# Short-Term Wind Power Forecasting Using Nonnegative Sparse Coding

Yu Zhang
Dept. of ECE and the DTC
University of Minnesota
Minneapolis, MN 55455, USA
Email: yuzhang@umn.edu

Seung-Jun Kim
Dept. of Computer Sci. and Electrical Engr.
University of Maryland, Baltimore County
Baltimore, MD 21250, USA
Email: sjkim@umbc.edu

Georgios B. Giannakis
Dept. of ECE and the DTC
University of Minnesota
Minneapolis, MN 55455, USA
Email: georgios@umn.edu

***Abstract*—State-of-the-art statistical learning techniques are adapted in this contribution for real-time wind power forecasting. Spatio-temporal wind power outputs are modeled as a linear combination of "few" atoms in a dictionary. By exploiting geographical information of wind farms, a graph Laplacian-based regularizer encourages positive correlation of wind power levels of adjacent farms. Real-time forecasting is achieved by online nonnegative sparse coding with elastic net regularization. The resultant convex optimization problems are efficiently solved using a block coordinate descent solver. Numerical tests on real data corroborate the merits of the proposed approach, which outperforms competitive alternatives in forecasting accuracy.**

## I. Introduction

With environmental and economical ramifications of fossil fuel-based generation, renewable energy sources, such as wind and solar, have been actively pursued over the last few decades. Achieving over a 30% annual growth rate, wind power generation reached 318 GW worldwide capacity by the end of 2013 [7]. Both the United States and European Union set a goal to recruit wind energy to meet up to 20% of electricity demands by 2030 and 2020, respectively [1].

However, full benefits of wind power can only be achieved by properly mitigating its inherent variability and limited predictability. In both forward and spot electricity markets, renewable asset owners make their strategic bidding decisions according to wind predictions [19]. Wind forecasts may also serve for unit commitment and economic dispatch implemented by the independent system operators to ensure the grid's economic and reliable operation. In addition, accurate wind forecasts are useful for planning the maintenance of conventional power plants, onshore/offshore wind farms, as well as transmission lines [18]. Thus, to boost the wind power penetration in the future grid, it is critical for the system operators, wind power producers, as well as the utility companies to obtain accurate wind energy forecasts.

Wind power generation depends on various meteorological factors (e.g., wind speed/direction and air density/pressure), as well as turbine deployment conditions, such as the type and the height. In addition, annual, seasonal, diurnal and hourly patterns change dramatically across regions. All these factors make wind forecasting quite challenging.

Depending on various applications, wind speed or power forecasting can be classified into three categories in terms of the time scale, namely short-term (up to several hours ahead), medium-term (day ahead), and long-term (several days/months ahead). Existing prediction methods mainly aim at forecasting the wind speed and generation of a single turbine or multiple wind farms via AR(I)MA time-series models [12], [17], neural networks [11], [3], [22], support vector regression [23], [14], and $k$-nearest neighbors regression [21]. Markov-switching autoregressive models were proposed for forecasting of offshore wind power fluctuations [20]. However, schemes of forecasting wind speed and mapping to wind generation are likely to have error magnifications incurred by the nonlinear wind-speed-to-wind-power mappings. Approaches relying on separate processing per time series fail to capitalize on spatial correlations. A space-time Kalman filtering approach was studied to incorporate spatial correlation in forecasting ozone concentration [8], but the computational burden is substantial. State-of-the-art wind forecasting techniques are surveyed in recent articles [15], [5].

In this paper, spatially correlated wind power outputs are directly forecasted using the state-of-the-art machine learning and compressive sensing techniques. Measurements of wind power generation may not always be available due to e.g. meter failures. Hence, imputation of missing observations must be taken into account in the learning task. The framework of dictionary learning [16] is adapted here to learn the spatio-temporal patterns of the wind power outputs. Leveraging the topology information of multiple wind farms, a Laplacian-based regularizer is utilized to aid in spatial interpolation. Nonnegative sparse coding with an elastic-net regularizer is proposed for the interpolation and extrapolation. In addition to a batch algorithm, an online alternative featuring low burden of memory and computations is also developed to capture the temporal correlation of the underlying process.

The remainder of the paper is organized as follows. Section II introduces the system model and the problem. In Section III, the spatio-temporal dictionary learning algorithms are developed for interpolating missing measurements and

This work was supported by the Initiative for Renewable Energy & the Environment (IREE) grant RL-0010-13, Univ. of Minnesota, and NSF grants CCF-1423316 and CCF-1442686.

