Applied Energy 188 (2017) 56–70

Contents lists available at ScienceDirect

Applied Energy

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

Deep learning based ensemble approach for probabilistic wind power forecasting

Huai-zhi Wang [a], Gang-qiang Li [a], Gui-bin Wang [b,*], Jian-chun Peng [a], Hui Jiang [c], Yi-tao Liu [a]

[a] The College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518060, China
[b] Shenzhen Key Laboratory of Urban Rail Transit, The College of Urban Rail Transit, Shenzhen University, Shenzhen 518060, China
[c] The College of Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China

HIGHLIGHTS

- Convolutional neural network is designed for probabilistic wind power forecasting.
- Ensemble technique is used to cancel out the diverse errors of point forecasters.
- The model misspecification and data noise in wind power are separately evaluated.
- The competitive performance and robustness of the proposed method were proved.

ARTICLE INFO

Article history:
Received 9 September 2016
Received in revised form 8 November 2016
Accepted 26 November 2016
Available online 8 December 2016

Keywords:
Convolutional neural network
Ensemble
Probabilistic wind power forecast
Deep learning
Wavelet transform

ABSTRACT

Due to the economic and environmental benefits, wind power is becoming one of the more promising supplements for electric power generation. However, the uncertainty exhibited in wind power data is generally unacceptably large. Thus, the data should be accurately evaluated by operators to effectively mitigate the risks of wind power on power system operations. Recognizing this challenge, a novel deep learning based ensemble approach is proposed for probabilistic wind power forecasting. In this approach, an advanced point forecasting method is originally proposed based on wavelet transform and convolutional neural network. Wavelet transform is used to decompose the raw wind power data into different frequencies. The nonlinear features in each frequency that are used to improve the forecast accuracy are later effectively learned by the convolutional neural network. The uncertainties in wind power data, i.e., the model misspecification and data noise, are separately identified thereafter. Consequently, the probabilistic distribution of wind power data can be statistically formulated. The proposed ensemble approach has been extensively assessed using real wind farm data from China, and the results demonstrate that the uncertainties in wind power data can be better learned using the proposed approach and that a competitive performance is obtained.

1. Introduction

Due to the continuous decrease in the storage capacity of fossil fuel, the energy crisis is becoming more significant than ever [1]. Therefore, to mitigate the energy crisis, regulatory acts that encourage the use of renewable energy have been promoted worldwide. Among the renewable energy resources, wind energy, as an alternative to fossil energy, has attracted much attention due to its beneficial impacts on climate change mitigation and environmental pollution reduction [2]. Coupled with its mature technology, wind energy has experienced an unexpected annual growth on a global scale. Wind energy can be used to drive engines directly and provide rural energy services. In [3], a novel mean flow acoustic engine with a cross-junction configuration was designed to convert wind energy in a pipeline into acoustic energy, and its efficiency was numerically analyzed in [4] by using computational fluid dynamics method. In practice, wind energy is mainly utilized to mechanically power generators for electricity. The annual growth rate of worldwide wind power has been between 20% to 35% per year since 2000 [5]. However, due to the chaotic nature of the earth's atmosphere, wind generated power always exhibits nonlinear and non-stationary uncertainties, which pose great challenges for the management and operations of electric power and

* Corresponding author.
E-mail addresses: wanghz@szu.edu.cn (H.-z. Wang), ligq@szu.edu.cn (G.-q. Li), wanggb@szu.edu.cn (G.-b. Wang), jcpeng@szu.edu.cn (J.-c. Peng), huijiang@szu.edu.cn (H. Jiang).

http://dx.doi.org/10.1016/j.apenergy.2016.11.111
0306-2619/

Nomenclature

ACE	average coverage error
CNN	convolutional neural network
DBM	deep restricted Boltzmann machine
MWWF	milky way wind farm
PI	prediction interval
QR	quantile regression
SIWF	Shangchuan island wind farm
WPF	wind power forecasting
A_n	wavelet approximation signal
D_n	wavelet detail signal
DS_{mu}	dataset used for model uncertainty evaluation
GD	Gaussian distribution
I_i^a	PI at time step i given PINC = $100(1-\alpha)\%$
M_{du}	mean of data noise uncertainty
N_E	number of ensembles
N_S	number of training samples
$\boldsymbol{W}$	CNN's weight matrix
$\boldsymbol{W_{log}^L}$	weight matrix at Lth logistic regression layer
T	length of the signal required to be decomposed
$\boldsymbol{b}$	CNN's bias matrix
$\boldsymbol{b_{log}^L}$	bias matrix at Lth logistic output layer
$\boldsymbol{c}$	CNN's additive bias matrix
d	output vector size of training samples
$f(\cdot)$	output activation function
h_j^i	the jth target in ith training sample
m	mini-batch size of training sample
t	discrete time step
$up(\bullet)$	up-sampling function
$\boldsymbol{w}_{i,j}^l$	weight matrix at lth layer connecting the ith input map and jth output map
$\boldsymbol{x}_{i,j,k}$	the ith input in jth input map at kth layer
$\widehat{y}(\boldsymbol{x_i})$	mean of the estimated model uncertainty
$\boldsymbol{y_j^l}$	the jth output map at lth layer
$y(\boldsymbol{x_i})$	output of the jth deep CNN model
α	confidence level parameter
$\boldsymbol{\beta_{sub}^l}$	multiplicative bias matrix at lth sub-sampling layer
$\boldsymbol{\gamma_{len,wid}}$	average filter parameter matrix with size $len \times wid$
$\varepsilon(\boldsymbol{x}_i)$	uncertainty given the input $\boldsymbol{x}_i$
κ	translation variable
$\phi(\bullet)$	mother wavelet function
σ_{mu}^2	variance of model uncertainty
σ_ε^2	variance of the uncertainty signal $\varepsilon(\boldsymbol{x}_i)$
BP	back-propagation algorithm
CRPS	continuous ranking probability score
IS	interval sharpness
NN	neural network
PINC	prediction interval nominal confidence
SAE	stacked auto-encoder
SVM	support vector machine
WT	wavelet transform
CDF_i	cumulative distribution function at time step i
DS_{du}	dataset used for data noise uncertainty evaluation
E_m	squared-error loss function considering m batches
$H(\cdot)$	indicative function
L_h^α	lower bound of PI given target h and PINC
M_ε	mean of the uncertainty signal $\varepsilon(\boldsymbol{x}_i)$
N_M	number of selected input maps
U_h^α	upper bound of PI given target h and PINC
$\boldsymbol{W_{con}^l}$	weight matrix at lth convolution layer
WS_i^a	wind speed at time step i given PINC
b_j^l	bias of jth output map at lth layer
$\boldsymbol{b_{con}^l}$	bias matrix at lth convolution layer
c_j^l	additive bias of jth output map at lth layer
$\boldsymbol{c_{sub}^l}$	additive bias matrix at lth sub-sampling layer
$down(\cdot)$	down-sampling function
$g(\cdot)$	signal required to be decomposed by wavelet
len	length of a given map
r_i	indicator of prediction interval coverage probability
$\boldsymbol{u^L}$	output vector of the neurons in $(L-1)$ layer
wid	width of a given map
$\boldsymbol{x_i^l}$	the ith input map at lth layer
$\boldsymbol{x}^{L-1}$	the output of the neurons at $(L-1)$th layer
$\boldsymbol{y}_{i,j,k}$	the ith output in jth output map at kth layer
y_j^i	the jth output in ith training sample
$z_{1-\alpha/2}$	critical value of a Gaussian distribution function
$\boldsymbol{\beta}$	CNN's multiplicative bias matrix
β_j^l	multiplicative bias of the jth output map at lth layer
δ_i^a	width of the PI at time step i given PINC
η	learning rate
υ	scaling variable
σ_{du}^2	variance of data noise uncertainty
σ_h^2	variance of total forecasting error

energy systems. It is demonstrated in [6] that the impact of these uncertainties on power system operations can be, to a certain degree, mitigated via advanced WPF methods, which are considered to be the most promising solutions for the integration of a large amount of wind energy into power grids. Aimed at this task, three typical methodologies for WPF have been proposed in the literature, including physical modeling, statistical methods and soft-computing techniques.

Physical modeling methods try to establish an accurate mathematical model for WPF using various geographical and meteorological information. However, this type of approach may not be applicable for practical real-time prediction tasks due to the high amount of calculation costs involved [7,8], whereas statistical approaches manage to develop an optimal relationship between future wind power and historical samples via error minimization. In [9], a generalized WPF model was proposed based on time-varying threshold autoregressive moving average, and the efficiency was numerically analyzed. In [10], a hybrid statistical approach in combination with empirical wavelet transform, partial auto-correlation function and Gaussian process regression was proposed. Simulation results indicate that the suggested approach, i.e., the generalized WPF model, performed the best among the three compared methods. In addition, soft-computing techniques, such as the artificial neural network [11,12] and Elman neural network [13], were utilized for WPF. In [14], a WPF model based on extreme machine learning was presented to evaluate wind power density. In [15], a multi-layer neural fuzzy network was mooted for hour-ahead WPF, and the model parameters were well-trained by using simultaneous perturbation stochastic approximation. In [16], a hybrid model based on wavelet packet technique and artificial neural network was originally proposed, and the model parameters were optimized by using crisscross optimization algorithm. In [17], reproducing kernel Hilbert space based probabilistic WPF method was proposed and the performance was evaluated by CRPS. In [18], the randomness and uncertainty of wind energy were quantitatively evaluated using Gaussian process regression and teaching learning optimization. In [19], a general framework based on k-nearest neighbors algorithm and kernel

density estimator were constructed for probabilistic WPF. The main merit of the soft-computing based WPF models is their potential abilities regarding data-mining and feature extracting [20]. A detailed overview of probabilistic WPF and wind speed forecasting is presented in [21].

However, though the afore mentioned prediction models can be categorized into individual forecasters, they may suffer from instability, high sensitivity to initial parameters and over-training problems. Nonetheless, these problems can be effectively mitigated by ensemble strategy [22], because the involved diverse errors from model misspecification and data noise can cancel each other out via the aggregating progress. Another advantage with regard to ensemble technique is the increased accuracy and provided probabilistic uncertainties, which play a decisive role as the operators prepare for future unknown conditions. Therefore, ensemble technique based WPF approach has attracted extensive attention in recent years. In [23], a weather ensemble prediction framework for wind power forecast was constructed, and it demonstrated superior performance. In [24], an ensemble approach based on empirical mode decomposition, sample entropy technique and extreme learning machine, was proposed for probabilistic WPF. In [25], an ensemble model consisting of 52 neural networks and five Gaussian process sub-models were put forward to better quantify the wind power distribution information.

