Characterizing Time Series Data Diversity for Wind Forecasting

Preprint

Cong Feng,[1] Bri-Mathias Hodge,[2]
Erol Kevin Chartan,[2] and Jie Zhang[1]
[1] *University of Texas at Dallas*
[2] *National Renewable Energy Laboratory*

Presented at the IEEE/ACM International Conference on Big Data Computing, Applications, and Technologies
Austin, Texas
December 5–8, 2017

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Conference Paper
NREL/CP-5D00-71412
June 2018

Contract No. DE-AC36-08GO28308

NOTICE

The submitted manuscript has been offered by an employee of the Alliance for Sustainable Energy, LLC (Alliance), a contractor of the US Government under Contract No. DE-AC36-08GO28308. Accordingly, the US Government and Alliance retain a nonexclusive royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for US Government purposes.

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Characterizing Time Series Data Diversity for Wind Forecasting

Cong Feng
The University of Texas at Dallas
cong.feng1@utdallas.edu

Erol Kevin Chartan
National Renewable Energy Laboratory
ErolKevin.Chartan@nrel.gov

Bri-Mathias Hodge
National Renewable Energy Laboratory
Bri.Mathias.Hodge@nrel.gov

Jie Zhang
The University of Texas at Dallas
jiezhang@utdallas.edu

ABSTRACT

Wind forecasting plays an important role in integrating variable and uncertain wind power into the power grid. Various forecasting models have been developed to improve the forecasting accuracy. However, it is challenging to accurately compare the true forecasting performances from different methods and forecasters due to the lack of diversity in forecasting test datasets. This paper proposes a time series characteristic analysis approach to visualize and quantify wind time series diversity. The developed method first calculates six time series characteristic indices from various perspectives. Then the principal component analysis is performed to reduce the data dimension while preserving the important information. The diversity of the time series dataset is visualized by the geometric distribution of the newly constructed principal component space. The volume of the 3-dimensional (3D) convex polytope (or the length of 1D number axis, or the area of the 2D convex polygon) is used to quantify the time series data diversity. The method is tested with five datasets with various degrees of diversity.

KEYWORDS

wind forecasting; time series analysis; data diversity; big data visualization; machine learning

1 INTRODUCTION

As a renewable energy resource, notable progress has been made in wind energy in the past decade. However, the uncertain and variable characteristics of the wind resource pose challenges to further increases in wind penetration. These challenges can be partially addressed by improving the accuracy of wind speed and power forecasting. Accurate wind forecasting benefits wind integration by assisting economic and reliable power system operations from different perspectives. Significant improvements in wind forecasting have been achieved by developments in forecasting models. Wind forecasting models can be classified into differing categories based on the algorithm principles, and are generally divided into physical models (e.g., numerical weather prediction models), statistical models (e.g., machine learning models), and hybrid physical and statistical models.

Different types of statistical methods have been applied in wind forecasting, including traditional statistical methods (e.g., time series methods), machine learning methods, and deep learning methods. Traditional statistical methods, such as autoregressive integrated moving average (ARIMA) [1], have been initially adopted for wind forecasting. Then the machine learning algorithms have been recently used for wind forecasting due to their powerful learning abilities, such as the neural networks, support vector machines, etc [2]. Another group of statistical methods is deep learning methods. Wang et al. developed both the deterministic and probabilistic models based on the deep learning methods recently [3, 4]. Compared to shallow machine learning methods, deep learning methods are expected to capture hidden invariant structures in wind speed/power. More details about the wind forecasting methods are reviewed in [5–8].

Besides the learning abilities, the performance of these statistical methods varies greatly based on locations, forecasting horizons, training data sizes, and other factors. For example, the SVM algorithm was reported to outperform the backpropagation neural network in [9]. However, the SVM models with linear and polynomial kernels were worse than the radial based function neural network model in [10]. Additionally, ARIMA performed better than ANN in [1] but was worse than ANN in [11]. The situation becomes more complicated when several algorithms are hybridized to improve the forecasting. The conflicting results are largely due to the small validation datasets utilized for the studies. For instance, data from only one location is used to test the LSSVM-GSA model in [9]. Even though three locations' data was applied in the case studies in [11], but case one only had 100 samples in the testing data and the total length of cases two and three was only fifteen days. Since the superiority of different data-driven algorithms hasn't been proved theoretically, the data selected for case studies is especially important. To the best of our knowledge, the generality of the experimental data has not been well quantified and evaluated in the literature. To bridge this gap, this paper proposes a method to visualize and quantify the generality and diversity of the time series datasets, which is validated by five wind time series datasets.

The remainder of the paper is organized as follows. Section 2 develops the method to characterize the diversity and generality of the dataset. The testing datasets with different diversity are described in Section 3. Section 4 presents the experimental results and discussion. The conclusions are drawn in Section 5.

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

BDCAT'17, December 5–8, 2017, Austin, TX, USA

© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.

ACM ISBN 978-1-4503-5549-0/17/12…$15.00
https://doi.org/10.1145/3148055.3148065

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

2 DATA DIVERSITY JUSTIFICATION METHOD

Machine learning methods for wind forecasting proposed in the literature are usually evaluated by data for a limited number of locations with a relatively small length of test data, which is usually insufficient for general applications. To justify the diversity and generality of the data, a time series characteristic analysis (TSCA) technique is developed for the target forecasted time series, i.e., wind speed or wind power. First, the characteristic indices (*CI*s) of a time series are extracted to represent its features from different perspectives. Then principal component analysis (PCA) is performed to reduce the dimension of the *CI* space. The reduced *CI* space is visualized and quantified by the geometric distribution.