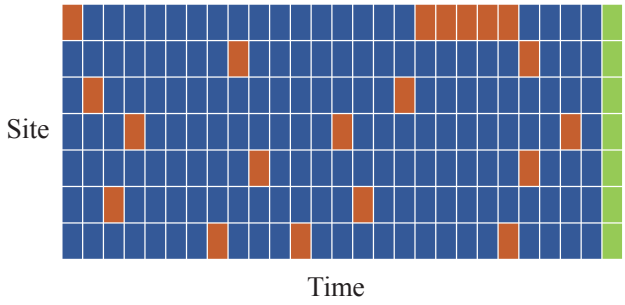


Fig. 1. Wind power prediction problem: Given the observations (blue), interpolate the missing measurements (brown), and forecast the future values (green).

extrapolating future outputs. Numerical results are reported in Section IV, and concluding remarks are given in Section V.

**Notation**. Boldface lower (upper) case letters represent vectors (matrices); $\mathcal{R}^{N\times M}$ and $\mathcal{R}^{N}$ stand for spaces of $N\times M$ and $N\times 1$ real matrices and vectors, respectively; $(\cdot)'$ indicates transpose; $\otimes$ the Kronecker product; and $\succeq$ vector entrywise inequality. $\mathbb{1}_{\{A\}}$ is the indicator function equal to 1 if the condition $A$ is satisfied, and 0 otherwise. $\mathrm{diag}(\mathbf{x})$ returns a square diagonal matrix with the elements of vector $\mathbf{x}$ on the main diagonal. Finally, $\mathbf{x}^+ := \max\{\mathbf{0},\mathbf{x}\}$ denotes the projection of $\mathbf{x}$ onto the nonnegative orthant.

## II. Problem Statement

Consider the wind power generation at sites $\mathcal{N} := \{1,2,\ldots,N\}$, and across time slots $\mathcal{T} := \{1,2,\ldots,T\}$. Let $\mathbf{p}_t := [p_{1,t},\ldots,p_{N,t}]' \in \mathcal{R}^N$ denote the power outputs of all $N$ sites at time $t$. It is postulated that $\mathbf{p}_t$ can be represented as a linear combination of a small number of bases taken from a *sufficiently rich* dictionary $\mathbf{D} := [\mathbf{d}_1,\ldots,\mathbf{d}_M] \in \mathcal{R}^{N\times M}$. Mathematically, this can be written as

$$\mathbf{p}_t = \mathbf{D}\mathbf{s}_t, \quad \forall t\in\mathcal{T} \tag{1}$$

where vectors $\{\mathbf{s}_t\}_{t=1}^T$ are sparse. At each time $t$, a subset $\mathcal{N}_t^{\text{obs}} \subseteq \mathcal{N}$ of wind farms observe $N_t^{\text{obs}} := |\mathcal{N}_t^{\text{obs}}|$ wind power outputs. These measurements can be collected in a vector $\mathbf{y}_t^{\text{obs}} \in \mathcal{R}^{N_t^{\text{obs}}}$ given as

$$\mathbf{y}_t^{\text{obs}} = \mathbf{O}_t\mathbf{p}_t + \mathbf{z}_t, \quad \forall t\in\mathcal{T} \tag{2}$$

where $\mathbf{O}_t \in \mathcal{R}^{N_t^{\text{obs}}\times N}$ is the observation matrix consisting of the $n$-th row of the $N\times N$ identity matrix for all $n\in\mathcal{N}_t^{\text{obs}}$; and $\mathbf{z}_t \in \mathcal{R}^{N_t^{\text{obs}}}$ is the measurement noise vector.

Given the past and current observations $\{\mathbf{y}_\tau^{\text{obs}}\}_{\tau=1}^t$, the task here is to predict the missing wind power $p_{n,t}$ for $n\in\mathcal{N}_t^{\text{miss}} := \mathcal{N}\setminus N_t^{\text{obs}}$, and the future outputs $\mathbf{p}_{t+1}$, as shown in Fig. 1. To solve this problem, a graph based dictionary learning approach will be developed in the ensuing section.

