Hybrid Deep Learning-Based Model for Wind Speed Forecasting Based on DWPT and Bidirectional LSTM Network

AMIRHOSSEIN DOLATABADI\textsuperscript{1}, (Graduate Student Member, IEEE),
HUSSEIN ABDELTAWAB\textsuperscript{2}, (Member, IEEE),
AND YASSER ABDEL-RADY I. MOHAMED\textsuperscript{1}, (Fellow, IEEE)

\textsuperscript{1}Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2V4, Canada
\textsuperscript{2}School of Engineering, Penn State Behrend, Erie, PA 16563, USA
Corresponding author: Amirhossein Dolatabadi (adolatab@ualberta.ca)

This work was supported in part by the Future Energy Systems Research through the Canada First Research Excellence Fund (CFREF), and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

\section*{ABSTRACT} Accurate wind speed forecasting is essential for the reliability and security of the power system, and optimal operation and management of wind integrated smart grids. However, it is still a challenging task due to the highly uncertain and volatile nature of wind speed. Accordingly, in this work, a novel deep learning-based model integrating the discrete wavelet packet transform (DWPT) and bidirectional long short-term memory (BLSTM) is developed to precisely capture deep temporal features and learn the time-varying relationship of wind speed time series. In the proposed method, by applying the DWPT, both approximations and details parts are decomposed by passing through the filters to choose the frequency band related to the features of the original signal more adaptively. The BLSTM networks are incorporated to deal with the uncertainties more effectively as they have bidirectional memory capability (feedforward and feedback loops) to investigate both previous and future hidden layers data. To simultaneously improve the forecasting performance and decrease the learning complexity, the reconstructed state space of historical wind data is employed to reflect the evolution laws of wind speed. Two case studies using real-world wind speed datasets gathered from Flatirons campus (M2) of National Renewable Energy Laboratory (NREL) located in Colorado, USA and weather station of Edmonton, Canada are implemented to demonstrate the effectiveness and superiority of the proposed hybrid method compared to the shallow architectures and state-of-the-art deep learning models in the recent literature.

\section*{INDEX TERMS} Bidirectional LSTM, deep learning, time series, Takens’ embedding theorem, wavelet packet decomposition, wind speed prediction.

\section*{I. INTRODUCTION} In recent decades, the research and development of renewable energies have gradually increased around the world as an appealing solution to the high greenhouse gases’ emissions of fossil fuel-based energy resources, which raised worldwide concerns [1]. Due to the cleanliness and abundance, wind energy has attracted extensive attention compared to others in the realm of renewable energy sources. The total installed capacity of wind power in Canada has increased from 2,349 MW in 2008 to 12,816 MW in 2018 by an annual rate of 20\% in the past ten years [2]. Nonetheless, the most significant challenge for the large-scale penetration of wind energy in the power and energy systems is its uncertain and intermittent nature [3]. The wind power generation mainly depends on the wind speed, which can dramatically fluctuate in few seconds and directly affect the stability, resilience, and robustness of the power system. For this reason, accurate wind speed prediction facilitates wind power facilities integration into modern power systems.

Over the past few years, different wind time series prediction models were developed in the literature [4]–[7]. Based on the forecast time horizons, these models can be mainly classified into three categories [4]: 1) Short-term wind
forecasting refers to the prediction of wind data in a period starting from several minutes to hours ahead. Economic dispatch, grid regulations, and real-time electricity market clearing are dependent on this type [5]. 2) Medium-term prediction is mainly for time horizons in the range of several hours to a week. This type of prediction benefits reserve markets and unit commitment [6]. 3) Long-term prediction is for a period starting from one week to years ahead. Long-term studies of wind power plants, such as maintenance issues or expansion planning, utilize this forecasting type [7].

Generally, wind speed and wind power prediction methodologies are divided into two groups: physical and statistical methods [8]. The physical forecasting methods use the boundary conditions and physical parameters such as ambient temperature, atmospheric pressure, obstacles, and surface roughness [9]. These models often have excellent forecasting performance in the long-term period and large-scale areas due to the high computational burden. The computational fluid dynamics (CFD) and numerical weather prediction (NWP) are the most critical technologies in the physical models. The research work presented in [10] proposes a boundary layer scaling (BLS) technique based on the NWP model for long-term wind speed forecasting. The statistical forecasting models mainly use the wind time series data and try to find the mathematical relationships between the spatial-temporal samples or historical data, which yields accurate estimation results in the short-term prediction tasks. The Autoregressive (AR) model, autoregressive moving average (ARMA) model, and autoregressive integrated moving average (ARIMA) model are the most popular linear statistical approaches. In [11], an ARIMA model is introduced to represent the upper and lower bounds of the wind power generation. However, the linear nature of these methods limits the ability of them to deal with challenging wind data prediction problems and handle the nonlinear patterns. Recently, by developing artificial intelligence (AI) algorithms, different new methods for wind data forecasting have been quickly proposed. Compared with the linear statistical prediction methods, these models could provide a complicated nonlinear relationship for prediction tasks. Artificial neural network (ANN) is a promising AI tool for accurate time series forecasting [12]. Most ANN architectures introduced in the literature have only one single hidden layer and are shallow. For example, three types of typical shallow neural networks (SNNs), including adaptive linear element (ALE), backpropagation neural network (BPNN), and radial basis neural network (RBFNN) for one-hour-ahead wind speed prediction are employed and evaluated in [13]. It was observed that the performance is highly dependent on the hyperparameters of the networks, and none of the models can outperform others in terms of all criteria.

