four steps ahead	TGPLSTM	1.67	1.36	1.13	0.73
	MLP	1.69	1.61	1.22	–
	GLM	1.73	1.65	1.26	0.62

structure and the kernel hyperparameters θ of GPR structure. For the nested hybrid feed-forward network, the optimization is difficult. A semi-stochastic alternating gradient descent optimization procedure proposed by Al-Shedivat et al. [29] is adopted for training, which alternately updates θ and W on the mini-batch dataset using stochastic steps. Instead of using all training examples, i.e., full-batch, or a single training example (as in stochastic training), mini-batch training uses a subset of training examples to compute gradients, which achieves a faster convergence [30]. The pseudocode of algorithm is shown below. The kernel matrix ϕ of GPR layer is computed with initial W and θ parameters and full training sub-dataset of wind speed forecast. For a fixed ϕ, update W on a mini-batch using the derivatives of the negative log marginal likelihood with respect to W. Update the θ and kernel matrix ϕ on changed full vectors in the hidden space due to updated W. Repeat the process for the mini-batches until convergence.

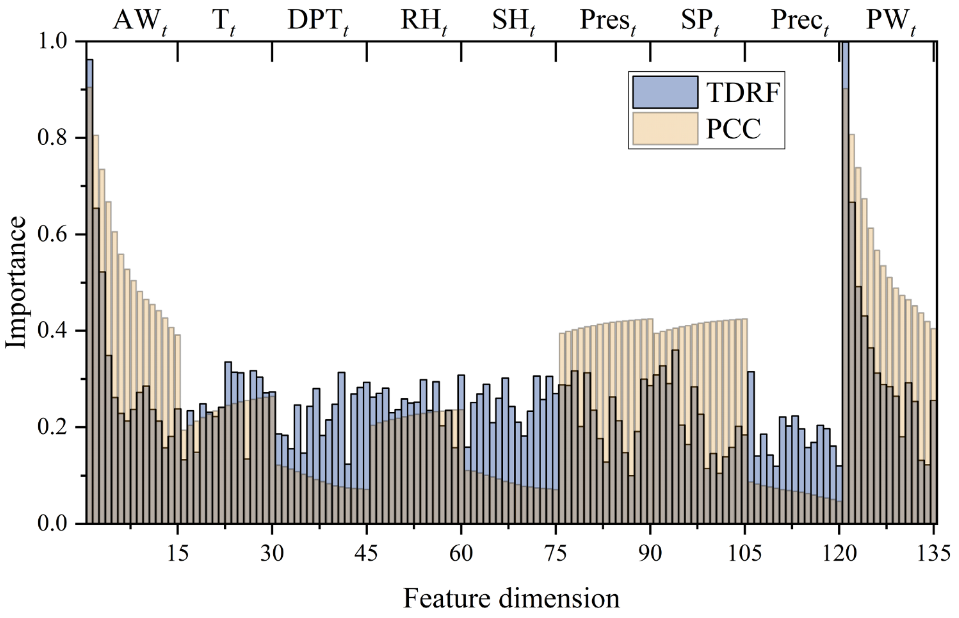

Fig. 2. The importance of 135 features using PCC and TDRF (the blue bar chart represents the TDRF results, the yellow bar chart represents the PCC results).

Table 6
Statistics of the forecast skill involving point forecast and interval forecast, dataset two.

Station	Model	RMSE	RSE	MAE	SS_{CRPS}
One step ahead	TGPLSTM	1.33	0.05	0.84	0.82
	MLP	1.43	0.06	0.88	–
	GLM	1.54	0.06	0.93	0.80
Two steps ahead	TGPLSTM	2.27	0.19	1.34	0.77
	MLP	2.14	0.15	1.32	–
	GLM	2.46	0.21	1.42	0.71
Three steps ahead	TGPLSTM	2.46	0.22	1.58	0.74
	MLP	2.50	0.21	1.57	–
	GLM	2.62	0.28	1.63	0.65
Four steps ahead	TGPLSTM	2.68	0.28	1.73	0.69
	MLP	2.95	0.42	1.84	–
	GLM	3.23	0.65	1.99	0.59

3. Experimental data

This paper implemented ultra-short-term wind speed forecasting for future one hour with a time resolution of 15 min. Available data are obtained from the National Wind Energy Technology Center (NWTC) M2 Wind Tower (https://www.osti.gov/biblio/1052222). The original sampling step is one minute. Data include average wind speed at 80 m (AWS), peak wind speed at 80 m (PWS), temperature at 80 m (T), dew point temperature (DPT), relative humidity (RH), specific humidity (SH), station pressure (SP), sea-level pressure (SLP), accumulated precipitation (AP).