2.1 Characterizing Wind Time Series

The TSCA method has been used in time series classification [12], anomalous time series detection [13], and the forecasting domain [14]. A collection of time series *CI*s has been utilized in the literature to quantify the time series characteristics in the fields of demography, finance, and economics fields [15]. In this study, six *CI*s are selected based on the nature of the wind time series: the strength of trend, the strength of seasonality, the skewness and kurtosis of the wind time series distribution, the nonlinearity, and the spectral entropy. Seasonality and trend are two wind time series characteristics considered in time series forecasting models [16, 17]. Skewness and kurtosis provide information of the asymmetry and the tail of the wind distribution in wind forecasting, respectively [18]. Nonlinearity and spectral entropy represent the complexity and chaos of the wind series, respectively, which highly impact the forecasting performance. Hence, we believe these six *CI*s can comprehensively quantify the wind time series characteristics in a static manner. The mathematical explanations of the six *CI*s are described as follows.

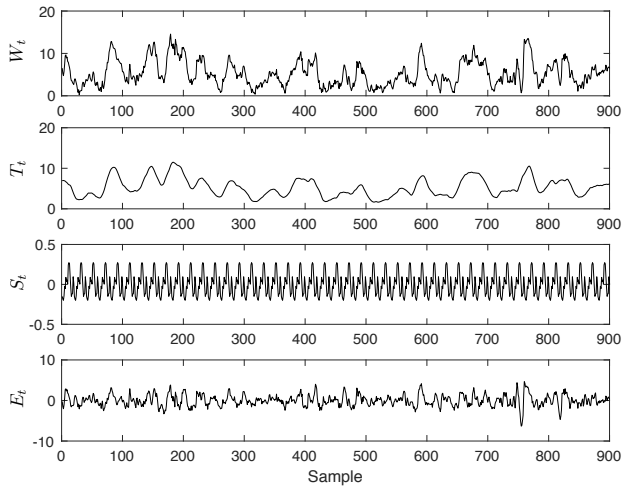


Figure 1: Decomposition of the wind speed series.

- Strength of trend CI_1: The trend is the long-run increase or decrease in the time series. To quantify the trend in the wind time series, additive decomposition is performed using seasonal trend decomposition based on Loess [19], which can be described as:

$$W_t = S_t + T_t + E_t \tag{1}$$

where W_t, T_t, S_t, and E_t are the original wind, trend, season, and remainder series, respectively, which are shown in Fig. 1. The strength of trend is defined as [14]:

$$CI_1 = 1 - \frac{var(E_t)}{var(W_t - S_t)} \tag{2}$$

where $(W_t - S_t)$ is the de-seasonalised series, E_t is the de-trended and de-seasonalised series, and var is the variance operator.

- Strength of seasonality CI_2: Seasonality is wavelike fluctuations of constant length. Similar to CI_1, the strength of seasonality is defined as [20]:

$$CI_2 = 1 - \frac{var(E_t)}{var(W_t - T_t)} \tag{3}$$

where $(W_t - T_t)$ is the de-trended series.

- Skewness coefficient CI_3: The skewness of a univariate distribution can be quantified by the Pearson's moment coefficient of skewness, which is defined as the third moment of this random variable [21]:

$$CI_3 = E\left[\left(\frac{W_t - \mu}{\sigma}\right)^3\right] \tag{4}$$

where E is the expectation operator, μ is the mean value, and σ is the standard deviation.

- Kurtosis coefficient CI_4: The kurtosis of the wind distribution is measured by the Pearson's moment coefficient of kurtosis, which is defined as the fourth moment of the random variable:

$$CI_4 = E\left[\left(\frac{W_t - \mu}{\sigma}\right)^4\right] \tag{5}$$

- Nonlinearity CI_5: Wind data often has a highly nonlinear nature, which increases the forecasting difficulty. The nonlinearity measures the nonlinear structure in the time series. In this study, Teräesvirta's neural network test is selected to quantify the nonlinearity [22].
- Spectral entropy CI_6: Entropy describes the uncertainty and complexity in the time series. A large entropy indicates a more uncertain and chaotic time series. To determine the entropy, the spectral entropy analysis is used to calculate the Shannon entropy of the wind time series [23]:

$$CI_6 = -\sum_w P(w) \log_2 [P(w)] \tag{6}$$

where $P(w)$ is the probability in the state w.

2.2 Principal Component Analysis (PCA) Dimension Reduction

PCA is a widely used feature selection and reduction method in the time series analysis [2]. After extracting *CI*s of each time series from the dataset, the normalization method is applied to standardize

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

every CI separately [24]. The principal components are extracted by the singular value decomposition (SVD) as [25]:

$$\boldsymbol{CI} = \boldsymbol{U}\Sigma\boldsymbol{W}^{\boldsymbol{T}} \tag{7}$$

where $\boldsymbol{CI} \in \mathbb{R}^{N\times 6}$ is the normalized CI matrix, $\boldsymbol{U} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{W} \in \mathbb{R}^{6\times 6}$ are the left and right orthogonal matrices conforming $\boldsymbol{U^T U = I_N}$ and $\boldsymbol{W^T W = I_6}$, respectively, and $\Sigma \in \mathbb{R}^{N\times 6}$ is a rectangular diagonal matrix of positive numbers, $\sigma_i, i = 1, 2, \ldots, N$. $\boldsymbol{U\Sigma}$ (denoted as $\boldsymbol{T}$) is the principal component matrix and $\boldsymbol{W^T}$ gives the corresponding coefficients.