To compensate for the shortcomings of the above forward structures, recurrent neural networks (RNNs) have been proposed [8]. Unlike feedforward NNs, RNNs acquire the predicted value from the current inputs and the experience that leads to the better capturing of various patterns and temporal sequences. For example, to predict wind power time series, the authors of [8] use the two RNN-based model: the nonlinear autoregressive with exogenous inputs (NARX) and Elman. Reference [14] presented a non-parametric probabilistic wind power forecasting method based on the empirical dynamic modeling (EDM) and ‘Takens’ theorem. Difficulties of finding optimal structure in the search space of ANNs with several layers might be the main reason for using SNNs. However, SNNs are not capable of efficiently learning sophisticated features from the wind data; thus, they have the drawbacks of getting into local minimum and over-fitting. Recently, with the booming advancement of machine learning techniques as well as the development of graphics processing units (GPUs), neural network-based deep-learning methodologies as the new branch of ANNs have been gradually developed [15]. In contrast to SNNs, such models can effectively extract inherent abstract features of the highly varying time series data. In [16], a novel deep belief network (DBN) model is employed for both deterministic and probabilistic wind speed forecasting. A combination of secondary decomposition (SD) and bidirectional gated recurrent unit (BiGRU) presented in [17]. The presented model in [5] extracts unsupervised temporal features from wind speed data by restricted Boltzmann machines (RBM) and rough set theory. The long short-term memory (LSTM) architecture, which is a particular type of RNN with rich dynamics, was initially proposed by Hochreiter and Schmidhuber [18]. These networks overcome the vanishing gradient problem, causing losses of valuable information by introducing a gating mechanism and memory cells into RNNs [19]. Reference [20] evaluated the effect of performing deep-stacked LSTM and BLSTM in electricity load forecasting. Compared with abundant SNN forecasting researches, there are only a few studies related to deep learning-based forecasting for wind speed [21], [22]. In [23], the authors proposed a two-layer structure based on extreme learning machine (ELM), Elman NN and LSTM network to predict wind speed in 10-min and 1-hour ahead time intervals. As another example, in [24] a short-term wind speed forecasting model based on the combination of clustering and bidirectional LSTM (BLSTM) is proposed. All of these studies illustrated the effect of using deep networks on increasing the prediction accuracy.

Besides, the signal processing techniques also can considerably enhance the accuracy of the forecasting models by data transformation, data de-noising, and data feature extraction. For example, a short-term wind speed prediction model based on wavelet packet transform (WPT) and SNN was mooted in [25]. In very recent literature, [26] combines improved empirical wavelet transform (IEWT) and least square support vector machine (LSSVM) to forecast short-term wind speed. Although different signal processing methods have been widely used as a preprocessing approach in wind prediction models, minimal studies are made on integrating WPT and deep neural networks.
According to the models above, the study of using deep learning methods in the wind speed forecasting area is still not enough that improvement may be achieved from both the preprocessing and model aspects. To bridge the gap in wind data time series modeling, this work seeks to address two important issues for wind farm operators and energy market participants: 1) advantages and disadvantages of bidirectional deep learning models over other deep and shallow frameworks in wind data prediction; 2) impacts of using WPT as a special type of wavelet transform, in which BLSTM is investigated. Table 1 summarizes a novel short-term wind speed prediction model based on typical wavelet transform and WPT. To this end, in this study, a novel short-term wind speed prediction model based on the WPT, and BLSTM is investigated. Table 1 summarizes a taxonomy of recently proposed methodologies in forecasting area. We employ the WPT, which can effectively extract the meaningful components of the raw wind speed signal. Considering the chaotic and stochastic characteristics of wind speed time series, the BLSTM is introduced to explore the high-level nonlinear and non-stationary features of the wind speed time series. Besides, instead of the general correlation method, delay embedding theorem is applied to handle the chaotic historical wind speed data. To the best of our knowledge, this is the first study that introduces WPT-BLSTM deep learning model in wind speed time series prediction. Eventually, two real-world wind speed datasets have been used to evaluate the performance of the proposed prediction methodology and comparison with some benchmark models are provided in simulations. The normalized embedding delay $\lambda$ comes down to defining two parameters: normalized embedding dimension $d$ and normalized embedding delay $\lambda$, which determines each delay vector's optimal autocorrelation value, and embedding dimension $d$, which means the size of the set of most recent observations.