Before the experiment, the quality of the original data is carefully checked. Two different wind farm datasets are extracted for a more comprehensive assessment. The first dataset is from 0:00 on April 1 to 11:59 on April 30, 2015. In this dataset, the wind speed values are relatively small, the minute wind speed has a 0.99 quantile of 12.94 m/s. The second dataset is from 0:00 on October 13 to 11:59 on November 12, 2015. The wind speed values are relatively large, and the minute wind speed has a 0.99 quantile of 21.18 m/s, up to the nine Beaufort wind force scale.

The statistical characteristics of the two datasets are shown in Tables. 1 and 2. 1 and 2 list the mean, minimum, 1% quantile, 5% quantile, 25% quantile, 50% quantile, 75% quantile, 95% quantile, 99% quantile and maximum values. The statistical results show that the data quality is acceptable, and there are no outlier values and invalid values. In addition, the magnitudes of different types of feature factors vary widely, and SLP and SH are even 1000 times different in magnitude. Therefore, data are normalized before the training in this paper. Data of 15 min are calculated, the obtained total data length is 2880 and 2976, respectively. According to the chronological order, the first 2500 samples are selected as the training set, and the remaining samples are used as the testing set.

4. Results

This section starts with an evaluation of feature importance, based on which a subset is selected. Using the selected features, the forecast performance is evaluated with a validation data set. In addition, prediction interval analysis and probabilistic forecast error analysis are performed.

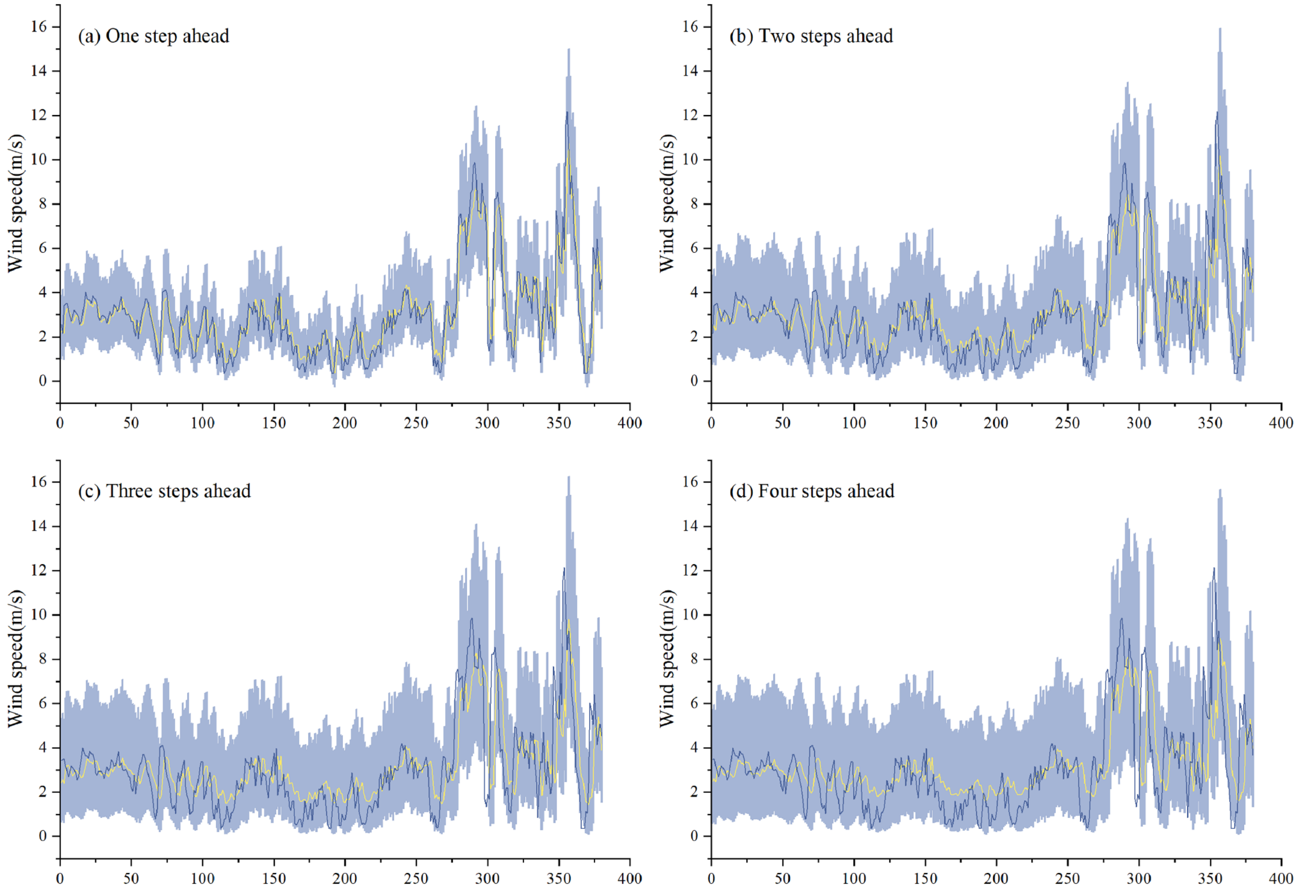

Fig. 3. The TGPLSTM forecast and observed value with 95% confidence intervals for one, two, three and four steps ahead forecast (the dark blue line is observed value, the yellow line is forecast, the blue band is 95% confidence interval).