In the data dimension reduction, the cumulative contributions of principal components are used to select the useful principal components by:

$$\begin{cases} \sum_{i=1}^{p} \sigma_i' / \sum_{i=1}^{6} \sigma_i & \sigma_i' \geq \xi \\ \sum_{i=1}^{p-1} \sigma_i' / \sum_{i=1}^{6} \sigma_i & \sigma_i' < \xi \end{cases} \tag{8}$$

where σ' is the descending σ array, ξ is the pre-specified threshold (that is 80% in this paper), and p is the number of principal components.

The reduced principal component matrix with the selected principal components can be derived from Eq. 7, given by:

$$\boldsymbol{T_r} = \begin{bmatrix} \boldsymbol{PC_1} & \boldsymbol{PC_2} & \ldots & \boldsymbol{PC_p} \end{bmatrix} \subseteq \boldsymbol{T} = \boldsymbol{CI \cdot W} \tag{9}$$

where PI_i is the ith principal component.

2.3 Diversity Visualization and Quantification

To further measure the diversity of each dataset, a two-step diversity justification method is developed for visualization and quantification. The proposed method is based on the geometric characteristic, therefore is adaptable with different instance space dimensions (determined by p value). The visualization and quantification method in a 3-dimension (3D) space case is detaily described, and other space dimension cases are also briefly discussed.

In the 3D space, the distribution of the scatter points characterizes the diversity. First, the convex polytope of the finite point set is constructed by a combination of the two-dimensional Quick-hull Algorithm and the general-dimension Beneath-Beyond Algorithm, which is described by [26]:

$$Conv(\boldsymbol{S}) = \left\{ \sum_{i=1}^{|S|} \alpha_i x_i \,|\, (\forall \alpha_i \geqslant 0) \bigwedge \sum_{i=1}^{|S|} \alpha_i = 1 \right\} \tag{10}$$

where $\boldsymbol{S} \subseteq \mathbb{R}^3$ is a collection of points in the 3D space; x_i means the ith point; α_i is the corresponding coefficient. Second, the volume of the convex polytope (Vol_S) formed by the convex hull is defined as the diversity (Div) of S, which is solved by the Delaunay triangulation algorithm [27].

For lower- or higher-dimensional spaces, this diversity quantification approach can be adjusted. Considering the 1D case, the length of the 1D scatter points on the axis represents the diversity of the dataset. For the 2D space, the minimum polygon of the 2D scatter points is constructed and its area quantifies the diversity of the dataset. In case of an instance space with dimension higher than three, the 3D projections of the high-dimension data characterize the diversity of the dataset and the average value of Vol_Ss measure the overall diversity of the dataset.

3 EXPERIMENTAL DATASETS

To validate the proposed TSCA method, the diversity of five datasets are quantified, which are the Global Energy Forecasting Competition 2012 (GEFCom2012) dataset *, the Global Energy Forecasting Competition 2014 (GEFCom2014) dataset †, the Surface Radiation Budget Network (SURFRAD) dataset ‡, the Wind Integration National Dataset (WIND) Toolkit dataset §, and the Comparison of Numerical Weather Prediction (CompNWP) dataset [28]. These datasets contain measurements or simulated wind power/speed data and meteorological data in Australia and the United States. Each dataset contains data from several locations with various time spans. The variables and other standard information are summarized in Table 1. The combination (COMB) of the five datasets is also included in the visualization and quantification step for better comparison. The detailed dataset information and selection criteria are described in the rest of this section.

3.1 The Global Energy Forecasting Competition 2012 (GEFCom2012) Dataset

The GEFCom2012 dataset contains three years of hourly measured wind power data from seven wind farms in the same region. Additional meteorological data was obtained from the European Centre for Medium-range Weather Forecasts (ECMWF) model. The wind power data is normalized between 0 and 1. Since the GEFCom2012 data was prepared for the competition, there are periods with intentionally missing data points [29]. The only completely available variable is the wind power, which is used in this study.

3.2 The Global Energy Forecasting Competition 2014 (GEFCom2014) Dataset

The GEFCom2014 dataset contains hourly wind farm data from 10 locations in Australia, spanning from 2012-01-01 to 2012-10-01. The variables in this dataset include the zonal and meridional wind components forecasted by ECMWF at 10 and 100 meters height ($U_{10}, V_{10}, U_{100}, V_{100}$), and the wind power ($WP$) generation data. The wind power data is normalized by the nominal capacities of the wind farm. More details about this dataset can be found in [30].

3.3 The Surface Radiation Budget Network (SURFRAD) Dataset

SURFRAD was established to support climate research. The SURFRAD dataset collects meteorological data in climatologically diverse regions around the continental US, based on ground-based sensors. In this paper, the hourly data from seven locations is used, spanning from 2015-01-01 to 2015-12-31 [2]. The data contains five variables, which are the wind speed, wind direction, relative humidity, atmosphere pressure, and temperature measured at a height below 10 m (far below the height of large-scale wind turbines).