## III. Spatio-Temporal Wind Power Prediction Using Dictionary Learning

Some off-the-shelf dictionaries such as the Fourier, Hadamard, or wavelet bases exhibit good performance in many applications with natural or artificial signals. Instead of using existing dictionaries, a data-driven one is proposed to learn from historical data and the network topology information. Such dictionary learning techniques have been successfully applied to image processing [16], network load prediction [6], and cognitive radio spectrum sensing [13].

### A. Batch Algorithm

1) *Training phase*: Let $\mathbf{S} := [\mathbf{s}_1,\ldots,\mathbf{s}_T] \in \mathcal{R}^{M\times T}$ collect all sparse codes across the time slots $\mathcal{T}$. In order to capture the spatial correlation across wind farms, consider an undirected weighted graph $\mathcal{G}(V,E)$ with $V$ and $E$ denoting the sets of vertices and edges, respectively. The weighted adjacency matrix $\mathbf{W}\in\mathcal{R}^{N\times N}$ has its $(i,j)$-entry $w_{ij}\geq 0$, which can be set inversely proportional to the distance between sites $i$ and $j$. Then, the graph Laplacian matrix $\mathbf{L}$ is defined as

$$\mathbf{L} := \mathrm{diag}(\mathbf{W}\mathbf{1}_N) - \mathbf{W} \tag{3}$$

where $\mathbf{1}_N$ is $N$-dimensional all-ones vector. Dictionary $\hat{\mathbf{D}}$ can then be learned by solving a joint optimization problem

$$\hat{\mathbf{D}} := \underset{\mathbf{D}\in\mathcal{D},\mathbf{S}\in\mathcal{S}}{\arg\min} \frac{1}{T}\sum_{t=1}^T f_t(\mathbf{D},\mathbf{s}_t) \tag{4}$$

where

$$f_t(\mathbf{D},\mathbf{s}_t) := \frac{1}{2}\|\mathbf{y}_t^{\text{obs}} - \mathbf{O}_t\mathbf{D}\mathbf{s}_t\|_2^2 + \lambda_1\|\mathbf{s}_t\|_1 + \frac{\lambda_2}{2}\|\mathbf{s}_t\|_2^2 + \frac{\lambda_L}{2}\mathbf{s}_t'\mathbf{D}'\mathbf{L}\mathbf{D}\mathbf{s}_t \tag{5}$$

$$\mathcal{D} := \{\mathbf{D} \mid \mathbf{d}_m \succeq \mathbf{0}, \|\mathbf{d}_m\|_2^2 \leq 1,\ m=1,\ldots,M\} \tag{6}$$

$$\mathcal{S} := \{\mathbf{S} \mid \mathbf{s}_t \succeq \mathbf{0},\ t=1,\ldots,T\}. \tag{7}$$

The objective (5) consists of three parts: the least-squares data fitting error, the elastic net, and the Laplacian-based regularizers. The elastic net regularizer is a linear combination of the sparsity-promoting $\ell_1$-norm (Lasso) and the $\ell_2$-norm (ridge) of $\mathbf{s}_t$, which can yield a better performance than plain Lasso [24]. The Laplacian-based regularizer can be equivalently rewritten as

$$\mathbf{s}_t'\mathbf{D}'\mathbf{L}\mathbf{D}\mathbf{s}_t = \sum_{i,j=1}^N w_{ij}(p_{i,t}-p_{j,t})^2 \tag{8}$$

which clearly encourages the difference between $p_{i,t}$ and $p_{j,t}$ to be small if the two sites $i$ and $j$ are close. The norm of each atom $\mathbf{d}_m$ is bounded in (6) to prevent the degeneracy of obtaining arbitrarily large $\mathbf{D}$ by replacing $(\mathbf{D},\mathbf{s}_t)$ with $(\frac{1}{c}\mathbf{D}, c\mathbf{s}_t)$ for $c\ll 1$. Since the wind power outputs $\{\mathbf{p}_t\}_{t=1}^T$ are inherently nonnegative, naturally $(\mathbf{D},\mathbf{S})$ are constrained to be nonnegative as well.