4.1. Feature importance and selection

Nine types of features with fifteen historical time lags are included as the initial feature set. The dimension of features is 135, which are shown in Table 3.

TDRF algorithm is used to calculate the importance of 135 features. The Pearson correlation coefficients (PCC) is also computed. Fig. 2 shows the importance from TDRF and PCC. The yellow bar plot presents the PCC results, and the blue bar plot presents the TDRF results. The PCC importance of the same type of features are very similar. For example, the features of dimensions 1–15 are average wind speeds with 1–15 lag times and their PCC importance decreases slowly as the time lag increases. Whereas the TDRF importance of the same types of features vary greatly. Therefore, if the PCC method is adopted for feature selection, a series of factors in the same type is included, which could increase the redundancy. It is demonstrated that TDRF automatically eliminates the redundancy.

Table 4 shows the dimensions of the thirty most important features.

4.2. Forecast and validation

To forecast wind speed at future one hour with a time resolution of 15 min, multi-step ahead TGPLSTM forecast is developed. The ten most important features selected by TDFR are included as the input, and the output is the average wind speed in the future at 15 min, 30 min, 45 min and 60 min. The prediction step is 15 min. Multilayer perceptron (MLP) model and generalized linear regression (GLM) model are also developed to make a comparison. The input and output sets of MLP and GLM are the same as TGPLSTM. Standard MLP is a point forecast model without probability information. GLM is specified by univariate independent response variables and the canonical link function. The expected value is linked to a linear predictor by a known monotone function. It provides the interval forecast based on the asymptotic variance of the maximum likelihood estimate. Using the first 2500 samples as a training set, and the remaining samples as the testing set. The prediction performances of TGPLSTM, MLP, and GLM are evaluated.

Table 5 presents forecast evaluation on dataset one. It involves point forecast metrics *RMSE*, *RSE*, *MAE*, and interval forecast metrics SS_{CRPS}. For *RMSE*, *RSE*, *MAE*, a better forecast is detected with lower *RMSE*, *RSE*, *MAE*. For one step ahead forecast, the *RMSE*, *RSE*, *MAE* of TGPLSTM are 0.90, 0.24 and 0.58, respectively. The values of MLP are 0.91, 0.25 and 0.57, of GLM are 1.07, 0.29, 0.63. The *RMSE*, *RSE*, *MAE* values of TGPLSTM are low. TGPLSTM improves the forecast accuracy by 15.9%, 17.2%, 7.9% compared to GLM. For two steps ahead forecast, TGPLSTM increases forecast accuracy by 1.6%, 4.9%, 4.6% compared to MLP, 6%, 9.4%, 9.9% compared to GLM. For three steps ahead forecast, TGPLSTM increases forecast accuracy by 1.3 %, 9.5%, 5.6% compared to MLP, 4.5%, 12.8%, 9.8% compared to GLM. For four steps ahead forecast, TGPLSTM increases forecast accuracy by 1.3 %, 15.5%, 7.4% compared to MLP, 3.5%, 11.5%, 10.3% compared to GLM. Probabilistic forecast results are evaluated by SS_{CRPS}, a higher SS_{CRPS} value represents a better uncertainty forecast. Here TGPLSTM and GLM are capable of providing probabilistic results. For one step ahead forecast, SS_{CRPS} of TGPLSTM and GLM are 0.82 and 0.81, the values are closing. However, for two steps ahead forecast, SS_{CRPS} of TGPLSTM is 0.78, of GLM is 0.71. TGPLSTM increases the accuracy of probabilistic forecast by 9.8%. Consistently, TGPLSTM improves the performance of