*http://www.drhongtao.com/gefcom/2012
†http://www.drhongtao.com/gefcom/2014
‡https://www.esrl.noaa.gov/gmd/grad/surfrad/
§http://www.nrel.gov/grid/wind-toolkit.html

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Table 1: Dataset summary

Dataset	No. of locations	Variable (forecasted variable)	Length
GEFCom2012	7	W_P (W_P)	<1 year
GEFCom2014	10	W_P, U_{10}, V_{10}, U_{100}, V_{100} (W_P)	<1 year
SURFRAD	7	W_S, H, T, WD, P (W_S)	1 year
WIND Toolkit	5 (selected from 126, 000+)	W_P, W_S, H, T, WD, P (W_P)	7 years
CompNWP	8	W_S (W_S)	1 - 4 year(s)

Note: W_P means wind power, U_{10} V_{10} U_{100} V_{100} are zonal and meridional wind components at 10 m and 100 m heights, W_S means wind speed, H means relative humidity, T means temperature, WD means wind direction, and P means pressure. The WIND Toolkit dataset has the simulated forecasted variable, while the other datasets have measured forecasted variables. The SURFRAD data is measured at 10 m height or below, and the data in other datasets is measured/simulated at different turbine-scale heights.

3.4 The Wind Integration National Dataset (WIND) Toolkit Dataset

WIND Toolkit was developed for the next generation of wind integration studies. The WIND Toolkit dataset is composed of meteorological dataset, generated by the Weather Research and Forecasting model with a 2 km grid, and the wind power dataset [31]. The dataset contains seven years' data, spanning from 2007-01-01 to 2013-12-31, at more than 126,000 wind locations with a 5-min resolution. In this paper, five wind farms near Dallas, New York City, Chicago, Miami, and Los Angeles are selected for the sake of topographical diversity. The data is averaged from five-minute to an hourly resolution.

3.5 The Comparison of Numerical Weather Prediction (CompNWP) Dataset

The CompNWP dataset is a collection of hub-height wind speed measurements at eight locations across the United States used in our previous research [28]. The dataset is created based on several criteria: (i) the data is collected from locations with different topography and climates; (ii) the data is measured at different hub-heights (all above 50 m); (iii) the data has a variety of time periods at different locations. The location and topographical information can be found in [28].

4 RESULTS AND DISCUSSION

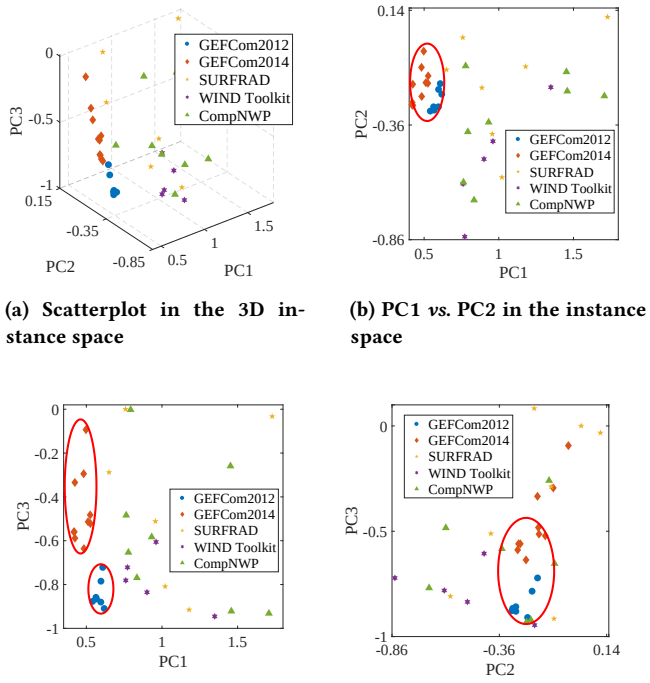
4.1 Characterizing Data Diversity

The *CI*s of time series in each dataset are extracted using Eqs. 1 - 6 first. Then, PCA is utilized to map the six-dimension space to a smaller principal component space. Using Eq. 8, it is found that the first three principal components (PCs) cover 82.13% of the information in the original data. The linear transformation from the CI space to the first three PCs is given by:

$$\boldsymbol{T_r} = \begin{bmatrix} \boldsymbol{PC_1} & \boldsymbol{PC_2} & \boldsymbol{PC_3} \end{bmatrix} = \boldsymbol{CI}^T \boldsymbol{W_r} \tag{11}$$

where $\boldsymbol{CI} = \begin{bmatrix} \boldsymbol{CI_1} & \boldsymbol{CI_2} & \boldsymbol{CI_3} & \boldsymbol{CI_4} & \boldsymbol{CI_5} & \boldsymbol{CI_6} \end{bmatrix}^T$, and $\boldsymbol{W_r}$ is the reduced right orthogonal matrix.