Problem (4) is nonconvex, and hence difficult to solve in general. However, if either $\mathbf{D}$ or $\mathbf{S}$ is fixed, the problem is convex in the remaining variable. This motivates the use of the block coordinate descent (BCD) solver, with convergence

**Algorithm 1** Batch dictionary learning

1: Initialize $\hat{\mathbf{D}} = \mathbf{D}_0$
2: **repeat**
3: Find sparse codes with fixed $\hat{\mathbf{D}}$:
4: **for** $t = 1, 2, \ldots, T$ **do**
5: $\hat{\mathbf{s}}_t = \arg\min_{\mathbf{s}_t \succeq \mathbf{0}} f_t(\hat{\mathbf{D}}, \mathbf{s}_t)$
6: **end for**
7: Update the dictionary with fixed $\{\hat{\mathbf{s}}_t\}_{t=1}^T$:
8: $\hat{\mathbf{D}} = \arg\min_{\mathbf{D}\in\mathcal{D}} \sum_{t=1}^{T} \frac{1}{2}\|\mathbf{y}_t^{\text{obs}} - \mathbf{O}_t\mathbf{D}\hat{\mathbf{s}}_t\|_2^2 + \frac{\lambda_L}{2}\hat{\mathbf{s}}_t'\mathbf{D}'\mathbf{L}\mathbf{D}\hat{\mathbf{s}}_t$
9: **until** convergence

guaranteed. Specifically, BCD updates at the $k$-th iteration are

$$\{\hat{\mathbf{s}}_t(k)\} := \arg\min_{\mathbf{S}\in\mathcal{S}} \sum_{t=1}^{T} f_t(\hat{\mathbf{D}}(k-1), \mathbf{s}_t) \tag{9}$$

$$\hat{\mathbf{D}}(k) := \arg\min_{\mathbf{D}\in\mathcal{D}} \sum_{t=1}^{T} f_t(\mathbf{D}, \hat{\mathbf{s}}_t(k)). \tag{10}$$

The overall steps of the batch dictionary learning are tabulated in Algorithm 1. Note that here $\|\mathbf{s}\|_1$ is equivalent to the sum of all its entries $\sum_{m=1}^{M} s_m$ because $\mathbf{s} \succeq \mathbf{0}$. Therefore, the problem in step 5 is essentially a convex quadratic program (QP) constrained in the positive orthant, which can be efficiently solved by the BCD or off-the-shelf QP solvers. Closed-form updates are available for step 8 via further applying BCD over the columns of $\mathbf{D}$; see e.g., [13].

2) *Operational phase*: Upon obtaining $\hat{\mathbf{D}}$ from historical data, the power outputs $\{\hat{\mathbf{p}}_t\}$ can be predicted in the operational phase. The "optimal" sparse code corresponding to $\hat{\mathbf{D}}$ and $\mathbf{y}_t^{\text{obs}}$ is first obtained by solving the problem

$$\hat{\mathbf{s}}_t = \arg\min_{\mathbf{s}_t\succeq\mathbf{0}} \frac{1}{2}\|\mathbf{y}_t^{\text{obs}} - \mathbf{O}_t\hat{\mathbf{D}}\mathbf{s}_t\|_2^2 + \lambda_1\|\mathbf{s}_t\|_1 + \frac{\lambda_2}{2}\|\mathbf{s}_t\|_2^2 + \frac{\lambda_L}{2}\mathbf{s}_t'\hat{\mathbf{D}}'\mathbf{L}\hat{\mathbf{D}}\mathbf{s}_t. \tag{11}$$

Then, the wind power generation $\mathbf{p}_t$ with possible missing entries can be recovered by $\hat{\mathbf{p}}_t = \hat{\mathbf{D}}\hat{\mathbf{s}}_t$.

### *B. Online Algorithm*

An online dictionary learning algorithm is useful to track the time-varying signals, and reduce the computational complexity [13]. Specifically, both the dictionary $\hat{\mathbf{D}}$ and the sparse code $\hat{\mathbf{s}}_t$ are updated when a new observation $\mathbf{y}_t^{\text{obs}}$ arrives at time $t$, which can be obtained by solving

$$\min_{\mathbf{D}\in\mathcal{D},\{\mathbf{s}_\tau\succeq\mathbf{0}\}} \sum_{\tau=1}^{t} \beta^{t-\tau}\left(\frac{1}{2}\|\mathbf{y}_\tau^{\text{obs}} - \mathbf{O}_\tau\mathbf{D}\mathbf{s}_\tau\|_2^2 + \lambda_1\|\mathbf{s}_\tau\|_1 + \frac{\lambda_2}{2}\|\mathbf{s}_\tau\|_2^2 + \frac{\lambda_L}{2}\mathbf{s}_\tau'\mathbf{D}'\mathbf{L}\mathbf{D}\mathbf{s}_\tau\right) \tag{12}$$