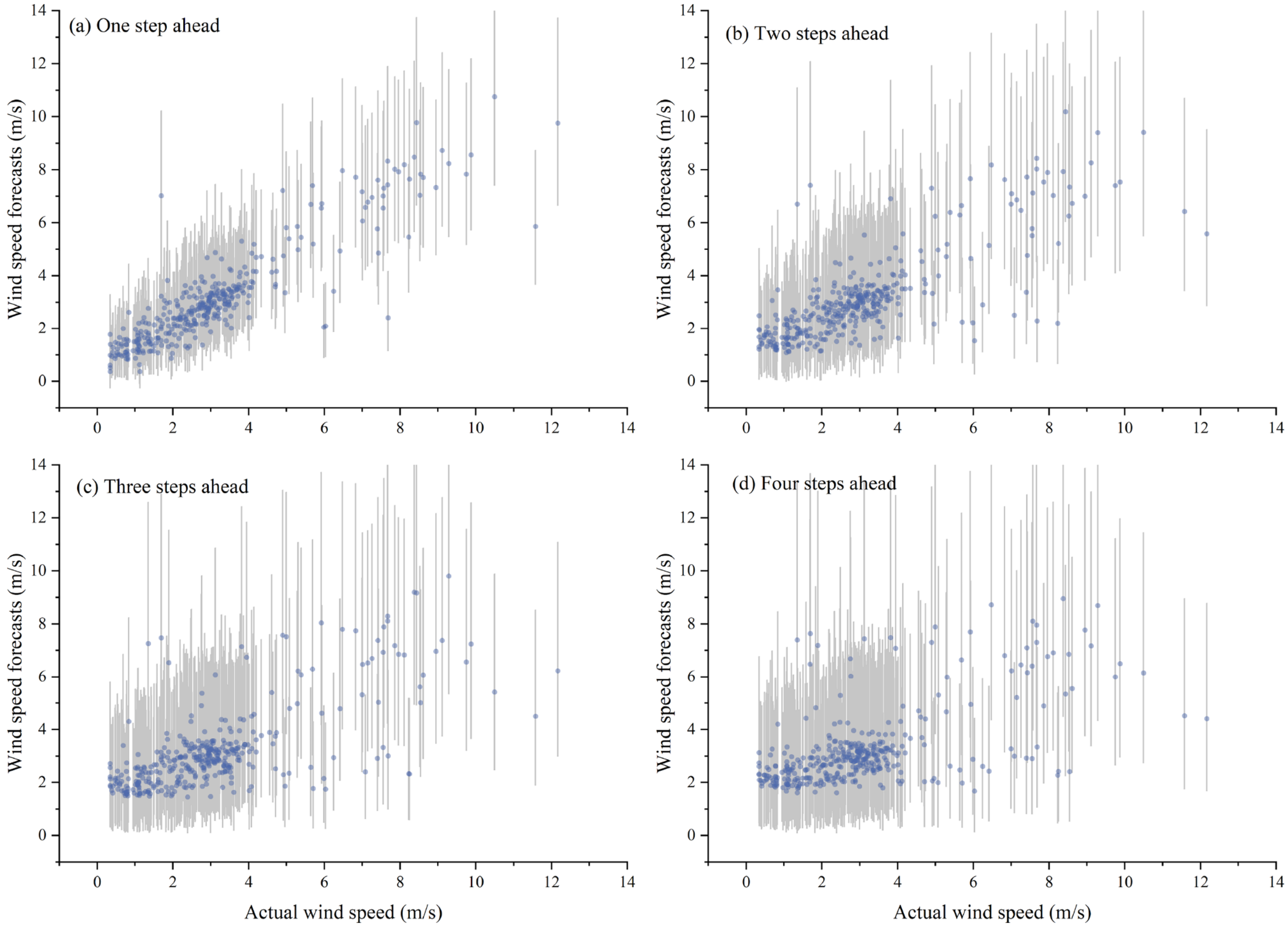

Fig. 4. The TGPLSTM forecast shown according to the the rank of actual wind speed value (the blue scatter represents the data pair of forecast mean and observed value, the grey bar is the related forecast interval).

three steps ahead and four steps ahead interval forecast by 18.5% and 17.7%, respectively. It can be concluded that the proposed TGPLSTM shows significantly improved forecast skills compared with state-of-the-art models, and also provides satisfactory prediction interval.

Table 6 presents the results of the dataset two forecast. For one step ahead forecast, the RMSE, RSE, MAE of TGPLSTM are 1.33, 0.05, 0.84, respectively. The values of MLP are 1.43, 0.06 and 0.88, of GLM are 1.54, 0.06, 0.93. TGPLSTM improves the forecast accuracy by 13.6%, 16.7%, 9.7% compared to GLM. For two, three and four steps ahead forecast, TGPLSTM has also improved forecast accuracy through the same analysis. For the evaluation of probabilistic forecast, SSCRPS of TGPLSTM is higher than that of GLM for all steps ahead forecast. Thus, the probabilistic forecast of TGPLSTM is better than GLM. Consistently, TGPLSTM improves the prediction performance of higher wind speeds.