The rest of the paper is structured as follows. Section II motivates and explains the delay embedding methodology, WPT approach, and the BLSTM network. It describes how they are employed to develop the proposed wind prediction model. Section III conducts two 1-hour and 10-minute real-world wind speed datasets to demonstrate the efficiency and applicability of the proposed algorithm. Section IV draws the conclusion and outlines interesting main directions for future works.

II. METHODOLOGY

In this section, first, the notation and concept of Takens' embedding theorem and are explained which are used in this paper as an effective tool to form the reconstructed state space. Then, the proposed DWPT-BLSTM framework is introduced to capture deep temporal features from the reconstructed wind speed time series.

A. DELAY EMBEDDING AND DYNAMIC RECONSTRUCTION THEORY

In a chaotic system, the initially unobservable dynamics of interest can be reconstructed by employing Takens’ embedding and dynamic reconstruction theories [14]. According to the Takens’ theorem, a new state space can be constructed such that its evolution of observations is equivalent to that of the original one. Building a delay embedding comes down to defining two parameters: normalized embedding delay $\lambda$, which determines each delay vector’s optimal autocorrelation value, and embedding dimension $d$, which means the size of the set of most recent observations.

In a dynamical system and discrete-time environment, the observable output $y_t$ is described as follows:

$$y_t = f(x_t)$$

where $f(\cdot)$ and $x_t$ represent the dynamics of the system and nonlinear scalar-valued function, respectively. Based on the delay embedding theorem, reconstructed dynamics with embedding dimension $d$ and normalized embedding delay $\lambda$, $\{y_t^{rec}\}$, can be formulated as follows:

$$y_t^{rec} = [y_t, y_{t-\lambda}, \ldots, y_{t-(d-1)\lambda}]$$

The normalized embedding delay $\lambda$ is determined heuristically based on the average mutual information (AMI) method [27]. By this method, the first minimum of the mutual

| Reference | Research field | Decomposition | Input vector selection | Network |
|-----------|---------------|---------------|-----------------------|---------|
| [5]       | Wind          | ×             | Mutual information    | RBM     |
| [14]      | Wind          | ×             | Takens’ embedding theorem | EDM     |
| [17]      | Wind          | SD            | Trial-and-error       | BiGRU   |
| [20]      | Load          | ×             | Trial-and-error       | BLSTM   |
| [23]      | Wind          | ×             | Trial-and-error       | ELM+LSTM|
| [24]      | Wind, load, price | ×           | Trial-and-error       | BLSTM   |
| Proposed  | Wind          | DWPT          | Takens’ embedding theorem | BLSTM   |
information between \( y_t \) and \( y_{t-\lambda} \) is the optimal value of \( \lambda \). Besides, the false nearest neighbors (FNN) technique is applied to find the proper value of embedding dimension \( d \). Furthermore, the first minimum of the FNN determines the acceptable minimum value of \( d \) under changes in the embedding dimension from \( d \to d + 1 \) which satisfies the sufficient condition \( d \geq 2D + 1 \) (\( D \) represents the state space dimension of the unknown dynamics) [28]. Based on the discussion as mentioned above, both evolutions \( y_{t+1}^{rec} \to y_{t+1}^{rec} \) and \( x_t \to x_{t+1} \) are similar.

Therefore, to handle the forecasting problem of time series \( \{x_t\} \), it is better to forecast the time series \( \{y_t^{rec}\} \). The following mapping can represent this:

\[
y_{t+1}^{predicted} = F \left( y_t^{rec} \right) \tag{3}
\]

where \( y_{t+1}^{predicted} \) is the forecasted value of the time series \( \{y_t\} \) for the next time-step. It is worth noting that with different \( F \), equation (5) can be extended to a multi-step prediction form.

\[
y_{t+\tau}^{predicted} = F' \left( y_{t}^{rec} \right) \tag{4}
\]

Since the wind data time series shows a chaotic behavior from a dynamic system point of view, the reconstructed state-space model is employed to transform it into a suitable form of machine learning methods.

**B. WIND SPEED DECOMPOSITION**

1) **WAVELET TRANSFORM**

Wavelet transform (WT) represents an excellent tool to capture the wind speed dynamics and temporal patterns since wind speed has a time-varying nature and spreading frequency spectrum. By using WT, an initial wind data signal is decomposed into a set of wavelets, which in turn represent a better behavior than the original wind data series.