Another demerit regarding the physical modeling, statistical approaches and intelligent methods previously presented is their shallow learning models. As presented in [26], considering the wind power data is complicated in nature, these shallow models may be insufficient to extract the corresponding deep nonlinear and non-stationary traits. In addition, with the rapid development of the smart grid, the environmental sensors and related technologies are becoming increasingly more widely used than ever before, which compel us into a big-data era. Accordingly, it becomes even more difficult to extract the deep features of wind power data [27]. However, one effective way to address the shallow model issue is the use of deep learning [28], due to the ability to discover the inherent abstract features and hidden high-level invariant structures in data. The characteristics specific to feature extraction results in deep learning being much more attractive for WPF [29]. In summary, the unsatisfactory feature mining of the shallow model and existing problems for individual forecasters inspire us to rethink the WPF problem based on deep learning architecture and the ensemble technique.

Deep learning mainly consists of SAE, DBM and CNN. SAE consists of an unsupervised learning subpart that use auto-encoder as its building blocks and a logistic regression layer for data fitting [26]. SAE aims to learn a dimensionality reduction representation for the input data. DBM is a generative graphical model composed of multiple layers of hidden Boolean units with connections between the layers but not between the units within each layer [30]. DBM tries to learn a desirable probability distribution from the input units to the output units by minimizing a user-defined energy function. Compared to conventional neural network, the merit of SAE and DBM is in their training process, which can be used to effectively alleviate the local minimum dilemma. The training process of SAE and DBM consists of a layer-wise pre-training process and a fine-tuning process. The former is used to provide good initial values for all parameters, while the latter is used to search the optimum based on the given initial states of the network [31]. The earlier implementations of SAE and DBM tailored for wind power prediction were reported in [26,31,32]. However, the drawback of SAE and DBM is their limited ability to search for a global optimum, especially when the number of parameters to be optimized substantially increases as the neural network architecture goes deeper or the dimensionality of the neurons expands. Another typical deep learning algorithm is CNN, an approach in which the connectivity principles between the neurons are inspired by the organization of animal visual cortex. Compared to SAE and DBM, CNN not only has fewer parameters to be estimated because of the weight sharing technique, but it also can effectively extract the hidden structure and inherent features. Nevertheless, to the authors' knowledge, CNN designed for WPF has not yet been considered in the published literature. Therefore, this research is devoted to investigating a deep framework for probabilistic WPF. Therefore, a hybrid approach based on WT and CNN, that can enhance the forecasting performance and prediction efficiency, is originally proposed. The main contributions of this paper are presented as follows:

(1) For the first time, CNN is introduced and tailored to comprehensively extract the deep invariant structures and hidden high-level nonlinear features exhibited at any wind power frequency.
(2) A hybrid approach based on WT, CNN and ensemble technique is proposed to quantify the wind power uncertainties with respect to model misspecification and data noise.
(3) Based on the quantified model misspecification and data noise, a WPF framework is originally formulated to probabilistically evaluate the randomness and stochastics in wind power data from the perspective of sharpness, reliability and overall skills.

The proposed probabilistic WPF approach has been thoroughly tested and benchmarked using real wind power data in time domain under various time-scales and performance criteria.

2. Deep convolutional neural network

As one of the means to effectively extract the invariant structures and inherent hidden features in data, deep CNN has been successfully applied in various fields, including image-identification, classification, and feature-mining [33]. A typical deep CNN architecture consists of alternating the convolution layer and subsampling layer with a fully connected layer as the output.

2.1. Convolution layer

The convolution layer is a two-layer feed-forward neural network that adopts a convolution operation to map the low-level maps with local features into several high-level maps with global features. In CNN architecture, no connections exist between the neurons in the same layer. Moreover, weight sharing technique is employed between the neurons in different CNN layer to simplify the feed forward and back propagation process. Concretely, the feature maps in the previous $(l-1)$th layer are convolved with the shared weights, also termed kernels, and then passed through a user-defined activation function to generate the lth layer with several output feature maps, as follows,

$$\boldsymbol{y}_j^l = f\left(\sum_{i\in N_M} \boldsymbol{x}_i^{l-1} \otimes \boldsymbol{w}_{i,j}^l + b_j^l\right) \tag{1}$$

where $\otimes$ denotes a convolution operation. In addition, the output activation function $f(\cdot)$ is chosen to be the sigmoid function.

The weight sharing technique is the main attribute that differs from other deep learning algorithms, i.e., SAE and DBM. In this technique, all neurons in the same output feature map share the same weight but receives inputs from the neurons at different locations. It has been demonstrated that the memory footprints and number of parameters to be estimated can be reduced significantly and the hidden invariant features in data can be effectively extracted via the use of weight sharing technique [34].

2.2. Sub-sampling layer

Classical convolution layers are usually intersected with sub-sampling layers to gradually build up the high-level invariant structures in data without sacrifice of specificity and also to reduce the calculation time. The output maps of a sub-sampling layer are more concise representations of input maps. Specifically, a sub-sampling layer generates a down-sampled version of the input maps, as follows

$$\boldsymbol{y}_j^l = f\left(\beta_j^l \,\mathrm{down}\left(\boldsymbol{x}_j^{l-1}\right) + c_j^l\right) \tag{2}$$

where down-sampling function down(·) adopts the average function, described as follows,

$$\boldsymbol{y}_{i,j,k} = \sum_{len,wid} \gamma_{len,wid} \boldsymbol{x}_{i+p,j+q,k} \tag{3}$$

The down-sampling function will average the sum of each $n \times n$ area in each input image, which causes the output image to be n-times compacted in both spatial dimensions.

2.3. Deep CNN for wind power prediction

The building block of deep CNN consists of a pair of hidden plies, namely, a convolution layer and a sub-sampling layer. Stacking the building blocks one-by-one hierarchically and adding a fully connected layer, such as a classifier, at the end of the stacks creates a typical deep CNN. The input of deep CNN contains a number of localized 2-D maps. The resolution of input maps becomes increasingly smaller as more convolution and down-sampling operations are applied [35]. The parameters in deep CNN, including the weights and bias, are trained in a data-driven fashion via the use of back propagation algorithm.

Deep CNN differs from SAE and DBM in two features. One is that CNN adopts a more concise deep architecture with fewer memory footprints and fewer parameters because of the use of weight sharing technique. Another is that the pre-training process in SAE and DBM is not required for CNN training, making CNN more appropriate for real-time implementation. These two features, coupled with its effective feature extraction and model recognition, make CNN an attractive option for WPF. However, classical deep CNN cannot be applied directly to WPF because wind power data is 1D data in a time domain. Therefore, 1D wind power data should be first converted into a 2D image for feature extraction, and then reconverted to a 1D vector for prediction. Therefore, a deep CNN architecture for WPF is proposed in this paper, as illustrated in Fig. 1. The architecture includes a 1D-Data-to-2D-Image layer, several building blocks consisting of a convolution layer and a sub-sampling layer, a 2D-Image-to-1D-Data layer and a logistic regression layer. The convolution layer and sub-sampling layer are used to extract the hidden features in wind power data, and the logistic regression layer is employed for WPF.

3. Deep CNN based point forecast of wind power

Owing to the chaotic nature of the weather system on earth, wind power data always exhibit high variability and volatility. Therefore, a hybrid point forecaster for WPF is proposed in this research to alleviate the impact of the uncertainties on WPF accuracy, The forecaster is a hybrid of WT and deep CNN, as shown in Fig. 2. The raw wind power data are first normalized and then decomposed into several frequencies. Each of the frequencies exhibits better data outliers with smaller degrees of uncertainty than the original signal. An independent deep CNN network is then designed for each frequency and trained in a back propagation manner to predict the behavior of each frequency as accurately as possible. Consequently, synthesizing all of the forecasting frequencies via wavelet reconstruction and anti-normalization produces the final results for the deterministic point forecast of wind power. Fig. 2 presents the schematic diagram of the proposed point forecaster. The details of each subpart are given below.

3.1. Wavelet decomposition

Generally, raw wind power data series may contain nonlinear and non-stationary features in the form of spikes and fluctuations. Indisputably, these features are the main attributes to deteriorate the accuracy of WPF [36]. One way to mitigate the impact of spikes and fluctuations on WPF accuracy is the use of WT [37]. Generally, WT can be defined in a discrete form to improve the computational efficiency, as follows

$$Wavelet(\upsilon,\kappa) = 2^{-(\upsilon/2)} \sum_{t=0}^{T-1} g(t)\phi[(t-\kappa 2^{\upsilon})/2^{\upsilon}] \tag{4}$$

In this paper, 4th Daubechies function is chosen to be the mother wavelet because it provides a proper balance between the wavelength and smoothness [38]. Therefore, in this research, a fast discrete WT, termed Mallat algorithm, is adopted. This algorithm is based on decomposition filters and reconstruction filters. Each filter corresponds to a low pass and a high pass. Thereafter, the raw wind power data series can be decomposed into one approximation (A_n) and several details (D_n). Compared to original wind power data series, the obtained approximation and details exhibit better outliers and lower uncertainties, making the decomposed signals easier for WPF.

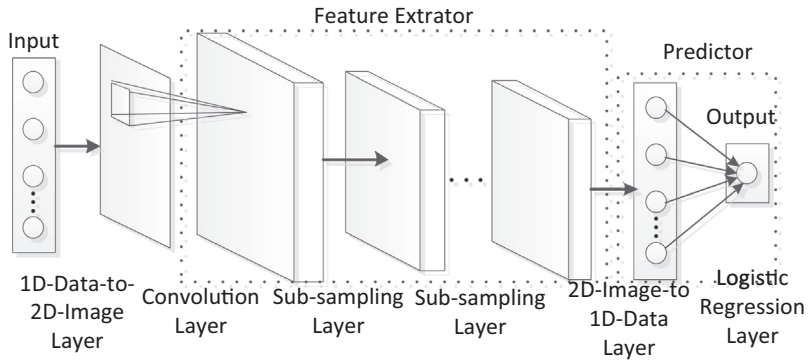

Fig. 1. Deep CNN architecture for wind power prediction.