4.3. Prediction interval analysis

Fig. 3(a), (b), (c), (d) depict the details of TGPLSTM forecast for one, two, three, and four steps ahead, respectively. Each figure illustrates the actual wind speed record vs. the predicted wind speed. It can be seen that the deterministic prediction and 95% confidence range keeps up with the fluctuation of the actual values. It can be seen that the 95% interval width varies with the wind speed. Generally speaking, the higher the wind speed, the wider the interval width, and vice versa. This is consistent with the actual situation that the greater wind speed leads to the greater forecast error and sample variance, so the uncertainty interval increases. Moreover, the prediction interval is asymmetrical to the forecast mean. The Gaussian process regression provides a target variable satisfying conditional normal distribution. Therefore, the forecast means and 5%, 95% quantile values should be symmetrical. However, to generate the variables conforming to the normal distribution to satisfy the basic assumptions of incorporating GPR model, Box-Cox transformation [31] is adopted to normalize wind speed series. Since the Box-Cox transformation is a non-linear transformation, 5%, and 95% quantiles become asymmetrical after the inverse transformation.

Fig. 4 shows the prediction plots with the order of true wind speeds, rather than sample order. The blue points are predicted mean and the grey bars are prediction interval. The Fig. 4(a) shows more clearly that more points of the one-step-ahead prediction are gathered on the y-x line. As the steps increase, the distribution of points is looser. The heteroscedastic property is apparent and the uncertainty interval is small when the actual wind speed is small, and vice versa.

4.4. Probabilistic forecast errors analysis

Probability in truth (PIT) of each sample is calculated to measure the probabilistic errors. PIT refers to the statistical consistency of forecasts and observations. It is the probability of the observed wind speed value in the predicted distribution. If the PIT is around 0.5, the forecast and observation are close. If the PIT is closer to 0 or 1, the forecast and observation are very different. Fig. 5 depict the PIT value of one, two, three and four steps ahead forecast. The grey circles represent PIT values of GLM forecast, and red circles represent PIT values of TGPLSTM forecast. It can be seen that most PIT circles of TGPLSTM forecast are distributed between 0.3 and 0.7 horizontal lines, the probabilistic forecast errors are acceptable. Whereas more PIT circles of GLM forecast are closer to 0 and 1. Therefore, from the probabilistic