By performing the previous steps, the data of each location is represented by one point in a 3D space, as shown in Fig. 2a. Different markers represent different datasets. Each point stands for the target time series of one location. The projection drawings of the 3D plot are shown in Figs. 2b - 2d. It is observed that some datasets,



(a) Scatterplot in the 3D instance space

(b) PC1 *vs.* PC2 in the instance space

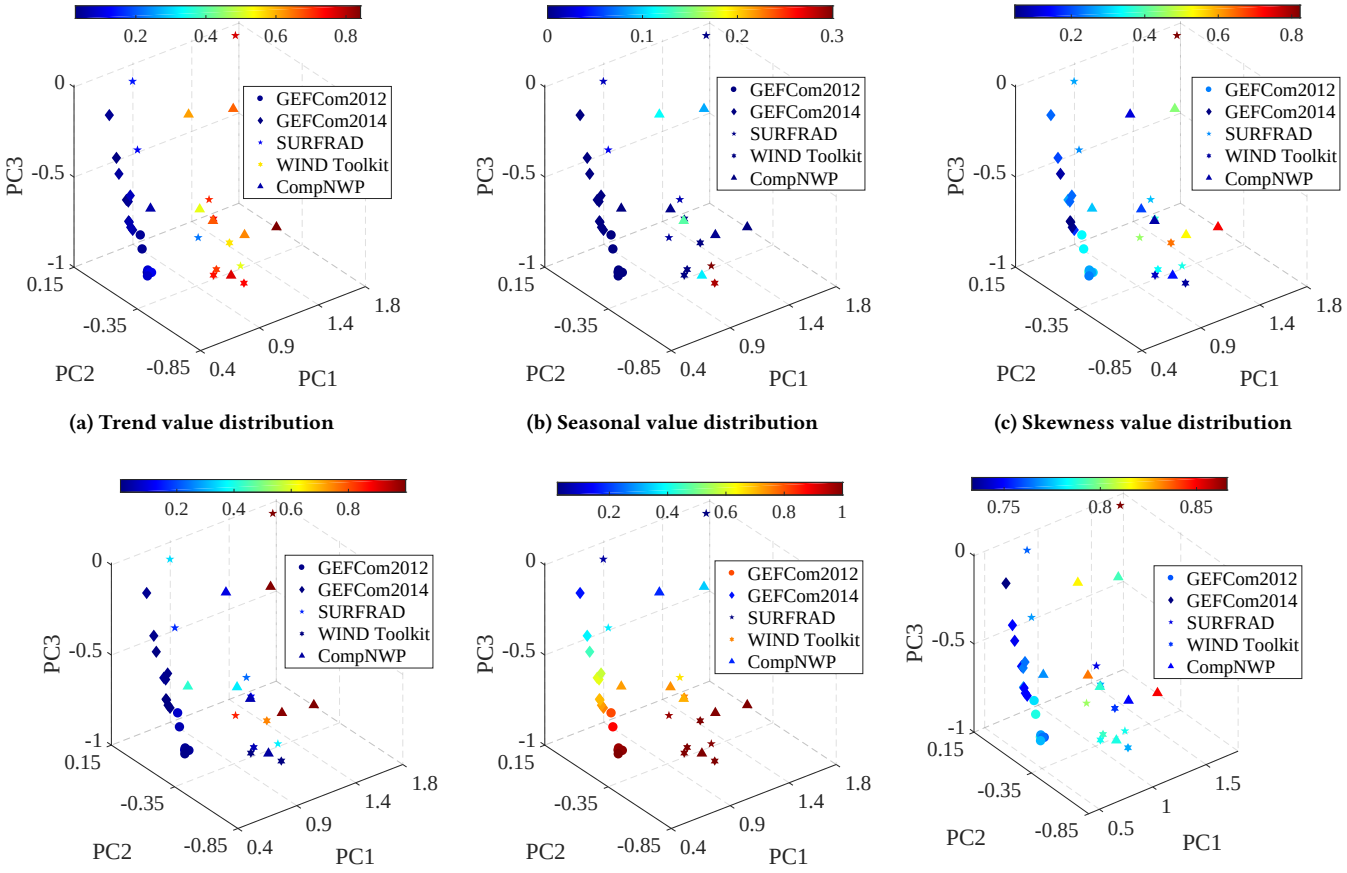
(c) PC1 *vs.* PC3 in the instance space

(d) PC2 *vs.* PC3 in the instance space

Figure 2: Instance space of the target wind series in the WFSD dataset.

such as GEFCom2012 and GEFCom2014, are concentrated in a small region in the 3D space, which means the different data within these two datasets has similar characteristics. This may be due to the highly topological similarity of the Australian locations, where the data was measured. Comparing the GEFCom2012 and GEFCom2014 scatter points, the GEFCom2014 is more diverse in the PC3 direction. By comparing the WIND Toolkit dataset to the GEFCom2012 and GEFCom2014 datasets, it is found that the simulated WIND Toolkit data is more diverse than the measured data. However, the WIND Toolkit dataset is less diverse in the PC3 direction than the GEFCom2014, SURFRAD, and CompNWP datasets. The SURFRAD and CompNWP datasets contain wind speed measurements at different heights. The SURFRAD data is measured at a low height (<

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.



(a) Trend value distribution **(b) Seasonal value distribution** **(c) Skewness value distribution**

(d) Kurtosis value distribution **(e) Nonlinearity value distribution** **(f) Entropy value distribution**

Figure 3: CI distributions of the wind series characteristic indices in a 3D space. Color bars indicate the CI value.