Compared to other signal decomposition methods, wavelet analysis can better reveal temporal features of the wind speed sequential data such as discontinuities in higher derivatives, breakdown points, self-similarity, and trends [29]. Additionally, signal de-noising or compressing without any remarkable degradation are the other essential features of the WT. The WT is categorized into two groups: continuous wavelet transform (CWT) and discrete wavelet transform (DWT) [27]. A CWT of the signal \( f(t) \) is described as follows [34]:

\[
CWT_f (\alpha, \beta) = \left\langle f(t) , \psi_{\alpha,\beta} (t) \right\rangle = \int_{-\infty}^{+\infty} f(t) \psi^*_{\alpha,\beta} (t) \, dt \tag{5}
\]

\[
\psi_{\alpha,\beta} (t) = \frac{1}{\sqrt{\alpha}} \psi \left( \frac{t - \beta}{\alpha} \right) \tag{6}
\]

where \( \psi \) (\( t \)) and \( \Psi \) denote the mother wavelet and set of wavelets, respectively. \( \alpha \) as a scaling coefficient determines the spread of the wavelet and \( \beta \) as a translation coefficient, controls the central position. Compared to the Fourier transform (FT) which represents the signal as a combination of sines and cosines, by using the CWT, a set of wavelets is generated associated to a mother wavelet, \( \psi \), and predefined values of the scale and translation coefficients respect to the original non-stationary signal [30]. However, the CWT method is not easily applicable to the desired tasks due to substantial redundant information and a very high computational burden. According to (5) and (6), CWT is continuously achieved by continuously scaling and translating the mother wavelet and shifting it over the signal to obtain the correlation between them. Furthermore, there is not any analytical solution for most cases, which leads to numerical calculation methods and, consequently, higher computational complexity. DWT, as a digital counterpart of CWT is introduced to address these issues. Therefore, instead of following the proposed procedure, the signal is analyzed at different resolutions with various frequency bands. This type of WT applies a binary system to subsample the CWT, decreasing the redundancy information while retaining the principal characteristics. It dramatically improves efficiency and keeps accuracy just as same as the CWT [31]. The DWT is expressed as (7), where \( \nu \) and \( k \) denote integers. \( \beta_0 \) and \( \alpha_0 \) are fixed dilation step and the translation factor, respectively. There are two different sets of functions in DWT, wavelet, and scaling functions, which are related to high-pass and low-pass filters, respectively, as presented in (8) and (9).

\[
\psi_{\nu,k} (t) = \frac{1}{\sqrt{\alpha_0}} \psi \left( \frac{t - k \beta_0 \alpha_0^\nu}{\alpha_0} \right) \tag{7}
\]

\[
\psi (2^\nu t) = \sum_{k=1}^{K} g_{\nu+1} (k) \psi \left( 2^{\nu+1} t - k \right) \tag{8}
\]

\[
\psi (2^\nu t) = \sum_{k=1}^{K} h_{\nu+1} (k) \psi \left( 2^{\nu+1} t - k \right) \tag{9}
\]

where \( g(k) \) and \( h(k) \) denote the wavelet and scaling filters, respectively. \( \psi \) and \( \varphi \) are the wavelet and scaling functions, respectively. Subsequently, a signal \( f(t) \) is written as follows:

\[
f(t) = \sum_{k=1}^{K} \xi_{\nu-1} (k) \psi \left( 2^{\nu-1} t - k \right) \tag{10}
\]

\[
+ \sum_{k=1}^{K} \zeta_{\nu-1} (k) \varphi \left( 2^{\nu-1} t - k \right)
\]

\[
\xi_{\nu-1} (k) \quad \text{and} \quad \zeta_{\nu-1} (k) \text{ are the coefficients calculated using the inner products of wavelet and scaling functions with the signal as follows:}
\]

\[
\xi_{\nu-1} (k) = \left\langle f(t), \psi_{\nu,k} (t) \right\rangle \tag{11}
\]

\[
\zeta_{\nu-1} (k) = \left\langle f(t), \varphi_{\nu,k} (t) \right\rangle \tag{12}
\]

Based on the multiresolution approach developed by Mallat, a signal can be broken down into “approximations,” associated with the general trend of the analyzed signal (A) and “details” related to the high-frequency parts (D) [32]. Then, Approximations are consecutively decomposed into lower resolution components to obtain a multilevel decomposition process. Fig. 1 displays the analytical approach of the DWT method.
This problem limits the capability of the network to infinity/zero over time due to the multiple gradient calculations that training back-propagated errors in the steepest descent problem. Gradient exploding/vanishing refer to the situations that training back-propagated errors in the steepest descent problem have weak ability of learning and addressing long-range dependencies. Furthermore, RNNs, which mainly consist of sequence-based architectures to find the temporal correlations between past circumstances and the current information, have been widely applied in processing sequentially dependent data domain. RNNs try to establish a prediction framework for finding dependencies between inputs and outputs based on only a self-learning process instead of using a mathematical model. Although RNNs have recently obtained more accurate forecasting results compared to conventional feedforward networks, these models have two main drawbacks: 1) RNNs have weak ability of learning and addressing long-range dependencies due to the exploding or vanishing of gradient problem. Gradient exploding/vanishing refer to the situations that training back-propagated errors in the steepest descent algorithm, increase/decrease exponentially fast to infinity/zero over time due to the multiple gradient calculations. This problem limits the capability of the network to learn temporal correlations when the time horizon is extended reliably. 2) RNNs do not consider the future context’s information, which leads to requiring more backward relations modeling [34].