where $\beta \in (0, 1]$ is a forgetting factor that gradually diminishes the effect of past data. The spare coding is performed at each time $t$ with fixed $\hat{\mathbf{D}}_{t-1}$, which is given as

$$\hat{\mathbf{s}}_t = \arg\min_{\mathbf{s}\succeq\mathbf{0}} \frac{1}{2}\|\mathbf{y}_t^{\text{obs}} - \mathbf{O}_t\hat{\mathbf{D}}_{t-1}\mathbf{s}\|_2^2 + \lambda_1\|\mathbf{s}\|_1 + \frac{\lambda_2}{2}\|\mathbf{s}\|_2^2 + \frac{\lambda_L}{2}\mathbf{s}'\hat{\mathbf{D}}_{t-1}'\mathbf{L}\hat{\mathbf{D}}_{t-1}\mathbf{s}. \tag{13}$$

With fixed $\{\hat{\mathbf{s}}_\tau\}_{\tau=1}^t$, the dictionary can be updated in an online fashion. Define first the following quantities featuring recursive computations.

$$\mathbf{A}_t := \sum_{\tau=1}^{t} \beta^{t-\tau}\hat{\mathbf{s}}_\tau\hat{\mathbf{s}}_\tau' = \beta\mathbf{A}_{t-1} + \hat{\mathbf{s}}_t\hat{\mathbf{s}}_t' \tag{14}$$

$$\mathbf{A}_t^{(n)} := \sum_{\tau=1}^{t} \mathbb{1}_{\{n\in\mathcal{N}_\tau^{\text{obs}}\}}\beta^{t-\tau}\hat{\mathbf{s}}_\tau\hat{\mathbf{s}}_\tau' = \beta\mathbf{A}_{t-1}^{(n)} + \mathbb{1}_{\{n\in\mathcal{N}_t^{\text{obs}}\}}\hat{\mathbf{s}}_t\hat{\mathbf{s}}_t', \quad n \in \mathcal{N} \tag{15}$$

$$\mathbf{B}_t := \sum_{\tau=1}^{t} \beta^{t-\tau}\mathbf{O}_\tau'\mathbf{y}_\tau^{\text{obs}}\hat{\mathbf{s}}_\tau' = \beta\mathbf{B}_{t-1} + \mathbf{O}_t'\mathbf{y}_t^{\text{obs}}\hat{\mathbf{s}}_t' \tag{16}$$

$$\mathbf{V}_t^{i,j} := \sum_{\tau=1}^{t} \beta^{t-\tau}\hat{s}_{i,t}\hat{s}_{j,t}(\mathbf{O}_\tau'\mathbf{O}_\tau + \lambda_L\mathbf{L}) = \text{diag}([A_{ij,t}^{(1)}, \ldots, A_{ij,t}^{(N)}]) + \lambda_L A_{ij,t}\mathbf{L},\ i, j = 1, \ldots, M \tag{17}$$

where $\hat{s}_{i,t}$ denotes the $i$-th entry of $\hat{s}_t$. $A_{ij,t}^{(n)}$ and $A_{ij,t}$ represent the $(i, j)$-th entry of matrices $\mathbf{A}_t^{(n)}$ and $\mathbf{A}_t$, respectively. Let $\mathbf{b}_{i,t}$ denote $i$-th column of $\mathbf{B}_t$. Then, a column-wise BCD update of the dictionary $\hat{\mathbf{D}}_t$ is given as

$$\tilde{\mathbf{d}}_i = (\mathbf{V}_t^{i,i})^{-1}\left(\mathbf{b}_{i,t} - \sum_{j\neq i}\mathbf{V}_t^{i,j}\hat{\mathbf{d}}_{j,t}\right) \tag{18}$$

$$\hat{\mathbf{d}}_{i,t} = \frac{\tilde{\mathbf{d}}_i^+}{\max\{\|\tilde{\mathbf{d}}_i^+\|_2, 1\}}. \tag{19}$$

Under certain mild conditions, it can be shown that as $t \to \infty$, $\hat{\mathbf{D}}_t$ converges almost surely to the set of stationary points of the dictionary learning problem [16, Prop. 4]. The overall online approach is tabulated as Algorithm 2.