Table 2: The CI statistics of the WFSD dataset

		Time series characteristic index					
		CI_1	CI_2	CI_3	CI_4	CI_5	CI_6
GEFCom2012	μ	0.08	0	0.30	0.04	0.95	0.77
	σ	0.03	0	0.04	0.02	0.07	0.01
GEFCom2014	μ	0.11	0	0.17	0.03	0.57	0.75
	σ	0.19	0	0.09	0.06	0.17	0.01
SURFRAD	μ	0.44	0.06	0.39	0.44	0.44	0.79
	σ	0.25	0.10	0.19	0.31	0.40	0.04
WIND Toolkit	μ	0.69	0.06	0.30	0.18	0.95	0.78
	σ	0.07	0.11	0.21	0.28	0.10	0.01
CompNWP	μ	0.60	0.06	0.32	0.49	0.71	0.80
	σ	0.22	0.06	0.22	0.41	0.29	0.03

Note: μ is the mean value and σ is the standard deviation.

10 m) while the CompNWP data is recorded at above 50 m height. But both the SURFRAD and the CompNWP datasets show a high diversity in all the three directions.

More details can be found from the CI mean (μ) and standard deviation (σ) values of each dataset, which are listed in Table 2. The wind series in different datasets present various strength of trend (CI_1). For example, the average CI_1 of the GEFCom2012 data is 0.08, while the average CI_1 of the WIND Toolkit dataset increases to 0.69. The CI_1 standard deviation can be as high as 0.19, which means differences of the trend in different series within the same dataset are also distinct. Similar findings are observed by comparing the values in the CI_3, CI_4, and CI_5 columns. However, the wind time series show a relatively consistent seasonality (CI_2) and entropy (CI_6) within the same dataset and among different datasets.

The CI value distributions of the five datasets are visualized in Fig. 3, which provides a better insight of the data characteristics. The color of each point indicates the CI magnitude, and different markers represent datasets. Figure 3a shows that the GEFCom2012 and GEFCom2014 datasets have small trend values while the WIND Toolkit and CompNWP datasets have large trend values. Additionally, it is interesting to find that the low height (< 10 m) measurement series in SURFRAD dataset has a broader range of trend values. The seasonality of the data is consistently low in all datasets, especially in the GEFCom2012 and GEFCom2014 datasets. Scaled skewness and kurtosis are shown in Figs. 3c and 3d, which measure the asymmetry and peakness of the wind series distributions,

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

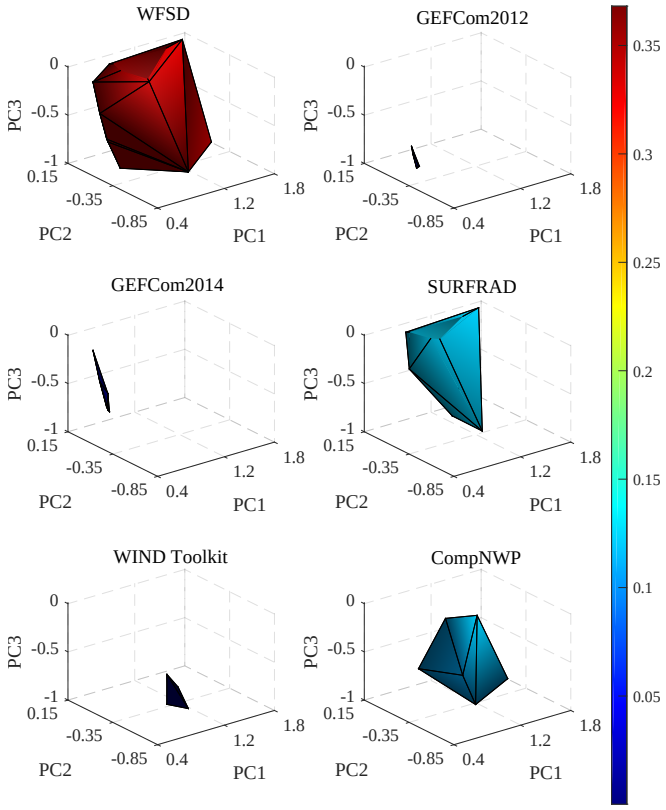


Figure 4: The minimum convex polytopes of different datasets in the 3D space. The color bar indicates the value of Div. The Div values are 3.7×10^{-1}, 1.3×10^{-4}, 6.4×10^{-4}, 1.3×10^{-1}, 6.2×10^{-3}, and 1.1×10^{-1} with respect to COMB, GEFCom2012, GEFCom2014, SURFRAD, WIND Toolkit, and CompNWP datasets, respectively.

respectively. A large skewness value indicates a clear asymmetry, and a large kurtosis value means a sharp distribution. It is observed from these two figures that most datasets have little asymmetry and low peakedness, except for the SURFRAD and CompNWP datasets. For the nonlinearity as shown in Fig. 3e, most data series have a high nonlinear characteristic. Figure 3e also shows that different data series in the same dataset may have a large variance, such as the SURFRAD data. All the series have relatively large entropy values, as shown in Fig. 3f. Moreover, the series in the same dataset has small variance, with σ less than 0.04 as shown in Table 2. Some other patterns are also observed through the 3D visualization. For example, the strength of trend increases along the PC1 direction and the nonlinearity decreases along the PC3 direction. This suggests that the trend and the nonlinearity have a strong linear relationship with PC1 and PC3, respectively.