LSTM architecture is employed here as an alternative network to tackle them efficiently. The first issue is that LSTM networks are more complex and improved RNNs with an internal state capable of propagating data through multiple time steps and temporal processing characteristics of time series data. Let \( l \in [1, L] \) be the layer of LSTM, where contains cyclically connected special blocks known as memory blocks. Fig. 3 illustrates the general structure of the LSTM block architecture, where each block has one or more memory cells, and three multiplicative units called an input, an output, and forget gates, representing operators for respectively continuous writing, reading and resetting of data in the cell. Also, Fig. 4 shows the information connection procedure during the subsequent time steps in an un-rolled LSTM network. The past state or the explanatory variables are the candidates of the new information memorized by the input gate. The output gate controls the impact of memory content on the node output, whereas the forget gate can discard irrelevant information. Succinctly, the forward pass associated with the LSTM architecture is formulated as follows [19]:

\[
\begin{align*}
    f_t &= \sigma \left( W_f^f x_t + W_f^h h_{t-1} + b^f \right) \quad (13) \\
    i_t &= \sigma \left( W_i^i x_t + W_i^h h_{t-1} + b^i \right) \quad (14) \\
    g_t &= \tanh \left( W_g^g x_t + W_g^h h_{t-1} + b^g \right) \quad (15) \\
    o_t &= \sigma \left( W_o^o x_t + W_o^h h_{t-1} + b^o \right) \quad (16) \\
    C_t &= g_t \odot i_t + C_{t-1} \odot f_t \quad (17) \\
    h_t &= \tanh \left( C_t \right) \odot o_t \quad (18)
\end{align*}
\]

where \( \sigma \) and \( \tanh \) represent the logistic sigmoid and the hyperbolic tangent activation functions, respectively, whereas \( f, i, o \) are the activation vectors related to the forget, input, and output gates, respectively. The weight matrices \( W \) and bias vector \( b \) to be optimized during the training procedure.

To solve the second problem of RNNs, the bidirectional concept is incorporated into the proposed LSTM model to capture the whole temporal horizon’s information. Based on Fig. 5, a structure including two different recurrent networks with the same output is capable of both forward and backward training process [35]. Proposed topology has been widely applied in the speech recognition domain due to its’ capability to efficiently recognize a word by using not only the previous words but also the whole sentence [36]. Motivated by the proposed principle, here, the necessary information is completely exploited by the explanatory variables during each time step. This process leads to having a better prediction performance. Moreover, besides better training time, an advantage of the proposed bidirectional networks over unidirectional RNNs is robustness to the biased inputs and model uncertainties [34].

Bidirectional LSTM (BLSTM) models as a combination of LSTM networks and bidirectional RNNs, can simultaneously
memorize long-term dependencies and process the information bidirectionally. More specifically, when deep structures are built, one can achieve much higher data representation capability compared to traditional RNNs or LSTMs.

III. REALISTIC WIND SPEED FORECASTING CASE STUDY DEFINITION

In this section, the details of real-world wind speed datasets, parameter settings and the well-known error criteria for evaluating the forecasting method are introduced.

A. DATASETS

The historical wind speed time series used in this work are measured from two different sites: 1) the Flatirons campus (M2) wind site of the National Renewable Energy Laboratory (NREL) located in Colorado, USA obtainable from the NREL National Wind Technology Center (NWTC) website [37]. The data were obtained by applying a next-generation mesoscale NWP system called Weather Research and Forecasting (WRF) developed for operational forecasting needs and atmospheric research tasks. 2) Edmonton, Canada historical wind speed data [38]. The historical weather data are courtesy of Environment and Climate Change Canada [39] and combined from multiple Environment and Climate Change Canada data sources to be accurate.

The chosen NREL datasets include wind speed data in 1-hour and 10-minute intervals from 1 January 2017 to 31 December 2018 and 1 January 2000 to 31 March 2000, respectively. The time period of Edmonton dataset is from 1 January 2017 to 31 December 2018 with the hour unit. Table 2 shows the impacts of choosing bigger datasets on the forecasting accuracy and computational time. As can be seen from this Table, using bigger training datasets increases the time up to double while having negligible impact on the learning capability. Whole datasets are further split into three subsets for different purposes, namely training set...
TABLE 2. Forecasting results of the proposed model (DWPT+BLSTM) with different training data size.

| Dataset   | Range (%)       | MAPE (%) | Training time |
|-----------|-----------------|----------|---------------|
| NREL 1-hour | Jan 2017 to 31 Dec 2018 | 6.16     | 00:09:11      |
| NREL 1-hour | Jan 2016 to 31 Dec 2018 | 6.04     | 00:16:48      |
| NREL 10-min | Jan 2000 to 31 Mar 2000 | 3.29     | 00:08:11      |
| NREL 10-min | Oct 1999 to 31 Mar 2000 | 3.24     | 00:17:05      |

for building up the model, validation set for an unbiased assessment during tuning the parameters of the model, and finally testing set for the last evaluation of the model built. In this study, the training and validation sets account for 70% and 20% of the dataset, respectively, and the remaining data are allocated for testing the forecasting performance of the proposed method. In other words, the data partitioning is 0.7/0.2/0.1.