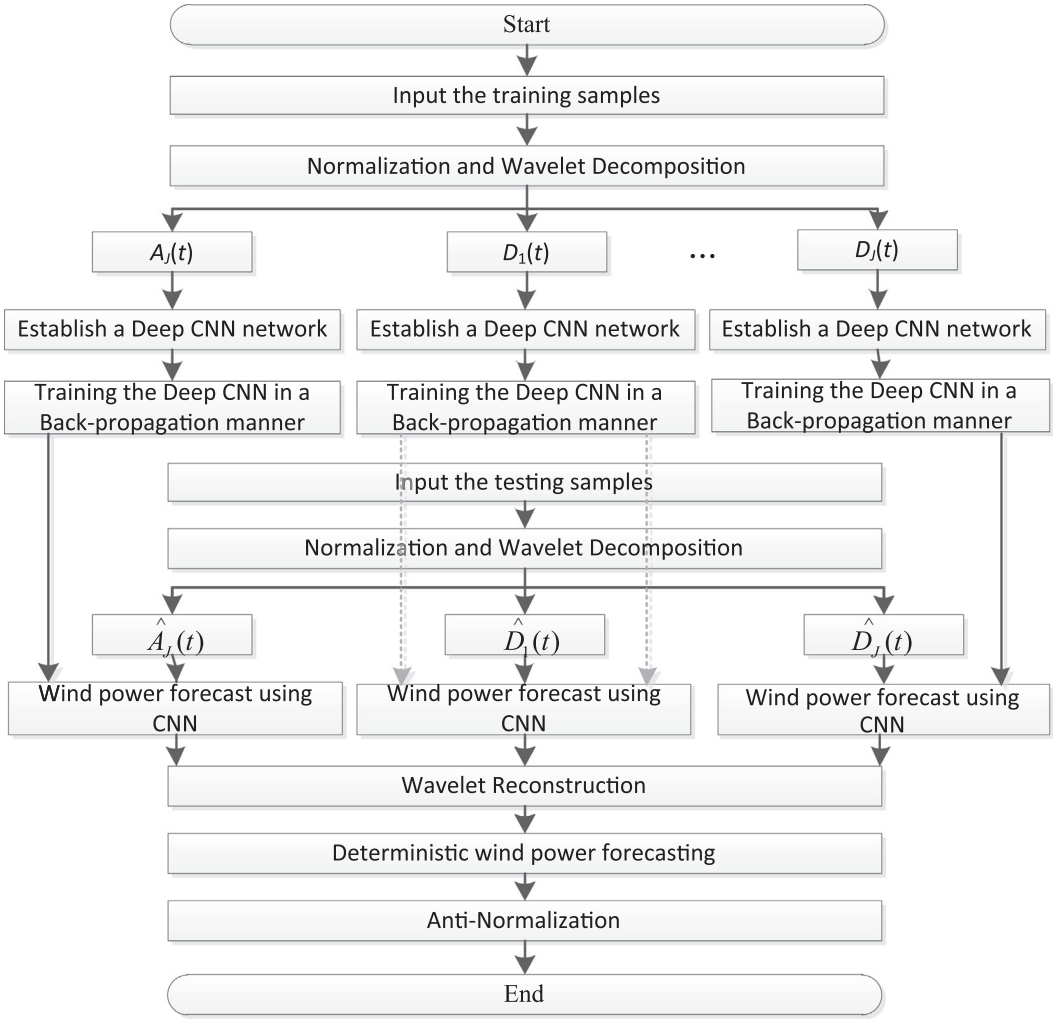

Fig. 2. Schematic diagram of the proposed point forecaster for WPF.

3.2. Conversion from 1D data to 2D image

As presented in Section 2.3, the wind power data series should first be converted to a 2D image to apply deep CNN for WPF. The 1D-Data-to-2D-Image layer is responsible for the conversion process. For simplicity, this paper proposes a rearrangement procedure to achieve the 1D-Data-to-2D-Image task, described as follows: (1) Determine the size of the image $p \times q$; (2) Given a wind power data sample, note that the length of data l should satisfy $l = p \times q$; (3) Sequentially arrange the first q data as the first row of the image, the next q data as the second row of the image, etc. Therefore, using these three steps, the 1D-Data-to-2D-Image process is easily completed through the 3 steps and its effectiveness is tested in Section 4. Similarly, the 2D-Image-to-1D-Data procedure is able to be achieved in a reversible way. In addition, the size of the image is adjusted in a trial-and-error manner to improve the forecast accuracy.

3.3. Back propagation training for deep CNN

To solve the WPF problem, the parameters of deep CNN, i.e., the weights and biases, should be trained by the back propagation rule applying stochastic gradient descent. Back propagation attempts to minimize the squared-error loss function E_m between the outputs and target as follows,

$$E_m = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{d}(h_j^i - y_j^i)^2 \tag{5}$$

Thus, the updating rule for weights $\boldsymbol{W}$ and biases $\boldsymbol{b}$, $\boldsymbol{\beta}$, $\boldsymbol{c}$ can be defined in an iterative form, as follows,

$$\boldsymbol{W} = \boldsymbol{W} - \eta \cdot \partial \boldsymbol{E}_m / \partial \boldsymbol{W} \tag{6}$$

$$\boldsymbol{b} = \boldsymbol{b} - \eta \cdot \partial \boldsymbol{E}_m / \partial \boldsymbol{b} \tag{7}$$

$$\boldsymbol{\beta} = \boldsymbol{\beta} - \eta \cdot \partial \boldsymbol{E}_m / \partial \boldsymbol{\beta} \tag{8}$$

$$\boldsymbol{c} = \boldsymbol{c} - \eta \cdot \partial \boldsymbol{E}_m / \partial \boldsymbol{c} \tag{9}$$

here, $\partial \boldsymbol{E}_m/\partial \boldsymbol{W}$, $\partial \boldsymbol{E}_m/\partial \boldsymbol{b}$, $\partial \boldsymbol{E}_m/\partial \boldsymbol{\beta}$, and $\partial \boldsymbol{E}_m/\partial \boldsymbol{c}$ are the partial derivatives of loss function with respect to perturbations of $\boldsymbol{W}$, $\boldsymbol{b}$, $\boldsymbol{\beta}$ and $\boldsymbol{c}$. However, three types of layers with parameters to be learned exist in deep CNN, i.e., the convolution layer, sub-sampling layer, and logistic regression layer. Therefore, because the update rules of the parameters in (6-9) depend on the type of layer, each layer should be analyzed separately.

Regarding logistic regression layer, the partial derivative $\partial \boldsymbol{E}_m/\partial \boldsymbol{b}_{\boldsymbol{log}}^{\boldsymbol{L}}$ is calculated as the element-wise product between the derivative of layer with respect to input and output error. Then, $\partial \boldsymbol{E}_m/\partial \boldsymbol{W}_{\boldsymbol{log}}^{\boldsymbol{L}}$ is calculated by scaling up the $\partial \boldsymbol{E}_m/\partial \boldsymbol{b}_{\boldsymbol{log}}^{\boldsymbol{L}}$ via the input of the layer [39], as follows,

$$\partial \boldsymbol{E}_m/\partial \boldsymbol{W}^{\boldsymbol{L}}_{\boldsymbol{log}} = \boldsymbol{x}^{L-1}(\partial \boldsymbol{E}_m/\partial \boldsymbol{b}^{\boldsymbol{L}}_{\boldsymbol{log}})^T \tag{10}$$

$$\partial \boldsymbol{E}_m/\partial \boldsymbol{b}^{\boldsymbol{L}}_{\boldsymbol{log}} = f'(\boldsymbol{u}^L) \circ (\boldsymbol{y}^d - \boldsymbol{h}^d) \tag{11}$$

where $\boldsymbol{u}^L = \boldsymbol{W}^L\boldsymbol{x}^{L-1} + \boldsymbol{b}^L$, and "$\circ$" denotes element-wise multiplication.

With respect to the convolution layer followed by a sub-sampling layer, it is difficult to estimate the involved partial derivatives since each neuron in an input map connects to only one neuron in the output map. In other words, the sensitivity maps for one output pixel correspond to a set of sensitivity input maps. Concretely, $\partial \boldsymbol{E}_{\boldsymbol{m}}/\partial \boldsymbol{b}^{\boldsymbol{l}}_{\boldsymbol{con}}$ is obtained by synthesizing all of the entries in the sensitivity output map. Each entry is calculated by up-sampling the sub-sampling layer's sensitivity maps to make it the same size as the convolutional layer's input map, and then multiplying the up-sampled map with the activation derivative map in an element-wise way. Thus, $\partial \boldsymbol{E}_m/\partial \boldsymbol{W}^{\boldsymbol{l}}_{\boldsymbol{con}}$ is then computed as the sum of the product between $\partial E_m/\partial \boldsymbol{b}^{\boldsymbol{l}}_{\boldsymbol{con}}$ and the corresponding patch vector [34] in $\boldsymbol{x}^{l-1}$ as follows,

$$\partial \boldsymbol{E}_m/\partial \boldsymbol{W}^{\boldsymbol{l}}_{\boldsymbol{con}} = \sum_{len,wid} \boldsymbol{\beta}^{l+1}(f'(\boldsymbol{u}^l) \circ up(\partial \boldsymbol{E}_m/\partial b^{\boldsymbol{l+1}}_{\boldsymbol{con}}))(\boldsymbol{patch}^{l-1}) \tag{12}$$

$$\partial \boldsymbol{E}_m/\partial \boldsymbol{b}^{\boldsymbol{l}}_{\boldsymbol{con}} = \sum_{len,wid} \boldsymbol{\beta}^{l+1}(f'(\boldsymbol{u}^l) \circ up(\partial \boldsymbol{E}_m/\partial b^{\boldsymbol{l+1}}_{\boldsymbol{con}})) \tag{13}$$

Generally, one effective way to implement the up-sampling function is the use of the Kronecker product, as given below,

$$up(\boldsymbol{x}) \equiv \boldsymbol{x} \otimes 1_{n\times n} \tag{14}$$

Considering the sub-sampling layer, the back propagation rules depend on the location of the layer. If the sub-sampling layer connects the logistic regression layer, then the update rules are the same with standard back propagation equations [39]. If not, the update rule for additive bias is the sum of the elements in the corresponding output patch. Accordingly, the update rule for multiplicative bias can be computed via a down-sampling operation, as follows,

$$\partial \boldsymbol{E}_m/\partial \boldsymbol{c}^{\boldsymbol{l}}_{\boldsymbol{sub}} = \sum_{j\in map} \left(f'\left(\boldsymbol{u}^l_j\right) \circ \left(\partial \boldsymbol{E}_m/\partial \boldsymbol{b}^{\boldsymbol{l+1}}_{\boldsymbol{j,con}} \otimes \boldsymbol{W}^{\boldsymbol{l+1}}_{\boldsymbol{j,con}}\right)\right) \tag{15}$$

$$\partial \boldsymbol{E}_m/\partial \boldsymbol{\beta}^{\boldsymbol{l}}_{\boldsymbol{sub}} = \left(\sum_{j\in map} \partial \boldsymbol{E}_m/\partial c^l_{sub,j} \circ down(\boldsymbol{x}^{l-1}_j)\boldsymbol{\beta}^{l+1}\right) \tag{16}$$

Therefore, all of the parameters in deep CNN can be periodically updated in a top-down manner based on (6-16) until they all converge. One period means the parameters of the entire model are updated once, resulting in smaller errors. The errors are then back propagated through the training set to re-correct the model parameters towards the optimal states. The training process is completed when all of the parameters find their optimal states.

4. Uncertainty evaluation for point forecast

Due to the chaotic nature of the weather system on earth, wind power data always exhibit nonlinear and non-stationary uncertainties. These uncertainties are mainly resulted from model misspecification and data noise, which are estimated as follows.