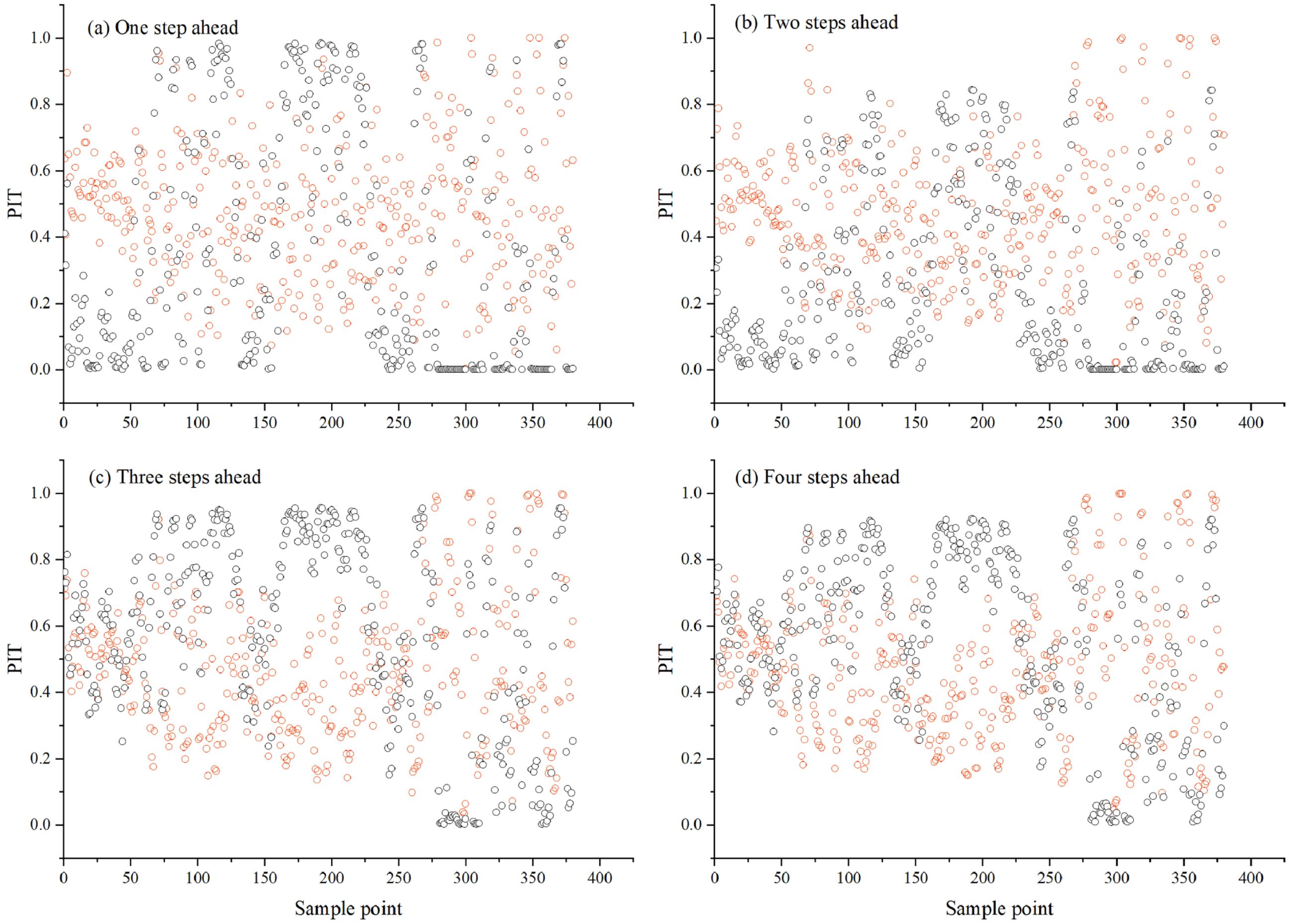

Fig. 5. PIT plots of TGPLSTM and GLM forecasts, measuring the probabilistic errors (The grey circles represent PIT values of GLM forecast, and red circles represent PIT values of TGPLSTM forecast).

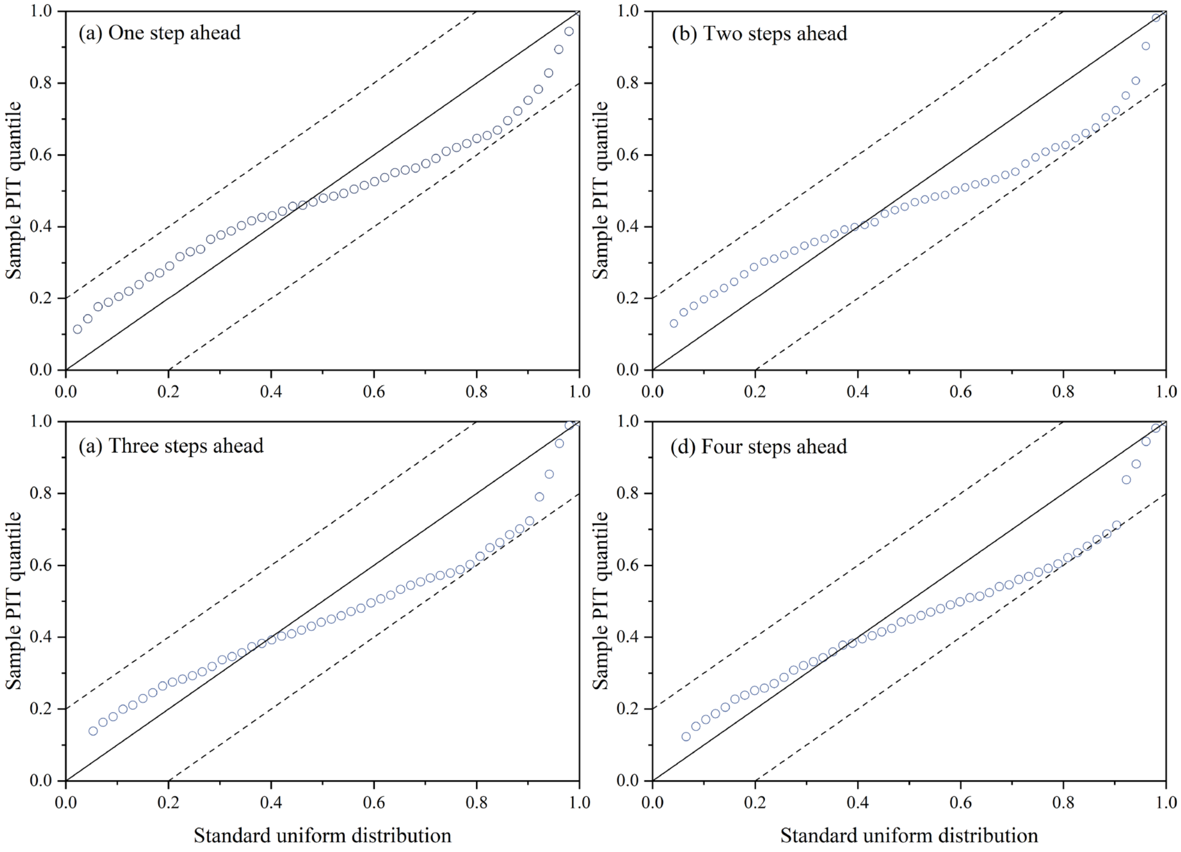

Fig. 6. PIT uniform probability plots for one, two, three and four steps ahead forecasts (the scatter represents the sampled quantile vs. uniform distribution quantile, the dotted lines are the 5% Kolmogorov significance bands).