B. PARAMETER DETAILS

It is worth noting that based on the state space reconstruction methodology, the input size of the network is determined by the dimension of delay vectors. From a dynamic systems point of view, the prediction task is considered the prediction of system states since a series of observations about the system is seen as a time series. In other words, the observations in the time series are a nonlinear projection of the system’s state variables onto the observation variables. To this end, a small set of the most recent previous observations is used as state variables to construct an equivalent version of the original state space [14]. In this way, two parameters of the space need to be calculated: embedding delay $\lambda$, which optimally determines the level of autocorrelation corresponding to each delay vector and embedding dimension $d$, which are mathematically equivalent to the size of proposed observations’ set. The utility functions of the TISEAN toolbox called $\text{mutual}$, and $\text{false_nearest}$ are employed in this paper to find the proper values of $\lambda$ and $d$, respectively [40]. Figs. 6 and 7 show the variations of the average mutual information and the percentage of false nearest neighbors, respectively. As shown in Fig. 6, the average mutual information between wind speed at times $t$ and $t - \lambda$ reaches its first local minimum at 18, which is chosen as the optimal value of $\lambda$. Moreover, as depicted in Fig. 7 and based on the first minimum of the false nearest neighbors percentage, seven can be selected as the minimum acceptable value of $d$. Note that this value is not in contrast to the seasonality of time series data, $\lambda$ means that the value time series at time $t$ and at a time $t - \lambda$ can participate in the reconstructed space as two consecutive members due to the essential independence. On the other hand, the independence level is not so much as to can say there is not any correlation between them. It is noteworthy that choosing two seasonal data points may lead to high redundant information, which is not desirable. Furthermore, wind speed data at time $t - 1$ to $t - 6$ (based on mutual information) are also added to the input vector to highlight the correlation of time series.

The proposed hybrid forecasting framework starts with determining the optimal values of embedding delay $\lambda$ and dimension $d$ parameters for the given datasets. To reduce the sensitivity of the BLSTM network to the data scale and accelerate the training procedure, the input vectors are normalized and scaled to the range of (0, 1) according to their nature. It is assumed that there is no false or missing data because of the normal performance of measuring instruments during this time period. Then, db4 three-level DWPT as a comprehensive signal decomposition method is adopted. As there is no global theory or clear method to determine hyper-parameters associated with the BLSTM network, and they are completely data-dependent. Therefore, a random search method is applied on set $\psi = \{10, 20, \ldots, 200\}$ to configure the proposed deep BLSTM network optimally. Although better-optimized configurations could be obtained by using heuristic algorithms or grid search method compared to the proposed random search, these search algorithms have a very high computational burden and remarkable running time [4]. The proposed BLSTM network is initialized with
100 units and 3 hidden layers, 50 epochs, and a batch size of 20. Besides, simulation of each model runs 30 times to alleviate the randomness influence and avoid a suboptimal solution. Adaptive Moment Estimation (Adam), as a computationally efficient optimizer, is applied to optimally calculate weights and biases of the network and minimize loss function [41]. Adam is an adaptive learning rate optimization method that shows slightly better performance in practical applications compared to other popular optimization algorithms such as RMSProp, stochastic gradient descent (SGD), Adadelta, and Adagrad. In this study, all the experiments are implemented in MATLAB 2019 software. The workstation used is configured with an Intel Core TM i7-8700 3.2 GHz CPU and 32 GB of RAM.

C. EVALUATION CRITERIA
The root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are employed as three evaluation metrics to evaluate the prediction results as follows:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (d(i) - y(i))^2} \tag{19}
\]

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |d(i) - y(i)| \tag{20}
\]

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{d(i) - y(i)}{d(i)} \right| \times 100\% \tag{21}
\]

where \(d(i), y(i)\) and \(N\) represents the desired output, the actual output, and the number of samples, respectively.

IV. EXPERIMENTAL RESULTS AND DISCUSSION
In this section, two practical case studies of wind speed prediction in Colorado, USA and Edmonton, Canada are carried out and the simulation results and comparison with benchmark methods are presented to validate the performance of the proposed model.

A. NREL M2 WIND SPEED DATASET
After tuning the parameters and specifying the optimal structure, the model is trained by using the training set. The average time required to train the network is entirely dependent on the structure complexity, and for the proposed model, it is about 10 minutes, which makes it applicable for real-time wind speed forecasting purposes. In the testing step, only the feed-forward process is used to find the output of the network; thus, it has negligible time complexity and is therefore so faster. Fig. 8 shows the forecasting results of the proposed model over the test data. In order to provide better visualization, the results for the last week are also shown in Fig. 9. As can be seen, the forecasted results (the orange line) follow the trend of the real measurements (the blue line) and are very close to them, which implies that the dynamic characteristics of the wind speed data are effectively captured by the model.