Figure 4 shows the constructed convex polytope of the five datasets and the COMB dataset (for comparison purpose). By comparing the size of the polytopes, it is observed that the GEFCom2012 and GEFCom2014 datasets have the lowest diversity. But the GEFCom2014 dataset is approximately five times more diverse than GEFCom2012, due to the larger span in the PC3 direction. The WIND Toolkit dataset also has relatively low diversity even though the selected locations are geographically diverse. Since the wind power series in the WIND Toolkit sub-dataset is converted from the simulated wind speed series, the low diversity may be due to the similar physical laws applied in the Weather Research and Forecasting (WRF) model at different locations [32]. It is important to note that only 5 out of over 126,000 WIND Toolkit locations are selected in this case study. The SURFRAD and CompNWP datasets have significantly larger diversity than the other three datasets. The COMB dataset is much more diverse than any of the single datasets.

4.2 Forecasting Uncertainty Validation

The diversity of every dataset has been quantified so far, and the results are validated in this section. The 1-h ahead forecasts are produced by gradient boosting machine (GBM). Two uncertainty metrics, Rényi entropy (H_R) and correlation coefficient (r), are used to measure the chaos in the forecasted series.

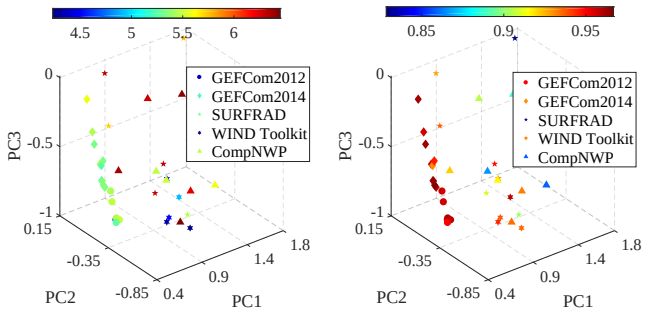
GBM is an ensemble machine learning algorithm, which does not need preprocessing compared to other machine learning algorithms such as ANN and SVM. GBM model relies on the combination of 'weak learners' to create an accurate learner. The combination is achieved by adding the weighted base learner to the previous model iteratively. The mathematical description of the GBM algorithm can be found in [2]. The GBM models are trained by 75% of the data in each time series, and are used to generate 1-h ahead forecasts for the rest 25% data.

Two evaluation metrics are chosen to measure the forecasting uncertainty. The forecasted series Rényi entropy (H_R) is able to quantify the chaos in the forecasted values. The correlation coefficient between the forecasted and the actual series (r) represents the linear relation between the two series [33]. A larger H_R value means the forecasted series is more chaotic and a smaller r value means it's more challenging to generate the forecasts from the original series. The distributions of the two metrics are shown in Fig. 5. In Fig. 5a, the SURFRAD and CompNWP datasets have forecasted series with larger H_R values (above 5.5) compared to GEFCom2012, GEFCom2014, and WIND Toolkit datasets. Comparing the r values in the five datasets, it is found that the correlation in the SURFRAD and CompNWP time series is smaller than the other three datasets. Both the two metrics indicate that the SURFRAD and CompNWP datasets are more diverse than the other three datasets. This is because the learning ability and the forecasting power of the same-algorithm model (i.e., GBM) is constant. Therefore, the dataset with large diversity will have more chaos and weak correlation with the input series. The forecasting results have shown that the proposed time series characteristic analysis method can successfully quantify the diversity of forecasting datasets.

5 CONCLUSION

This paper developed an approach to quantify the diversity of the time series dataset, based on the time series characteristic analysis (TSCA method). Five wind datasets with different diversity were

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.



(a) Forecasting series Rényi entropy

(b) Forecasting correlation coefficient

Figure 5: Distributions of forecasting uncertainty metrics in a 3D space. Color bars indicate H_R and r values. Different markers represent different datasets.

used for the numerical experiment. Six time series characteristic indices (*CI*s) were first extracted from each wind series. Then the principal component analysis (PCA) was used to reduce the *CI* dimension from six to three, by preserving 82.13% of the information. The diversity of the dataset was visualized and quantified by the *CI* distributions in the 3D space. To quantify the diversity, the volume of the minimum convex polytope formed by the scatter points was calculated, which was defined as the dataset diversity. The developed method was validated by evaluating the 1-h ahead gradient boosting machine forecasting uncertainty. The developed TSCA method is adaptive to be applied in other forecasting tasks, such as solar forecasting and electricity load forecasting. For future work, a systematic framework will be developed to adjust and apply the TSCA method in different time series forecasting.

ACKNOWLEDGMENTS

This work was supported by the National Renewable Energy Laboratory under Subcontract No. XGJ-6-62183-01 (under the U. S. Department of Energy Prime Contract No. DE-AC36-08GO28308). This work was authored in part by Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Water Power Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

REFERENCES

[1] Hui Liu, Hong qi Tian, and Yan fei Li. An EMD-recursive ARIMA method to predict wind speed for railway strong wind warning system. *Journal of Wind Engineering and Industrial Aerodynamics*, 141:27–38, jun 2015.