4.1. Uncertainties of WPF

The uncertainties exhibited in wind power data are inevitable due to the misspecification of the forecasting model and data noise [40]. The former, also termed model uncertainty, may be caused by a local minimum in the training process, finite training samples, and stochastically generated parameters. The latter may be because wind power time-series data have stochastic characteristics because of the chaotic weather system. These uncertainties have a significant side impact on prediction accuracy because it is difficult to estimate them in a deterministic manner. Therefore, in this research, these two types of uncertainties are probabilistically modeled as a Gaussian distribution that depends on the input, as follows,

$$\varepsilon(\boldsymbol{x_i}) \approx GD(M_\varepsilon(\boldsymbol{x_i}), \sigma^2_\varepsilon(\boldsymbol{x_i})) \tag{17}$$

Note that the adopted Gaussian distribution can fit different shapes of distributions. Even if the actual uncertainty is not Gaussian-like, it has been demonstrated in [41] that the prediction model based on Gaussian distribution can also be implemented with competitive performance.

4.2. Model uncertainty evaluation

The model uncertainty is estimated via an ensemble of individual deep CNNs, thus leading to a less biased approximation of the true regression of the measured targets. In the proposed ensemble approach, the adopted deep CNN model differs from the other models in terms of the number of hidden layers, the number of maps in each layer, and the length of input vector. Considering a training dataset consisting of N_S samples

$$DS_{mu} = (\boldsymbol{x}_1, \boldsymbol{h}_1), \ldots, (\boldsymbol{x}_i, \boldsymbol{h}_i), \ldots, (\boldsymbol{x}_N, \boldsymbol{h}_{N_S}) \tag{18}$$

and an ensemble of N_E deep CNN models, the true regression is estimated as the average output of the ensemble, as follows,

$$\widehat{y}(\boldsymbol{x_i}) = \frac{1}{N_E}\sum_{j=1}^{N_M} \widehat{y}_l(\boldsymbol{x_i}) \tag{19}$$

Then, the variance of the model uncertainty is evaluated as the variance of the output set of the well-trained CNN models, expressed as

$$\sigma^2_{mu}(\boldsymbol{x_i}) = \frac{1}{N_E - 1}\sum_{j=1}^{N_E}\left(\widehat{y}_l(\boldsymbol{x_i}) - \widehat{y}(\boldsymbol{x_i})\right)^2 \tag{20}$$

4.3. Data uncertainty evaluation

It is challenging to estimate the data uncertainty of wind power due to the erratic nature of the weather system. In this research, it is assumed that the mean and variance of data uncertainty are also conditioned on the input variables. Given the result of the point forecaster $\widehat{y}(\boldsymbol{x_i})$, the input dataset for data uncertainty is expressed as follows,

$$DS_{du} = (\boldsymbol{h}_1 - \widehat{y}(\boldsymbol{x_i})), \ldots, (\boldsymbol{h}_{N_S} - \widehat{y}(\boldsymbol{x_i})) \tag{21}$$

Then, the mean and variance of the data noise can be calculated as the follows,

$$M_{du} = \frac{1}{N_S}\sum_{l=1}^{N_S}(\boldsymbol{h}_l - \widehat{y}(\boldsymbol{x_l})) \tag{22}$$

$$\sigma^2_{du} = \frac{1}{N_S - 1}\sum_{l=1}^{N_S}((h_l - \widehat{y}(\boldsymbol{x_l})) - M_{du})^2 \tag{23}$$

5. Probabilistic WPF based on ensemble approach

In this section, a novel approach for probabilistic WPF based on ensemble technique is proposed. Moreover, the performance criteria for probabilistic WPF are also presented.

5.1. Probabilistic WPF approach

Due to the variability and volatility of wind power data, a novel approach is originally proposed to evaluate the probabilistic information in wind power data. This approach is a hybrid of WT, CNN and ensemble technique. Firstly, the input dataset are uniformly resampled to form N_E point forecasters, also termed N_E replicates. For each replicate, an independent point prediction model for WPF is designed. Then, based on the prediction results, the mean and variance of the model uncertainty and data noise are separately evaluated by using (18-20) and (21-23). Assuming that these two uncertainties are independent from each other, then the variance of total forecast error ${\sigma_h}^2$ can be obtained by synthesizing these two uncertainties, i.e.,

$$\sigma_h^2(\boldsymbol{x_i}) = \sigma_{mu}^2(\boldsymbol{x_i}) + M_{du} + \sigma_{du}^2 \tag{24}$$

Therefore, given the target h_i and confidence level $100(1-\alpha)\%$, the prediction intervals can be effectively constructed as a lower bound $L_h^{\alpha}(\boldsymbol{x_i})$ and an upper bound $U_h^{\alpha}(\boldsymbol{x_i})$, as follows,

$$L_h^{\alpha}(\boldsymbol{x_i}) = \widehat{y}(\boldsymbol{x_i}) - z_{1-\alpha/2}\sqrt{\sigma_h^2(\boldsymbol{x_i})} \tag{25}$$

$$U_h^{\alpha}(\boldsymbol{x_i}) = \widehat{y}(\boldsymbol{x_i}) + z_{1-\alpha/2}\sqrt{\sigma_h^2(\boldsymbol{x_i})} \tag{26}$$

The overall architecture for probabilistic WPF is illustrated in Fig. 3. The performance of the proposed approach has been completely benchmarked and tested in Section 4.

5.2. Performance criteria

In general, probabilistic WPF adopts average coverage error (ACE) and interval sharpness (IS) as the performance criteria. ACE measures how well the prediction quantiles matches the observed values, and IS is to comprehensively evaluate the sharpness of the PI, by rewarding the narrower PIs and penalizing wider PIs, expressed as follows.

$$ACE = \frac{1}{N_S}\sum_{i=1}^{N_S} r_i \times 100\% - \text{PINC} \tag{27}$$

$$IS = \frac{1}{N_S}\sum_{i=1}^{N_S}\begin{cases} -2\alpha\delta_i^{\alpha} - 4[L_i^{\alpha} - WS_i^a], & if\ WS_i^a < L_i^{\alpha} \\ -2\alpha\delta_i^{\alpha} & if\ WS_i^a \in I_i^{\alpha} \\ -2\alpha\delta_i^{\alpha} - 4[WS_i^a - U_i^{\alpha}], & if\ WS_i^a > U_i^{\alpha} \end{cases} \tag{28}$$

The width of PI δ_i^a can be computed as $(U_i^{\alpha} - L_i^{\alpha})$, and indicator r_i is defined as,

$$r_i = \begin{cases} 0, & WS_i^a \in I_i^a \\ 1, & WS_i^a \notin I_i^a \end{cases} \tag{29}$$

Another widely used performance criteria for probabilistic WPF is CRPS that considers both reliability and sharpness simultaneously [42]. Generally, given the cumulative distribution function CDF_i and the measurement y_i over the testing sample, the average value of CRPS can be calculated as [43]

$$CRPS = \frac{1}{N_s}\sum_{i=1}^{N_s}\int_{y=0}^{\infty}[CDF_i - H(y-y_i)]^2 dy \tag{30}$$

The value of indicative function $H(y\text{-}y_i)$ is 0 if $y < y_i$, otherwise, the value is 1.

6. Numerical results and analysis

In this research, the proposed approach for probabilistic WPF based on WT, deep CNN and ensemble technique is extensively evaluated and benchmarked using real data from the MWWF in Shandong Province, China, and the SIWF in Guangdong Province, China.

6.1. Investigations on milky way wind farm

6.1.1. Experimental settings

The MWWF has a rated capacity of 47.5 MW in combination with 19 wind turbines of 2.5 MW each. The wind power data from MWWF are sampled in 5-min intervals and cover the period from Jan. 2011 to Dec. 2011. The wind power data are divided into a training dataset and a testing dataset. The training dataset covers the days from the 1st to the 25th each month, and the remainder comprise the testing dataset. Based on the training dataset, the proposed approach specific for the training dataset is applied for probabilistic WPF. For each point forecaster, the input parameters are the wind power data at a current time step WP_t, and the previous 8, 15 or 24 values. The input wind power data series are then decomposed into three frequencies, including one approximation and two details, or four frequencies, including one approximation and three details. For each frequency, the building blocks consist-

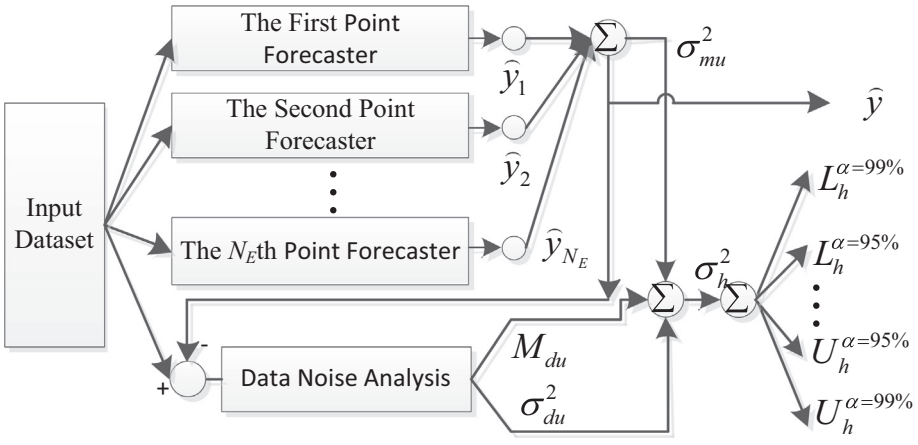

Fig. 3. The overall architecture of the proposed approach for probabilistic WPF.

ing of one convolution layer and one subsampling layer adopt 3, 4, 5 or 6. Therefore, there are $3 \times 2 \times 4 = 24$ replicates, i.e., $N_E = 24$. These parameters are determined in a trial-and-error manner. In addition, the wavelet parameters in each replicate adopt the default values in MATLAB wavelet toolbox because the results are not sensitive to these parameters. All of the model parameters in CNN, including the weights $\boldsymbol{W}$ and biases $\boldsymbol{b}$, $\boldsymbol{\beta}$, $\boldsymbol{c}$, are randomly initialized and updated based on the training process presented in Section 3.3 until they all converge. Moreover, to mitigate the impact of seasonal uncertainty on prediction accuracy, the proposed WPF approach is seasonally trained and evaluated because of the erratic nature of the weather system.

Furthermore, to validate the high-efficiency and high-accuracy of the proposed approach, the obtained results are compared to classical WPF methods, including persistence, SVM, and BP. The mean of the persistence model adopts the last available observation and the variance is calculated using the latest measurements. The SVM and BP are extended to probabilistic WPF via QR, which is frequently used for probabilistic WPF. The prediction algorithms are implemented in MATLAB R2014a and conducted on a personal computer with an Intel(R) Xeon(R) E3-1225 V2 3.2-GHz CPU and 16.00 GB of RAM.