forecast errors analysis, TGPLSTM forecast is better than GLM. As the forecast step size increases, it is found that the PIT circles shift to 0.5 horizontal line. It is known in the previous section that as the forecast step increases, the predicted probability distribution has a larger variance and a flatter distribution shape. Even if the predicted value and the observed value are far apart, the corresponding probability does not change much. Moreover, the probability prediction can be better.

4.5. Reliability of the forecast

Fig. 6 shows the PIT uniform probability plots for one, two, three and four steps ahead forecasts for the analysis of the reliability of the TGPLSTM wind speed forecast. The PIT sequences are simulated by the uniform probability distribution. The Q-Q maps of standard uniform distribution and sampled quantile are drawn, and the 5% Kolmogorov significance bands [32] are added to the graph to check whether it passes the test. Fig. 6 shows that the sampled quantiles are close to the diagonal line and within the significance bands. It implies that the distribution of PIT values is uniform. Hence, it is concluded that the predicted probability distribution is appropriate, generally unbiased, and has the appropriate bandwidth.

5. Discussion and conclusion

Long short-term memory (LSTM) wind speed forecast model has received a lot of attention in recent years due to its powerful fitting ability. But the probabilistic wind speed forecast based on LSTM is still insufficient as large scale ensemble technique based on resampling is difficult to implement. This paper proposed a more direct and reliable probabilistic wind speed forecast model TGPLSTM by coupling Gaussian process regression (GPR) to the internal structure of LSTM and using an intelligent inputs screening of top-down relevant feature search algorithm (TDRF).

Two datasets of different seasons are extracted to verify the model's predictability at different wind speed levels. The TDRF effectively eliminates irrelevant and redundant features. A probabilistic wind speed forecast is implemented using the most important features. The proposed TGPLSTM shows significantly improved forecast skills compared with state-of-the-art models, meanwhile, a satisfying interval is provided along with the point forecast. The accuracy of one step ahead prediction is best. As the forecast step increases, the accuracy of the forecast decreases.

Probability in truth (PIT) is used to measure the probabilistic errors. More PIT circles of TGPLSTM forecast are distributed nearby 0.5 horizontal lines than that of GLM. It is concluded that the probabilistic forecast error of TGPLSTM is lower than GLM. PIT uniform probability plots are made to examine the reliability of the TGPLSTM wind speed forecast. It indicated that the distribution of PIT values is very uniform. The predicted probability distribution of TGPLSTM is appropriate, generally unbiased and has the appropriate bandwidth.

Comprehensive experiments using practical wind farm data of different seasons have demonstrated the highly satisfactory results. The proposed TGPLSTM approach maintains the advanced prediction ability of deep learning and provides a direct confidence interval of prediction. It enriches the information content of wind speed prediction and makes the short-term wind speed prediction result more meaningful. Moreover, the TGPLSTM approach is a generalized framework for probabilistic forecasting of wind speed, and can provide an efficient support for power system applications such as probabilistic reserve determination, generation dispatch, wind farm control, electricity market trading, etc.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (51809242) and the Fundamental Research Funds for the Central Universities (G1323541875).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.enconman.2019.06.083.

References

[1] Jadidoleslam M, Ebrahimi A, Latify MA. Probabilistic transmission expansion planning to maximize the integration of wind power. Renewable Energy 2017;114:866–78.

[2] Zheng D, Shi M, Wang Y, Eseye A, Zhang J. Day-ahead wind power forecasting using a two-stage hybrid modeling approach based on SCADA and meteorological information, and evaluating the impact of input-data dependency on forecasting accuracy. Energies 2017;10(12):1988.

[3] Hao Z, Singh VP, Xia Y. Seasonal drought prediction: advances, challenges, and future prospects. Rev Geophys 2018;56(1):108–41.

[4] Tascikaraoglu A, Uzunoglu M. A review of combined approaches for prediction of short-term wind speed and power. Renew Sustain Energy Rev 2014;34:243–54.

[5] Wang H, Wang G, Li G, Peng J, Liu Y. Deep belief network based deterministic and probabilistic wind speed forecasting approach. Appl Energy 2016;182:80–93.

[6] Zhang CY, Chen CLP, Gan M, Chen L. Predictive deep boltzmann machine for multiperiod wind speed forecasting. IEEE Trans Sustain Energy 2017;6(4):1416–25.