Tables 3 and 4 compare the results of the proposed hybrid method with some benchmark models which have been recently employed in the literature, including different combinations of DWT and DWPT with BPNN, RNN, DBN, LSTM, MLSTM [42], and BLSTM for 1-hour and 10-minute, respectively. The procedure of designing an optimal structure for the other models is similar to that of the proposed hybrid model. The optimal number of layers and hidden nodes are determined to be 3 and 30 for the BPNN and RNN, respectively. The DBN network has three hidden layers with 30 hidden nodes in each layer. The structures of the typical LSTM and MLSTM models are considered to be the same as the BLSTM network for providing better comparison. According to the results, the RNN network outperforms BPNN due to
FIGURE 10. Forecasting results for the last day of NREL 10-minute test data. (a) DWT + LSTM. (b) DWPT + LSTM. (c) DWT + MLSTM. (d) DWT + BLSTM. (e) DWPT + MLSTM. (f) DWPT + BLSTM.

the dynamic behavior. DBN has generally higher forecasting accuracy compared to shallow architectures, i.e., BPNN and RNN. It has 9.15% and 8.74% MAPE improvements for the RMSE and MAPE over RNN in the 1-hour case. LSTM-based models, LSTM, MLSTM, and BLSTM, show promising prediction results by achieving significantly lower error metrics compared to BPNN RNN, and DBN.

As shown in Tables 3 and 4, the proposed method achieves the lowest RMSE, MAPE and MAE values and procures the best prediction performance compared to the other benchmark models. It can also be observed that models with DWPT have better performance than those with DWT. For example, DWPT + BPNN improves RMSE and MAE by 31% and 29%, respectively, compared to DWT + BPNN. These improvements result from the fact that applying the DWPT transformation yields more detail and approximation coefficients, and consequently, better forecasting.

Moreover, using BLSTM leads to considerable improvement, especially in the 10-minute case, compared to LSTM. BLSTM provides advantages mainly by reducing the forecast errors while increasing the training time and complexity of the model. Note that the training time for the models with BLSTM is still much less than the prediction time scale of one hour. To better show how different models predict the
highly volatile wind speed time series, the forecasting results of different networks for the 10-minute and 1-hour datasets are visualized in Fig. 10 and Fig. 11, respectively. As shown, the performance of the proposed hybrid model is much better than other shallow or deep learning-based models, especially when wind speed time series has an abrupt change due to using more meaningful information and better generalization capability.

To verify the performance of the reconstructed state-space model, the 1-hour case study is repeated for different input structures. Table 5 shows the results of some typical input structures for different values of \( d \) and \( \lambda \). As shown in this table, both the increase and the decrease in the input vector parameters reduce the performance according to the forecasting indices.

On the other hand, Table 6 shows the MAE of different models for 1-hour up to 3-hour ahead wind speed prediction task. The proposed architecture, DWPT+BLSTM, procures remarkably better results than other deep and shallow architectures for larger prediction time steps. DWPT+BLSTM improves MAE by 6.7% and 3% for 2-hour ahead prediction compared to DWPT+DBN and DWPT+MLSTM, respectively. Such improvement increases to about 10.7% and 6.8% for 3-hour ahead prediction. This shows the deep feature extraction ability of the proposed deep network. It is worth noting that, although the training time increases as the time horizon is extended, it is still negligible compared to the forecasting time step.
B. EDMONTON WIND SPEED DATASET

In this study, historical hourly wind speed data from Edmonton, Canada is used as the second dataset to show the applicability and effectiveness of the proposed model in dealing with different locations which have various wind speed characteristics. To perform a fair comparison, the time
TABLE 7. Comparison of forecasting performance for the Edmonton dataset.

| Forecasting method | RMSE (m/s) | MAE (m/s) | MAPE (%) | Training time |
|--------------------|------------|-----------|----------|---------------|
| DWT+BPNN [12]     | 0.758      | 0.602     | 15.79    | 00:02:36      |
| DWT+BPNN           | 0.507      | 0.583     | 8.97     | 00:03:04      |
| DWT+RNN [8]        | 0.615      | 0.481     | 12.22    | 00:03:19      |
| DWT+RNN            | 0.439      | 0.338     | 8.27     | 00:04:22      |
| DWT+DBN [16]       | 0.582      | 0.449     | 11.61    | 00:03:27      |
| DWT+DBN            | 0.385      | 0.273     | 7.88     | 00:04:41      |
| DWT+LSTM [22]      | 0.577      | 0.437     | 10.19    | 00:03:05      |
| DWT+LSTM           | 0.296      | 0.231     | 6.51     | 00:04:11      |
| DWT+MLSTM          | 0.526      | 0.392     | 9.21     | 00:09:02      |
| DWT+MLSTM [42]     | 0.277      | 0.221     | 6.44     | 00:09:56      |
| DWT+BLSTM          | 0.490      | 0.371     | 8.64     | 00:06:41      |
| Proposed (DWT+BLSTM) | 0.253      | 0.197     | 4.82     | 00:07:24      |