6.1.2. Numerical results

A series of simulations are conducted to demonstrate the feasibility and effectiveness of the proposed approach. Here, the performance of probabilistic WPF is evaluated in terms of ACE and IS. The 1-h ahead seasonal results obtained from the four benchmarks, i.e., ACE and IS, are presented in Tables 1 and 2 over a high confidence level ranging from 85% to 99% because of the high reliability required for power system optimization and operation. In this case, the constructed seasonal PIs with PINC 90% obtained from the proposed approach are graphically presented in Figs. 4–7. In addition, Figs. 8 and 9 depicts the reliability deviations and IS in spring and winter over the entire confidential level range with a step of 5%. Moreover, to further demonstrate the advantages of the proposed probabilistic WPF approach, a series of simulations using different time resolutions has been conducted and the numerical results are shown in Fig. 10. In this simulation, the training dataset covers the whole year and are obtained by interval sampling of the original wind power dataset. The results of SVM are only presented because of the relatively superior performance in 1-h ahead prediction tasks.

6.1.3. Analysis

From Table 1, it is evident that the ACEs obtained from the proposed ensemble approach perform the best in all four seasons. Quantitatively, the ACEs from the proposed WPF approach have a minimum of −1.81% and a maximum of 2.96% with an average of 0.16%. However, the average ACEs obtained from persistence, BP + QR, and SVM+QR correspond to −6.81%, −3.51%, and −1.72, respectively. The numerical results indicate that the resultant ACEs from the proposed approach are closer to the corresponding nominal confidence level, especially at the 95% and 99% PINC levels. Therefore, compared to the three benchmarks, the proposed approach exhibits higher prediction capability in term of reliability.

The presented ISs in Table 2 indicate that, compared to persistence, BP+QR, and SVM+QR, the ISs obtained from the proposed approach have been evenly improved by 58.94%, 60.56%, and 58.33% at PINC 85%, 61.81%, 62.67% and 58.48% at PINC 90%, 64.79%, 61.46% and 58.10% at PINC 95%, 79.26%, 58.02% and 61.94% at PINC 99%, respectively. These results prove that at a high confidence level, the sharpness errors between the observed probability and nominal confidence obtained from the proposed approach are at a minimum. Apparently, the IS performance demonstrates that the proposed approach has a higher forecast capability in term of sharpness and is thus more appealing when compared to the other three benchmarks.

Figs. 4–7 indicate that the measured wind power data are, for the most part, within the lower and upper bounds of the constructed PIs. In addition, it is noted that the shapes of the three lines in each of the four graphs are very similar to each other. Therefore, from the high PI coverage and the similarity of the lines, it can be concluded that the probabilistic performances obtained from the proposed approach are satisfactory. In addition, it is evident that the actual wind power data in spring and winter have more zero values than the data in summer and autumn. Meanwhile, the line trends of the constructed PIs in Figs. 5 and 6 exhibit stronger fluctuations than those in Figs. 4 and 7. This is because the weather systems of MWWF in summer and autumn are more erratic than those in spring and winter, and thus it is more difficult to make accurate prediction.

Fig. 8 indicates that the ACE deviations obtained from the proposed approach vary at a lesser value than those obtained from SVM in all confidential levels and seasons. The trends of the ACE deviations appear to be irregular and chaotic. From Fig. 9, it can be concluded that the sharpness of the quantiles obtained from the proposed approach is better than that obtained from the

Table 1
1-Hour ahead forecasting ACE for MWWF.

Method	Season	PINC 85%	PINC 90%	PINC 95%	PINC 99%
Persistence	Spring	−3.85%	−5.18%	−6.52%	−5.81%
	Summer	−1.23%	−4.66%	−8.09%	−8.95%
	Autumn	−5.94%	−8.32%	−7.57%	−7.90%
	Winter	−5.94%	−8.85%	−10.71%	−9.47%
BP+QR	Spring	−3.85%	−2.04%	−2.33%	−0.57%
	Summer	−2.28%	−0.99%	−1.81%	−0.05%
	Autumn	−4.37%	−3.09%	−4.95%	0.48%
	Winter	−3.83%	−10.42%	−12.28%	−3.71%
SVM+QR	Spring	−3.85%	−1.52%	1.28%	−1.09%
	Summer	−5.16%	2.67%	1.86%	−0.58%
	Autumn	−3.32%	−4.14%	1.86%	−0.57%
	Winter	−5.42%	−4.66%	−2.33%	−2.14%
Proposed approach	Spring	**−1.23%**	**−1.52%**	**−0.29%**	**−0.50%**
	Summer	**1.91%**	**0.05%**	**2.91%**	**−0.05%**
	Autumn	**2.96%**	**2.67%**	**−0.76%**	**−0.57%**
	Winter	**1.91%**	**−0.99%**	**−1.81%**	**−1.62%**

The bold values denote the best performance among the benchmarks.

Table 2
1-Hour ahead forecasting IS for MWWF.

Method	Season	PINC 85%	PINC 90%	PINC 95%	PINC 99%
Persistence	Spring	−11.90	−9.43	−6.51	−3.11
	Summer	−5.81	−4.82	−3.61	−2.17
	Autumn	−7.98	−6.32	−4.23	−1.89
	Winter	−11.89	−9.70	−6.81	−3.44
BP+QR	Spring	−11.38	−8.59	−5.00	−1.17
	Summer	−6.95	−5.83	−3.81	−0.92
	Autumn	−8.39	−6.49	−3.97	−1.09
	Winter	−12.40	−10.06	−6.55	−2.06
SVM+QR	Spring	−10.40	−7.81	−4.48	−1.03
	Summer	−6.89	−4.42	−2.83	−0.78
	Autumn	−8.22	−6.24	−3.77	−1.14
	Winter	−11.52	−9.37	−6.70	−2.83
Proposed Approach	Spring	**−5.68**	**−4.20**	**−2.72**	**−0.78**
	Summer	**−3.88**	**−3.02**	**−1.94**	**−0.66**
	Autumn	**−1.60**	**−1.16**	**−0.73**	**−0.21**
	Winter	**−4.27**	**−3.18**	**−2.06**	**−0.55**

The bold values denote the best performance among the benchmarks.

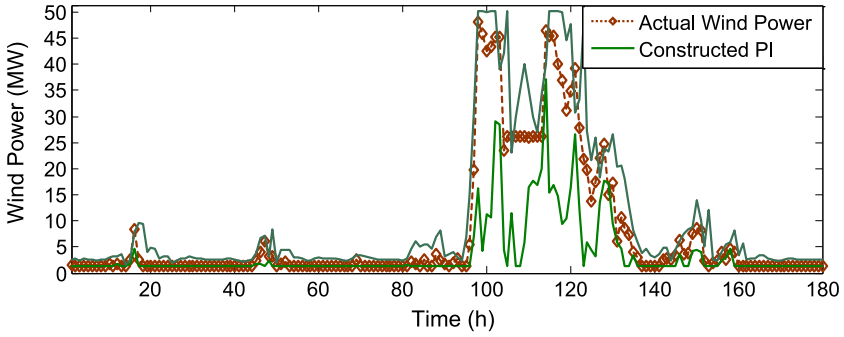

Fig. 4. PIs with PINC 90% in Spring 2011 at MWWF obtained from the proposed approach.

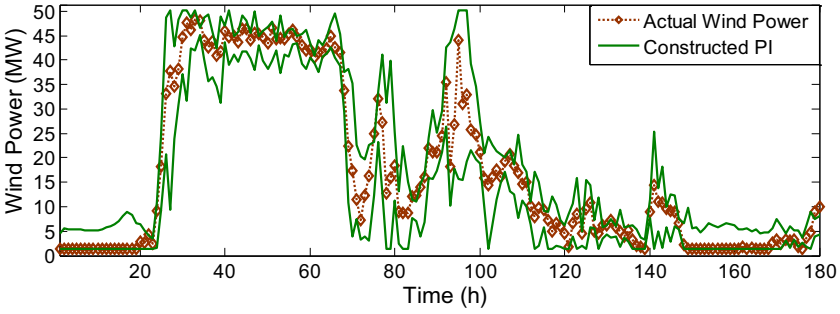

Fig. 5. PIs with PINC 90% in Summer 2011 at MWWF obtained from the proposed approach.

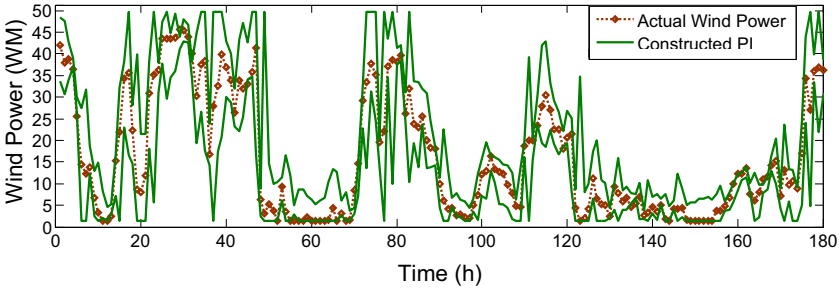

Fig. 6. PIs with PINC 90% in Autumn 2011 at MWWF obtained from the proposed approach.

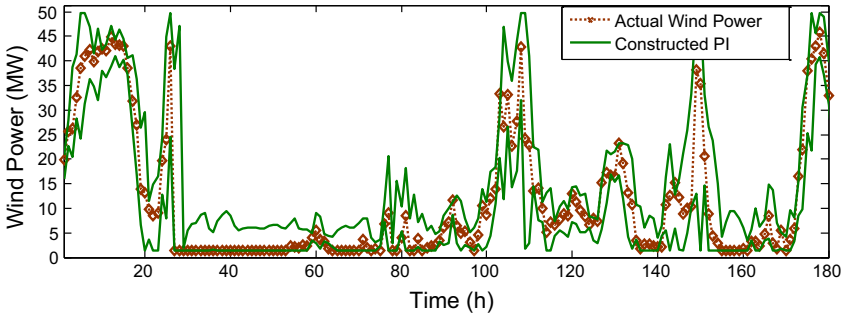

Fig. 7. PIs with PINC 90% in Winter 2011 at MWWF obtained from the proposed approach.

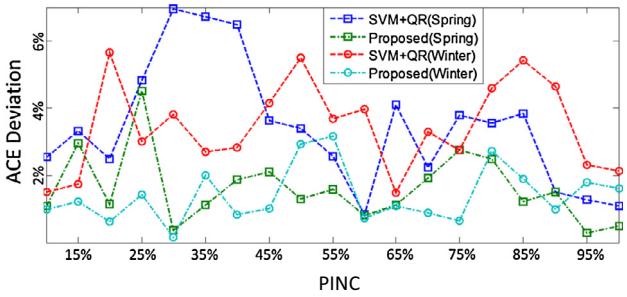

Fig. 8. The reliability deviation for 1-h-ahead prediction.

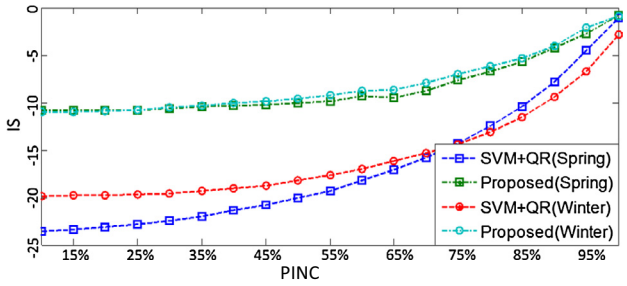

Fig. 9. The IS for 1-h-ahead prediction.