**Algorithm 2** Online dictionary learning

1: Initialize $\hat{\mathbf{D}}_0 = \mathbf{D}_0, \mathbf{A}_0 = \mathbf{0}, \mathbf{A}_0^{(n)} = \mathbf{0}, \forall n \in \mathcal{N}$, and $\mathbf{B}_0 = \mathbf{0}$
2: **for** $t = 1, 2, \ldots$ **do**
3: Obtain sparse code $\hat{\mathbf{s}}_t$ with fixed $\hat{\mathbf{D}}_{t-1}$ via (13)
4: Output prediction $\hat{\mathbf{p}}_t = \hat{\mathbf{D}}_{t-1}\hat{\mathbf{s}}_t$
5: Update matrices $\mathbf{A}_t$, $\mathbf{A}_t^{(n)}$, and $\mathbf{B}_t$ via (14)–(16)
6: Set $[\hat{\mathbf{d}}_{1,t}, \ldots, \hat{\mathbf{d}}_{M,t}] = \hat{\mathbf{D}}_{t-1}$
7: **repeat**
8: **for** $i = 1, \ldots, M$ **do**
9: Update $\hat{\mathbf{d}}_{i,t}$ as (17)–(19)
10: **end for**
11: **until** convergence
12: Set $\hat{\mathbf{D}}_t = [\hat{\mathbf{d}}_{1,t}, \ldots, \hat{\mathbf{d}}_{M,t}]$
13: **end for**

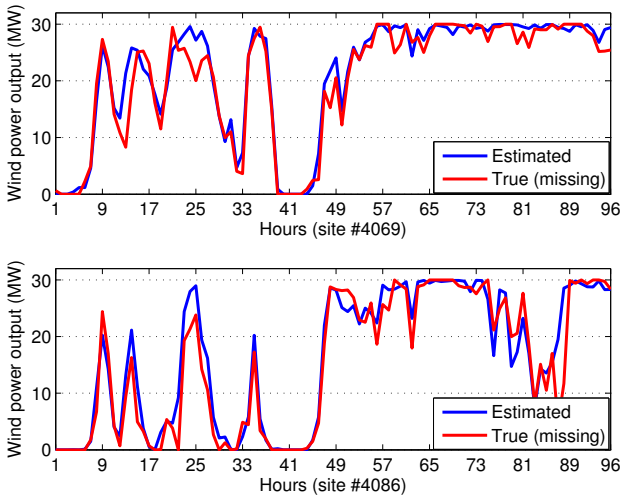


Fig. 2. Reconstruction of the missing wind power outputs of the sites #4069 and #4086 across 96 hours.

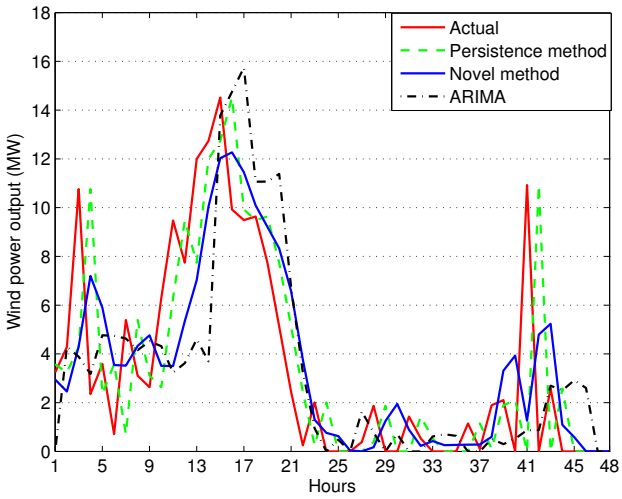


Fig. 3. Forecasting performance of the site \$4077 across 48 hours.

### C. Real-Time Wind Forecasting

Partially missing data can be easily recovered in the preceding procedure. However, in order to predict the future wind power outputs, temporal correlation has to be incorporated in the learning process. Here, the idea is to simply explore available information of the past measurements over a consecutive $T_c$ coherent slots [13]. Specifically, define an augmented measurement $\tilde{\mathbf{y}}_t^{\text{obs}}$, observation matrix $\tilde{\mathbf{O}}_t$, and Laplacian matrix $\tilde{\mathbf{L}}$ as

$$\tilde{\mathbf{y}}_t^{\text{obs}} := [\mathbf{y}_t^{\text{obs}\prime}, \ldots, \mathbf{y}_{t-T_c+1}^{\text{obs}}{}']' \tag{20}$$