[2] Cong Feng, Mingjian Cui, Bri-Mathias Hodge, and Jie Zhang. A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. *Applied Energy*, 190:1245–1257, 2017.

[3] HZ Wang, GB Wang, GQ Li, JC Peng, and YT Liu. Deep belief network based deterministic and probabilistic wind speed forecasting approach. *Applied Energy*, 182:80–93, 2016.

[4] Huai-zhi Wang, Gang-qiang Li, Gui-bing Wang, Jian-chun Peng, Hui Jiang, and Yi-tao Liu. Deep learning based ensemble approach for probabilistic wind power forecasting. *Applied Energy*, 188:56–70, 2017.

[5] Ma Lei, Luan Shiyan, Jiang Chuanwen, Liu Hongling, and Zhang Yan. A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, 13(4):915–920, 2009.

[6] Aoife M Foley, Paul G Leahy, Antonino Marvuglia, and Eamon J McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable Energy*, 37(1):1–8, 2012.

[7] Xin Zhao, Shuangxin Wang, and Tao Li. Review of evaluation criteria and main methods of wind power forecasting. *Energy Procedia*, 12:761–769, 2011.

[8] Jaesung Jung and Robert P Broadwater. Current status and future advances for wind speed and power forecasting. *Renewable and Sustainable Energy Reviews*, 31:762–777, 2014.

[9] Xiaohui Yuan, Chen Chen, Yanbin Yuan, Yuehua Huang, and Qingxiong Tan. Short-term wind power prediction based on lssvm–gsa model. *Energy Conversion and Management*, 101:393–401, 2015.

[10] Hassen Bouzgou and Nabil Benoudjit. Multiple architecture system for wind speed prediction. *Applied Energy*, 88(7):2463–2471, jul 2011.

[11] Hui Liu, Hong-qi Tian, Di-fu Pan, and Yan-fei Li. Forecasting models for wind speed using wavelet, wavelet packet, time series and artificial neural networks. *Applied Energy*, 107:191–208, 2013.

[12] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*, 7(4):349–371, 2003.

[13] Michael E Mann and Jonathan M Lees. Robust estimation of background noise and signal detection in climatic time series. *Climatic change*, 33(3):409–445, 1996.

[14] Yanfei Kang, Rob J Hyndman, and Kate Smith-Miles. Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, 33(2):345–358, 2017.

[15] Rob J Hyndman, Earo Wang, and Nikolay Laptev. Large-scale unusual time series detection. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pages 1616–1619. IEEE, 2015.

[16] Erasmo Cadenas and Wilfrido Rivera. Wind speed forecasting in three different regions of mexico, using a hybrid arima–ann model. *Renewable Energy*, 35(12):2732–2738, 2010.

[17] Heping Liu, Ergin Erdem, and Jing Shi. Comprehensive evaluation of arma–garch (-m) approaches for modeling the mean and volatility of wind speed. *Applied Energy*, 88(3):724–732, 2011.

[18] Jose Luis Torres, Almudena Garcia, Marian De Blas, and Adolfo De Francisco. Forecast of hourly average wind speed with arma models in navarre (spain). *Solar Energy*, 79(1):65–77, 2005.

[19] Robert B Cleveland, William S Cleveland, and Irma Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3, 1990.

[20] Xiaozhe Wang, Kate Smith, and Rob Hyndman. Characteristic-based clustering for time series data. *Data mining and knowledge Discovery*, 13(3):335–364, 2006.

[21] Stan Brown. Measures of shape: Skewness and kurtosis. *Retrieved on August*, 20:2012, 2011.

[22] Timo Teräsvirta, Chien-Fu Lin, and Clive WJ Granger. Power of the neural network linearity test. *Journal of Time Series Analysis*, 14(2):209–220, 1993.

[23] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[24] Cong Feng, Mingjian Cui, Meredith Lee, Jie Zhang, Bri-Mathias Hodge, Siyuan Lu, and Hendrik F Hamann. Short-term global horizontal irradiance forecasting based on sky imaging and pattern recognition. In *IEEE PES general meeting 2017*. IEEE PES, 2017.

[25] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[26] C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996.

[27] Der-Tsai Lee and Bruce J Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980.

[28] Jie Zhang, Caroline Draxl, Thomas Hopson, Luca Delle Monache, Emilie Vanvyve, and Bri-Mathias Hodge. Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods. *Applied Energy*, 156:528–541, oct 2015.

[29] Tao Hong, Pierre Pinson, and Shu Fan. Global energy forecasting competition 2012. *International Journal of Forecasting*, 30(2):357–363, apr 2014.

[30] Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J Hyndman. Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond, 2016.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

[31] Caroline Draxl, Andrew Clifton, Bri-Mathias Hodge, and Jim McCaa. The wind integration national dataset (wind) toolkit. *Applied Energy*, 151:355–366, 2015.
[32] William C Skamarock, Joseph B Klemp, Jimy Dudhia, David O Gill, Dale M Barker, Wei Wang, and Jordan G Powers. A description of the advanced research wrf version 2. Technical report, DTIC Document, 2005.
[33] Jie Zhang, Anthony Florita, Bri-Mathias Hodge, Siyuan Lu, Hendrik F Hamann, Venkat Banunarayanan, and Anna M Brockway. A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111:157–175, 2015.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.