SVM in all of the confidential levels and seasons. The values of IS tend to increase as the PINC increases. Therefore, the ACE and IS performances in Figs. 8 and 9 illustrate that the proposed approach performs the best in all confidential levels and seasons.

In Fig. 10, it is obvious that the proposed approach performs better than the three benchmarks with respect to time resolutions and confidential levels. The ACE deviation from the proposed approach ranges from 0.18% to 7.00% at PINC 85%, from 0.47% to 4.24% at PINC 90%, from 0.24% to 3.43% at PINC 95%, and from 0.48% to 3.19% at PINC 99%, respectively. The average value of the deviations is 1.91%. The absolute values of IS vary from 1.36 to 8.25 at PINC 85%, from 1.04 to 6.25 at PINC 90%, from 0.73 to 3.83 at PINC 95%, and from 0.24 to 1.20 at PINC 99%. The average value is 3.20. Compared to SVM, the performances of the ACE and IS of the proposed approach are evenly improved by 50.84% and 40.51%, respectively.

The ACE and IS are two typical metrics that are frequently used to evaluate the effectiveness and feasibility of probabilistic WPF. Based on the results presented herein, it is evident that the proposed probabilistic WPF approach outperforms the three benchmarks not only with respect to reliability but also from the perspective of sharpness. The high-accuracy and superiority of the proposed approach are mainly derived from the deep NN architecture, which provides an effective way to approximate the inherent invariant features and hidden structures. Therefore, the high-level nonlinear, non-stationary and non-smoothness exhibited in the wind power can be better extracted.

6.2. Investigations on Shangchuan island wind farm

6.2.1. Experimental settings

The SIWF is situated on the southern coast of China. The raw wind power data collected from SIWF cover the period from Jan. 2013 to Dec. 2013 with a 15-min resolution. The SIWF has a rated capacity of 48.45 MW with 57 wind turbines of 0.85 MW each. The dataset is divided into the training dataset and testing dataset. Similarly, the model for probabilistic WPF is seasonally designed because of the erratic nature of wind power data. The design process is similar to the process described in VI-A. Consequently, the obtained WPF model consists of 24 replicates, and the model parameters can thus be trained for WPF. The benchmarks for performance comparison also adopt persistence, BP and SVM.

6.2.2. Numerical results

To comprehensively demonstrate the overall privilege of the proposed approach, the performance criteria in this subsection adopts CRPS, which is widely used for probabilistic WPF. The advantage of CRPS is that it can simultaneously address both reliability and sharpness. Table 3 presents the CRPS data in terms of seasons and prediction horizons. The prediction horizons range from 15-min ahead to 8-h ahead. The CRPS indices are obtained by using the created quantiles ranging from 5% to 99%. The wind power density in each quantile is assumed to be uniform [42]. In the numerical simulation, a moving average process is adopted to obtain adequate wind power samples for probabilistic WPF over various prediction horizons. Moreover, Figs. 11–17 present the relationships between CRPS and its corresponding wind power. The presented samples are randomly selected from the result pool. Furthermore, the overall performance of the proposed approach somehow depends on the number of replicates, i.e., point forecaster. Therefore, a series of sensitivity analyses was carried out to provide more realistic and persuasive comparisons. The related CRPS is presented in Table 4. The number of replicates in the proposed approach adopts 8, 24, 48 and 96.

6.2.3. Analysis

From Table 3, the CRPS values obtained from the proposed approach has a minimum of 0.2809 and a maximum of 1.7813 with an average of 0.9410 at 15-min-ahead prediction, a minimum of 1.3444 and a maximum of 1.7476 with an average of 1.6101 at 30-min-ahead prediction, a minimum of 1.4532 and a maximum of 2.4334 with an average of 1.9381 at 1-h-ahead prediction, and a minimum of 2.1296 and a maximum of 2.7067 with an average of 2.3270 at 2-h-ahead prediction. In addition, the ranges of CRPS

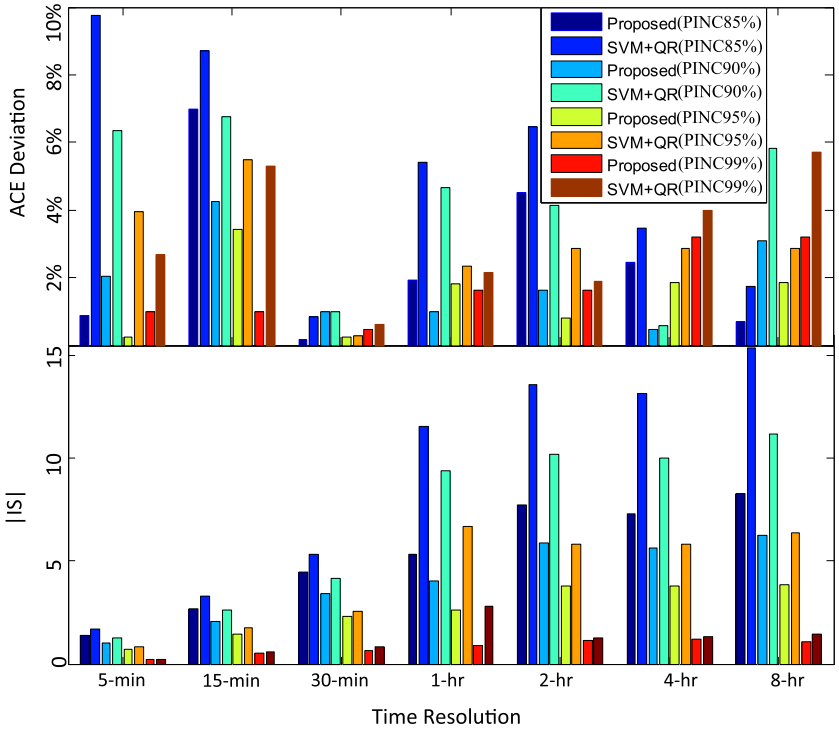

Fig. 10. ACE and IS in different time resolutions.

Table 3
The CRPS in different prediction horizons.

Method	Prediction horizon	Spring	Summer	Autumn	Winter
Persistence	15-min	2.2102	0.7196	0.3514	1.0693
	30-min	2.8667	2.0422	1.6968	2.9776
	1-h	2.5799	3.5908	2.5862	3.5578
	2-h	4.3073	2.4523	3.6326	6.0400
	4-h	5.4908	2.9488	5.5191	7.3005
	6-h	7.6553	3.3618	5.2071	7.2689
	8-h	9.2039	3.3239	5.7837	9.1547
BP+QR	15-min	2.4195	0.6995	0.4243	1.2663
	30-min	3.2270	3.4703	2.0680	3.1258
	1-h	3.0575	5.3135	3.2823	3.6227
	2-h	4.6495	4.1036	4.7229	6.3712
	4-h	6.5049	3.5162	5.9394	8.0313
	6-h	8.3627	3.9519	5.7148	9.0757
	8-h	10.0314	3.4295	6.2314	10.0169
SVM+QR	15-min	1.9264	0.6611	0.4486	1.1397
	30-min	2.6813	2.3440	1.6697	2.7172
	1-h	2.5362	3.4096	2.7210	3.2541
	2-h	4.4057	3.1566	4.0988	5.7870
	4-h	6.1283	3.2145	5.9401	6.3476
	6-h	7.0058	2.7965	4.7600	4.3214
	8-h	7.5087	2.6379	4.2695	5.4061
Proposed Approach	15-min	**1.7813**	**0.6397**	**0.2809**	**1.0621**
	30-min	**1.7476**	**1.6313**	**1.3444**	**1.7172**
	1-h	**1.4532**	**2.4334**	**1.8367**	**2.0293**
	2-h	**2.1489**	**2.1296**	**2.3229**	**2.7067**
	4-h	**2.8531**	**1.9691**	**2.7849**	**3.5409**
	6-h	**3.5699**	**1.9162**	**2.5432**	**3.6821**
	8-h	**4.2750**	**1.8777**	**2.8082**	**4.3385**

The bold values denote the best performance among the benchmarks.

are from 1.9691 to 3.5409 at 4-h-ahead prediction, from 1.9162 to 3.6821 at 6-h-ahead prediction, and from 1.8777 to 4.3385 at 8-h-ahead prediction. It is evident that the proposed approach exhibits the most ideal CRPS values in all of the prediction horizons. Quantitatively, at 15-min-ahead prediction, the CRPS performance of the proposed approach has been improved, on average, by 13.48%, 21.74%, and 9.86%, respectively, compared to persistence, BP and SVM. At 30-min-ahead prediction, the improvements are 32.79%, 45.84% and 31.57%, respectively. At 1-h-ahead prediction and 2-h-ahead prediction, the improvements correspond to 37.05%,

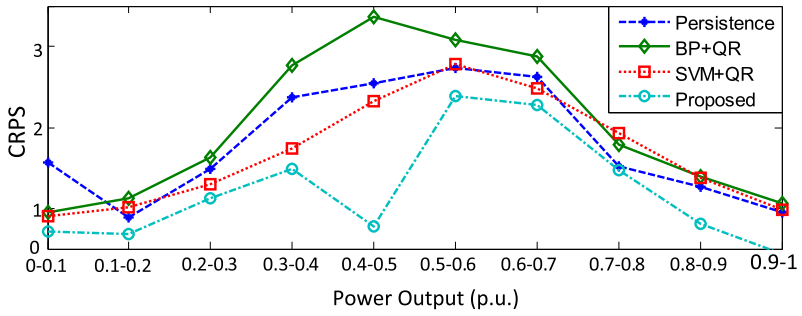

Fig. 11. The change of CRPS for 15-min-ahead WPF in winter.

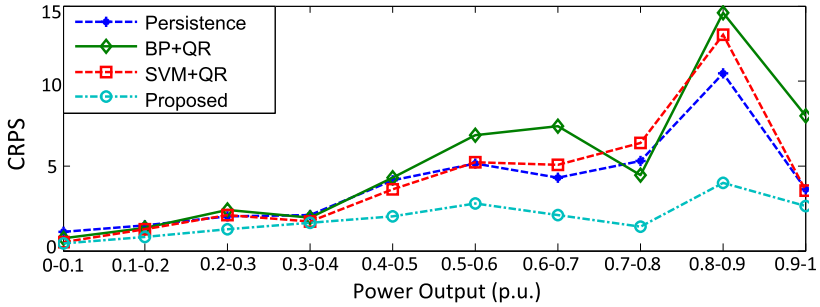

Fig. 12. The change of CRPS for 30-min-ahead WPF in spring.