[7] Yu C, Li Y, Bao Y, Tang H, Zhai G. A novel framework for wind speed prediction based on recurrent neural networks and support vector machine. Energy Convers Manage 2018;178:137–45.

[8] Chen Y, Zhang S, Zhang W, Peng J, Cai Y. Multifactor spatio-temporal correlation model based on a combination of convolutional neural network and long short-term memory neural network for wind speed forecasting. Energy Convers Manage 2019;185:783–99.

[9] Liu H, Mi X, Li Y. Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM. Energy Convers Manage 2018;159:54–64.

[10] Hu Y-L, Chen L. A nonlinear hybrid wind speed forecasting model using LSTM network, hysteretic elm and differential evolution algorithm. Energy Convers Manage 2018;173:123–42.

[11] Huang K, Ye L, Chen L, Wang Q, Dai L, Zhou J, Singh VP, Huang M, Zhang J. Risk analysis of flood control reservoir operation considering multiple uncertainties. J Hydrol 2018;565:672–84.

[12] Lei Y, Zhou J, Zeng X, Guo J, Zhang X. Multi-objective optimization for construction of prediction interval of hydrological models based on ensemble simulations. J Hydrol 2014;519:925–33.

[13] Ye L, Zhou J, Gupta HV, Zhang H, Zeng X, Chen L. Efficient estimation of flood forecast prediction intervals via single-and multi-objective versions of the lube method. Hydrol Process 2016;30(15):2703–16.

[14] Hlal MI, Ramachandaramurthya VK, Padmanaban S, Kaboli HR, Pouryekta A, Abdullah T, Ab Rashid T. NSGA-II and MOPSO based optimization for sizing of hybrid PV/wind/battery energy storage system. Int J Power Electron Drive Syst 2019;1(1):463–78.

[15] Huang K, Ye L, Chen L, Wang Q, Dai L, Zhou J, Singh VP, Huang M, Zhang J. Risk analysis of flood control reservoir operation considering multiple uncertainties. J Hydrol 2018;565:672–84.

[16] Wang H, Li G, Wang G, Peng J, Jiang H, Liu Y. Deep learning based ensemble approach for probabilistic wind power forecasting. Appl Energy 2017;188:56–70.

[17] Chen J, Zeng G, Zhou W, Du W, Lu K-D. Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization. Energy Convers Manage 2018;165:681–95.

[18] Wan C, Xu Z, Pinson P, Dong ZY, Wong KP. Probabilistic forecasting of wind power generation using extreme learning machine. IEEE Trans Power Syst 2014;29(3):1033–44.

[19] Wu W, Chen K, Qiao Y, Lu Z. Probabilistic short-term wind power forecasting based on deep neural networks. 2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS). IEEE; 2016. p. 1–8.

[20] Zhang J, Yan J, Infield D, Liu Y, Lien F. Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and gaussian mixture model. Appl Energy 2019;241:229–44.

[21] Kersting K, Plagemann C, Pfaff P, Burgard W. Most likely heteroscedastic gaussian process regression. Proceedings of the 24th international conference on Machine learning. ACM; 2007. p. 393–400.

[22] Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw 2010;36(11):1–13.

[23] Stańczyk U, Zielosko B, Jain LC. Advances in Feature Selection for Data and Pattern Recognition. Springer; 2018.

[24] Ishak S, Kotha P, Alecsandru C, Student G. Optimization of dynamic neural network performance for short-term traffic prediction. Transp Res Rec J Transp Res Board 2003;1836(1):27–31.

[25] Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. Lstm: a search space odyssey. IEEE Trans Neural Networks Learn Syst 2016;28(10):2222–32.

[26] Gibbs MN. Bayesian gaussian processes for regression and classification. Citeseer 1998. [Ph.D. thesis].

[27] Tolvanen V, Jylänki P, Vehtari A. Expectation propagation for nonstationary heteroscedastic gaussian process regression. Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on. IEEE; 2014. p. 1–6.

[28] Zhu S, Luo X, Xu Z, Ye L. Seasonal streamflow forecasts using mixture-kernel GPR and advanced methods of input variable selection. Hydrol Res 2018:nh2018023.

[29] M. Al-Shedivat, A.G. Wilson, Y. Saatchi, Z. Hu, E.P. Xing, Learning scalable deep kernels with recurrent structure, arXiv preprint arXiv:1610.08936.

[30] S.H.A. Kaboli, A.K. Alqallaf, Solving non-convex economic load dispatch problem via artificial cooperative search algorithm, Expert Systems with Applications.

[31] Osborne JW. Improving your data transformations: applying the box-cox transformation. Practical Assess Res Eval 2010;15(12):1–9.

[32] Laio F, Tamea S. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol Earth Syst Sci 2007;11(4):1267–77.