Furthermore, forecasting performance of different models with their corresponding errors for the last week of the Edmonton test data are shown in Fig. 12. From this figure, it can be seen that the errors of DWT-based models range from -1 to 1 m/s while for those with DWT can even approach to 5 m/s. Both DWT+MLSTM and DWT+BLSTM models have high potential to forecast the overall behavior of wind speed time-series and they can follow the sharp spikes accurately. However, DWT+MLSTM needs around 22% more training time compared to DWT+BLSTM while DWT+BLSTM has 33.6% less MAPE. In summary, as the forecasting results illustrate, we can find that the proposed DWT+BLSTM framework can predict Edmonton hourly wind data effectively better than the benchmark methods.

C. MULTIVARIATE FORECASTING TASK

In this subsection, the ambient temperature at 80 m altitude is considered as an extra feature. Fig. 13 compares the prediction performance of the proposed model with DWT+MLSTM and DWT+LSTM. As we expected, it can be seen that the multivariate frameworks demonstrate slightly higher forecast accuracy than the models without exogenous input.

In the multivariate forecasting, DWT+BPNN and DWT+RNN produce the MAPE of 10.27% and 8.63%, respectively. For DWT+DBN and DWT+LSTM methods, MAPEs of multivariate tasks are 7.75% and 6.92%, respectively. Similar improvements are made for DWT+MLSTM and DWT+BLSTM models. Comparing DWT+MLSTM with DWT+BLSTM, for the multivariate prediction, DWT+BLSTM method can improve MAPE by 14.52%. DWT+BLSTM can predict the most accurate multivariate wind speed forecasting task, which can get the best forecasting metrics in the real data. This finding can verify the superiority of the proposed DWT+BLSTM framework.

V. CONCLUSION

Accurate knowledge of the variability and availability of wind speed is a very crucial issue for the operation and scheduling of the smart grid. In this work, a new hybrid deep learning-based approach is proposed for short-term wind speed prediction. First, the DWT is applied to effectively extract the features of the signal by decomposing the raw wind speed time series into several sub-layers. The input vector is built by using the theory of dynamic reconstruction, which not only increases the accuracy of the results but also decreases the learning complexity by determining the optimal structure of inputs. Moreover, the BLSTM network as a combination of LSTM networks and bidirectional RNNs is incorporate to capture deep temporal features with high abstractions. The proposed model is evaluated on a publicly available real-world dataset, of which the forecasting accuracy is comprehensively compared to multiple benchmarks that exist in the literature. The proposed BLSTM+DWT framework shows the smallest metrics and generally achieves the best forecasting performance in the dataset. For example, BLSTM+DWT has demonstrated 34% and 32% improvement in RMSE and MAE when compared with BLSTM+DWT.

As for future research, the proposed model can be improved by taking into account more inputs such as humidity and atmospheric pressure. Further feature extraction methodologies such as data clustering methods will be tested to improve the wind speed prediction accuracy. Another Future research direction will focus on using spatiotemporal data and offshore wind power prediction while considering the sea current level.
AMIRHOSSEIN DOLATABADI (Graduate Student Member, IEEE) was born in Tabriz, Iran, in 1991. He received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Tabriz, Tabriz, Iran, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree with the University of Alberta, Edmonton, AB, Canada.

From 2016 to 2018, he was a Research Assistant with the Smart Energy Systems Laboratory (SES Lab), University of Tabriz. His research interests include artificial intelligence, machine learning, and data analytics applications in energy systems.

HUSSIN ABDELTAWAB (Member, IEEE) was born in Beni Suef, Egypt, in April 1987. He received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from Cairo University, in 2009 and 2012, respectively, and the Ph.D. degree in electrical engineering from the University of Alberta, Edmonton, AB, Canada, in 2017.

He is currently an Assistant Professor of electrical engineering with Penn State University, Erie, PA, USA. His research interests include energy management, control system applications in renewable energy, energy storage, and smart distribution systems. He is also a licensed Professional Engineer in Saskatchewan, Canada.

YASSER ABDEL-RADY I. MOHAMED (Fellow, IEEE) was born in Cairo, Egypt, in November 1977. He received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from Ain Shams University, Cairo, in 2000 and 2004, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 2008.

He is currently with the Department of Electrical and Computer Engineering, University of Alberta, AB, Canada, as a Professor. His research interests include dynamics and controls of power converters; grid integration of distributed generation and renewable resources, microgrids, modeling, analysis, and control of smart grids; and electric machines and motor drives.

Dr. Mohamed is an Associate Editor of the IEEE TRANSACTIONS ON POWER ELECTRONICS and an Editor of the IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, and IEEE Power Engineering Letters. He is a registered Professional Engineer in the Province of Alberta.