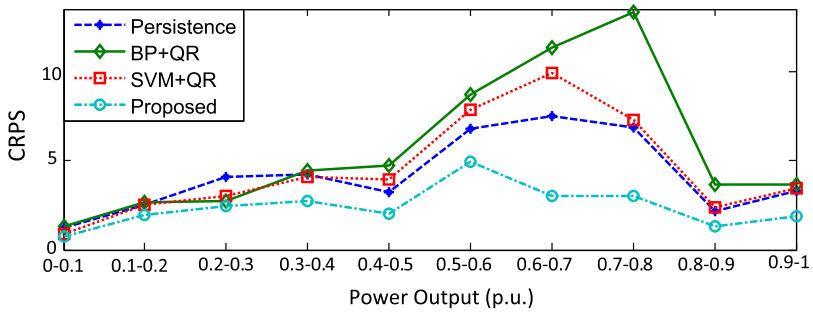

Fig. 13. The change of CRPS for 1-h-ahead WPF in autumn.

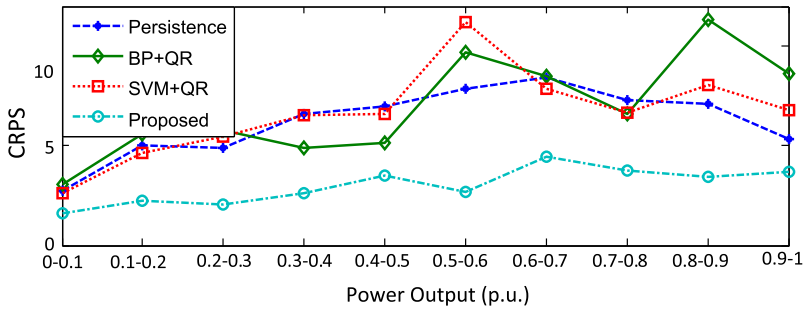

Fig. 14. The change of CRPS for 2-h-ahead WPF in summer.

49.25%, 34.97%, and 43.35%, 53.10%, 46.65%, respectively. In addition, compared to the three benchmarks, the CRPS performances have been evenly improved by 48.42%, 45.02%, and 45.10% when the prediction horizon ranges from 4-h ahead to 8-h ahead. The numerical results presented in Table 3 further demonstrate that the proposed approach has a higher forecasting ability with respect to overall skill, i.e., CRPS. Moreover, it can be easily concluded that the CRPS performance deteriorates substantially as the prediction horizon increases. This is because the long-term wind power data exhibit a more erratic nature and thus are more difficult to predict.

Figs. 11–17 indicates that the value of CRPS at power output 0–0.1 p.u. is approximately 2 in all cases. This value is relatively small when compared to the values at other output levels. The value of CRPS then continues to increase as wind power output increases until peaking at an output level between 0.5 p.u. and 0.7 p.u. Thereafter, the value of CRPS gradually decreases as the power out-

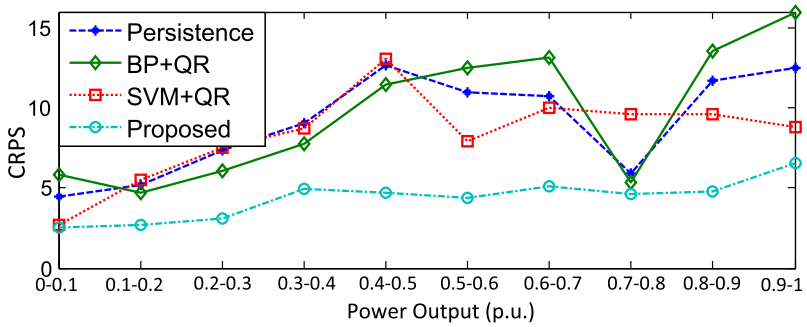

Fig. 15. The change of CRPS for 4-h-ahead WPF in spring.

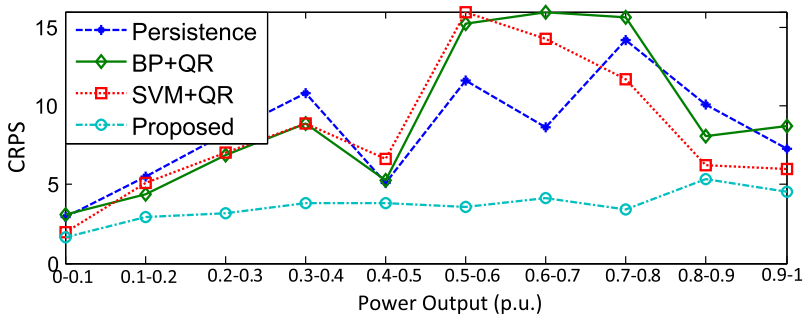

Fig. 16. The change of CRPS for 6-h-ahead WPF in winter.

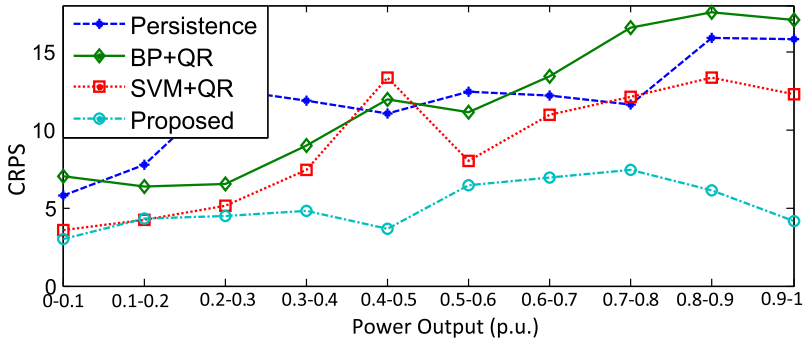

Fig. 17. The change of CRPS for 8-h-ahead WPF in summer.

Table 4
The Chang of CRPS with the number of replicates.

Prediction horizon	Season	Replicates			
		8	24	48	96
15-min	Spring	2.8731	1.7813	1.4254	1.2939
	Autumn	1.2845	0.2809	0.2503	0.2456
30-min	Spring	3.2492	1.7476	1.5602	1.3861
	Autumn	2.4735	1.3444	0.9880	0.9204
1-h	Spring	3.9830	1.4532	1.2122	1.0723
	Autumn	2.9028	1.8367	1.5039	1.3748
2-h	Spring	4.5097	2.1489	1.8023	1.7399
	Autumn	3.8721	2.3229	1.9304	1.7430
4-h	Spring	4.1662	2.8531	2.6303	2.5011
	Autumn	4.5396	2.7849	2.5116	2.5020
6-h	Spring	5.5102	3.5699	3.3307	3.1984
	Autumn	4.2019	2.5432	2.1787	2.0641
8-h	Spring	6.6488	4.2750	3.9215	3.7714
	Autumn	5.0361	2.8082	2.6601	2.5646

put increases. In addition, the values of CRPS at 0.9 to 1.0 p.u. output levels are usually below 5 and thus slightly greater than the CRPS value at 0–0.1 p.u. output level. Therefore, from the line trends of CRPS in Figs. 11–17, it can be concluded that the wind power data at a lower or higher output level are less difficult to predict and thus exhibit a relatively smaller degree of uncertainty than the data in the mid-range of output levels. Moreover, the CRPS obtained from the proposed approach is always the best regardless of seasons, power output levels or prediction horizons. The comparatively better performances also demonstrate that the proposed approach has the most robust prediction capability among the benchmarks. Furthermore, the CRPS performance of SVM is within the ideal range in same cases but may become unacceptable in other cases. Persistence and BP perform likewise. Therefore, the results demonstrate that the proposed approach not only significantly improves the prediction performance, but it also exhibits high stability and strong robustness. Thus the superiority and potentiality of the proposed approach are further illustrated.

From Table 4, it is evident that the value of CRPS in all of the cases gradually decreases as the number of replicates increases. This is because the increase of replicates undoubtedly reduces the uncertainties exhibited in wind power data, i.e., model misspecification and data noise. However, when the replicates increase to a certain level, the CRPS performance will no longer improve but rather suffer small fluctuations. It is further noted that the values of CRPS worsen as the prediction horizon increases. This is because long-term wind power data exhibit a more erratic nature and are thus more difficult to predict. On the other hand, the simulation for 15-min-ahead WPF accounts for 3 min. There are 24 replicates and $4 \times 24 \times 25 \times 3 = 7200$ samples in the simulation. When there are 48 and 96 replicates, although the CRPS performance improves, the calculation time doubles in size and the algorithm efficiency thus becomes unsatisfactory. Taken together, 24 replicates is a compromise solution between algorithm efficiency and overall performance. However, it should be noted that the proposed approach is not the most efficient among the benchmarks from the perspective of calculation time. This result is understandable because the CNN-based WPF architecture contains a signal decomposition process and four independent CNN networks, each of which would require a substantial amount of time.

The results in Tables 1–4 and Figs. 4–17 demonstrate that the proposed ensemble approach exhibits the best prediction capability not only from the perspective of reliability and sharpness, but also from that of overall performance. Two important aspects account for the relatively better performance of the proposed approach based on WT, deep CNN and the ensemble technique. First, the nonlinear, non-stationary features and deep hidden invariant structure exhibited in wind power data cannot be fully modeled by shallow NN models, such as BP and SVM. On the contrary, the primary advantage of CNN is the compact representation of a large set of hidden functions, which leads CNN to learn in a part-whole-decomposition manner that provides an effective way to extract the highly-varying inherent features and hidden invariant structure. Second, the ensemble technique is capable of mitigating the wind power uncertainties with respect to model misspecification and data noise. This is because the diverse errors of individual forecasters are stochastically distributed on the input space and can thus be cancelled out in the ensemble process. Therefore, considering forecasting reliability, sharpness and overall performance, the proposed approach exhibits much better comprehensive performances than the other three benchmarks.

6.3. Practical application of the proposed approach

The in-depth numerical simulations demonstrate that the proposed approach exhibits a more accurate and robust performance for probabilistic WPF than the benchmarks of the states. More accurate WPF approach is able to effectively reduce the uncertainty about the future availability of wind power, and thus play a significant role in the planning and operation of electric power and energy systems. For example, in [44], probabilistic WPF approach was used to develop a novel unit commitment model which could be applied to analyze different commitment strategies and thus to optimize the operation of electric power and energy systems. In [45], a probabilistic nonlinear interval optimization model for power system dispatch was proposed. In this model, wind power quantiles obtained from probabilistic WPF method were utilized to evaluate the effect of wind power uncertainty on power flow distribution, considering both the average and deviation of the objective functions. The feasibility of this model was demonstrated on a modified IEEE 30-bus system. Also for example, probabilistic WPF results can be employed to facilitate energy scheduling of system operators in electricity markets [46]. Three steps were required in this process. Firstly, wind power point forecasting results were applied to determine the day-ahead unit commitment and energy dispatch. Then, the commitment statuses of fast-starting units were real-time adjusted based on the probabilistic information of wind power. These two steps constitute a two-settlement electricity market model with clearing of day-ahead and real-time markets for energy and operating reserves. Lastly, the real-time energy dispatch was achieved based on the realized availability of wind power. The practicability of this energy scheduling strategy incorporating probabilistic WPF approach was demonstrated on power system in Illinois.

In addition, it has been demonstrated by the authors that wind power probabilistic information can also be implemented in economic dispatch to quantify the impacts of wind power uncertainties on total generation cost [31]. More accurate wind power forecasting results mean less uncertainty on the quantiles and thus lead to decrease the operation cost of wind-embedded electric power and energy systems. Therefore, from these presented examples, it can be seen that probabilistic wind power approach with high-accuracy is very appealing for practical applications.

7. Conclusions

To maximize the beneficial impact of wind energy on climate change mitigation and environmental pollution reduction, probabilistic wind power forecasting with high-accuracy is a pressing need. Aimed at the target, a novel hybrid approach based on wavelet transform, deep convolutional neural network and ensemble technique was originally proposed in this paper for probabilistic WPF. The developed hybrid approach was comprehensively compared with the benchmark persistence method and shallow NN models, such as BP and SVM. The wind power dataset used for WPF was collected from actual wind farms in China. The test results obtained from various seasons, time resolutions and prediction horizons demonstrate that the proposed approach outperforms all of the tested alternatives in terms of reliability, sharpness and overall skill, i.e., CRPS. These comparatively better performances in all of the cases examined herein further illustrate that the forecasted distribution from the proposed approach reflects the real wind power uncertainty. Therefore, the proposed approach exhibits high-stability and strong robustness and is superior to all of the alternatives with which it was compared. The superiority of the proposed probabilistic WPF approach is attributed to the deep CNN architecture and the ensemble technique. The former can effectively extract the nonlinear and stochastic nature exhibited in each wind power frequency, and the latter can cancel the diverse errors out. It is also evident that the proposed proba-

bilistic WPF approach demonstrates a high potential for practical applications in electrical power and energy systems.

Acknowledgement

This work was jointly supported by Natural Science Foundation of China (Nos. 51477104 and 51507103), Natural Science Foundation of Guangdong Province (Nos. 2015A030310316 and 2016A030313041), the Foundations of Shenzhen Science and Technology Committee (Nos. JCYJ20150525092941041 and JCYJ20160422165525693), Shenzhen International Cooperation Research Project (GJHZ20150313093836007), Shenzhen University Research and Development Startup Fund (Nos. 2016035 and 2015030), and National Basic Research Program (973 Program) (No. 2013CB228202).

References

[1] Zhao YN, Ye L, Li Z, Song XR, Lang YS, Su J. A novel bidirectional mechanism based on time series model for wind power forecasting. Appl Energy 2016;177:793–803.
[2] Jong P, Kiperstok A, Sanchez AS, Dargaville R, Torres EA. Integrating large scale wind power into the electricity grid in the Northeast of Brazil. Energy 2016;100:401–15.
[3] Sun DM, Xu Y, Chen HJ, Wu K, Liu KK, Yu Y. A mean flow acoustic engine capable of wind energy harvesting. Energy Convers Manage 2012;63:101–5.
[4] Yu Y, Sun DM, Wu K, Xu Y, Chen HJ, Zhang XJ, et al. CFD study on mean flow engine for wind power exploitation. Energy Convers Manage 2011;52(6):2355–9.
[5] The Global Wind Energy Council. Global wind energy outlook 2008. The Global Wind Energy Council, Belgium; Oct. 2008. Available: <http://www.gwec.net/fileadmin/documents/Publications/GWEO_2008_final.pdf> [accessed: Feb. 11, 2009].
[6] Tastu J, Pinson P, Trombe PJ, Madsen H. Probabilistic forecasts of wind power generation accounting for geographically dispersed information. IEEE Trans Smart Grid 2014;5(1):480–9.
[7] Zhao J, Guo ZH, Su ZY, Zhao ZY, Xiao X, Liu F. An improved multi-step forecasting model based on WRF ensembles and creative fuzzy systems for wind speed. Appl Energy 2016;162:808–26.
[8] Haque AU, Hashem NM, Mandal P. A hybrid intelligent model for deterministic and quantile regression approach for probabilistic wind power forecasting. IEEE Trans Power Syst 2014;29(4):1663–72.
[9] Ziel F, Croonenbroeck C, Ambach D. Forecasting wind power – modeling periodic and non-linear effects under conditional heteroscedasticity. Appl Energy 2016;177:285–97.
[10] Hu J, Wang J. Short-term wind speed prediction using empirical wavelet transform and Gaussian process regression. Energy 2015;93:1456–66.
[11] Liu H, Tian HQ, Pan DF, Li YF. Forecasting models for wind speed using wavelet, wavelet packet, time series and artificial neural networks. Appl Energy 2013;107:191–208.
[12] Li G, Shi J. On comparing three artificial neural networks for wind speed forecasting. Appl Energy 2010;87(7):2313–20.
[13] Liu H, Tian HQ, Pan DF, Li YF. Wind speed forecasting approach using secondary decomposition algorithm and Elman neural networks. Appl Energy 2015;157:183–94.
[14] Mohammadi K, Shamshirband S, Yee PL, Petkovic D, Zamani M, Ch S. Predicting the wind power density based upon extreme learning machine. Energy 2015;86:232–9.
[15] Higgins P, Foley AM, Douglas R, Li K. Impact of offshore wind power forecast error in a carbon constraint electricity market. Energy 2014;76:187–97.
[16] Meng A, Ge J, Yin H, Chen S. Wind speed forecasting based on wavelet packet decomposition and artificial neural networks trained by crisscross optimization algorithm. Energy Convers Manage 2016;114:75–88.
[17] Cristobal GC, Ricardo B, Laura C, Oscar LG. On-line quantile regression in the RKHS (Reproducing Kernel Hilbert Space) for operational probabilistic forecasting of wind power. Energy 2016;113:355–65.
[18] Yan J, Li K, Bai E, Yang Z, Foley A. Time series wind power forecasting based on variant Gaussian Process and TLBO. Neurocomputing 2016;189:135–44.
[19] Zhang Y, Wang J. K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting. Int J Forecast 2016;32(3):1074–80.
[20] Wang Y, Wang JZ, Wei X. A hybrid wind speed forecasting model based on phase space reconstruction theory and Markov model: a case study of wind farms in northwest China. Energy 2015;91:556–72.
[21] Giebel G. The state of the art in short term prediction of wind power – a literature overview. Deliverable 1.2b of the ANEMOS. Plus project; 2011. p. 4.
[22] Li S, Wang P, Goel L. Wind power forecasting using neural network ensembles with feature selection. IEEE Trans Sustain Energy 2015;6(4):1447–56.
[23] Alessandrini S, Sperati S, Pinson P. A comparison between the ECMWF and COSMO ensemble prediction systems applied to short-term wind power forecasting on real data. Appl Energy 2013;107:271–80.
[24] Zhang G, Wu Y, Wong KP, Xu Z, Dong ZY, Herbert HC. An advanced approach for construction of optimal wind power prediction intervals. IEEE Trans Power Syst 2015;30(5):2706–15.
[25] Lee D, Baldick R. Short-term wind power ensemble prediction based on Gaussian processes and neural networks. IEEE Trans Smart Grid 2014;5(1):501–10.
[26] Hu QH, Zhang RJ, Zhou YC. Transfer learning for short-term wind speed prediction with deep neural networks. Renew Energy 2016;85:83–95.
[27] Lv Y, Duan Y, Kang WW, Li ZX, Wang FY. Traffic flow prediction with big data: a deep learning approach. IEEE Trans Intell Transport Syst 2015;16(2):865–73.
[28] Chen Y, Lin Z, Zhao X, Wang G, Gu YF. Deep learning-based classification of hyperspectral data. IEEE J Selected Top Appl Earth Observ Remote Sens 2014;7(6):2094–107.
[29] Rashwan MAA, Al Sallab AA, Raafat HM, Rafea A. Deep learning framework with confused sub-set resolution architecture for automatic Arabic diacritization. IEEE Trans Audio Speech Language Proc 2015;23(3):505–16.
[30] Hinton G. Deep belief networks. Scholarpedia 2009;4(5):786–804.
[31] Wang HZ, Wang GB, Li GQ, Peng JC, Liu YT. Deep belief network based deterministic and probabilistic wind speed forecasting approach. Appl Energy 2016;182:80–93.
[32] Zhang CY, Chen CLP, Gan M, Chen L. Predictive deep Boltzmann machine for multiperiod wind speed forecasting. IEEE Trans Sustain Energy 2015;6(4):1416–25.
[33] Ding J, Liu H, Huang M. Convolutional neural network with data augmentation for SAR target recognition. IEEE Geosci Remote Sens Lett 2016;13(3):364–8.
[34] Liang YD, Wang J, Zhou S, Gong Y, Zheng N. Incorporating image priors with deep convolutional neural networks for image super-resolution. Neurocomputing 2016;194:340–7.
[35] Abdel HO, Mohamed AR, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. IEEE Trans Audio Speech Language Proc 2014;22(10):1533–45.
[36] Deo RC, Wen XH, Qi F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl Energy 2016;168:568–93.
[37] Tascikaraoglu A, Sanandaji MB, Poolla K, Varaiya P. Exploiting sparsity of interconnections in spatio-temporal wind speed forecasting using wavelet transform. Appl Energy 2016;165:735–47.
[38] Catalão JPS, Pousinho HMI, Mendes VMF. Hybrid wavelet-PSO-ANFIS approach for short-term wind power forecasting in Portugal. IEEE Trans Sustain Energy 2011;2(1):50–9.
[39] Bouvrie J. Notes on convolution neural networks. Available on-line <http://cogprints.org/5869/1/cnn_tutorial.pdf>.
[40] Wan C, Xu Z, Dong ZY, Wong KP. Probabilistic forecasting of wind power generation using extreme learning machine. IEEE Trans Power Syst 2014;29(3):1033–44.
[41] Muthén B. Moments of the censored and truncated bivariate normal distribution. Br J Math Stat Psychol 1990;43(1):131–43.
[42] Pinson P, Reikard G, Bidlot JR. Probabilistic forecasting of the wave energy flux. Appl Energy 2012;93:364–70.
[43] Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J Roy Stat Soc 2007;69(2):243–68.
[44] Botterud A, Zhou Z, Wang J, Sumaili J, Keko H, Mendes J, et al. Demand dispatch and probabilistic wind power forecasting in unit commitment and economic dispatch: a case study of Illinois. IEEE Trans Sustain Energy 2013;4(1):250–61.
[45] Li YZ, Wu QH, Jiang L, Yang JB, Xu DL. Optimal power system dispatch with wind power integrated using nonlinear interval optimization and evidential reasoning approach. IEEE Trans Power Syst 2016;31(3):2246–54.
[46] Zhou Z, Botterud A, Wang J, Bessa RJ, Keko H, Sumaili J. Application of probabilistic wind power forecasting in electricity markets. Wind Energy 2013;16(3):321–38.