$$\tilde{\mathbf{O}}_t := [\mathbf{O}_t', \ldots, \mathbf{O}_{t-T_c+1}']' \tag{21}$$

$$\tilde{\mathbf{L}} := \mathbf{I}_{T_c} \otimes \mathbf{L}. \tag{22}$$

A temporal dictionary $\hat{\mathbf{D}}_t$ can be first learned by replacing $\{\mathbf{y}_t^{\text{obs}}, \mathbf{O}_t, \mathbf{L}\}$ with $\{\tilde{\mathbf{y}}_t^{\text{obs}}, \tilde{\mathbf{O}}_t, \tilde{\mathbf{L}}\}$ in Algorithm 2. Upon defining a virtual observation $\tilde{\mathbf{y}}_{t+1}^{\text{obs,v}} := [\mathbf{y}_t^{\text{obs}\prime}, \ldots, \mathbf{y}_{t-T_c+2}^{\text{obs}}{}']'$, assuming the entire $\mathbf{y}_{t+1}^{\text{obs}}$ is missing. Then, compute the corresponding optimal sparse coefficient $\tilde{\mathbf{s}}_{t+1}^{\text{v}}$ by executing step 3 of Algorithm 2. The prediction $\hat{\mathbf{p}}_{t+1}$ can thus be obtained as

$$\hat{\mathbf{p}}_{t+1} = \hat{\mathbf{D}}_{[1:N,:],t}\tilde{\mathbf{s}}_{t+1}^{\text{v}} \tag{23}$$

where $\hat{\mathbf{D}}_{[1:N,:],t}$ is the first $N$ rows of $\hat{\mathbf{D}}_t$.

## IV. Numerical Tests

In this section, numerical results are presented to verify the performance of the novel inference approach. The forecasting performance was tested using the Western Wind Resources Dataset created by 3TIER with oversight and assistance from NREL [2]. Hourly wind power outputs were sampled across $N = 18$ neighboring wind farms that are located in the east of the city of Albuquerque, New Mexico. Each site has ten 3 MW Vestas V90 turbines, a total of 30 MW generation capacity. Wind speed of the region is often near the cut-out rate (25 m/s). Consequently, wind turbines cannot restart until the wind speed drops below a cut-back-in speed (20 m/s). Taking the possible turbine hysteresis effect into account, the "CorrectedScore" values in the dataset are accurate in the long-term, and hence were treated as the actual wind power outputs.

Let $i, j \in \mathcal{N}$ index the wind farms. The total 18 wind farms from the site #4069 ($i = 1$) to #4086 ($i = 18$) align in a straight line. To construct the Laplacian matrix $\mathbf{L}$ [cf. (3)], entries of the matrix $\mathbf{W}$ were simply chosen as $w_{ij} = 1$ for all $i \neq j, |i - j| \leq 2$, and 0 otherwise. The number of atoms in $\mathbf{D}$ is $M = 50$. For both the batch and online algorithms, The whole data of the year 2005 were used to train the initial dictionary $\mathbf{D}_0$. The prediction performance was tested across two evaluation periods in 2006.

Fig. 2 shows the performance of imputing the missing values via the batch algorithm. The values of $\lambda_1$, $\lambda_2$, $\lambda_L$ were set to 0.25, 0.002, and $2.5 \times 10^{-4}$, respectively. The measurements of the first and last sites are assumed to be missing across a consecutive period of 96 hours. It can be seen that the missing values are accurately recovered for the two sites.

The proposed online algorithm was used for forecasting the wind generation in the next hour. The relevant parameters were set as $\beta = 0.99$, $T_c = 3$, $\lambda_1 = 0.75$, $\lambda_2 = 0.95$, and $\lambda_L = 5 \times 10^{-5}$. Specifically, three methods were tested: (i) the proposed online dictionary learning approach; (ii) the persistent method, which simply repeats the last measurement; and (iii) the ARIMA model trained independently across different sites. Regarding the last one, the last 6-hour data were used to estimate the ARIMA model based on Akaike's information criterion (AIC) [4]. The functions `auto.arima` and `forecast` in the R package "forecast" [9] were used for the estimating the model, and forecasting the next-hour wind power outputs.

Fig. 3 depicts the forecasting performance of the site #4077 across an evaluation period of 48 hours. Clearly, the proposed novel approach outperforms both the persistence and ARIMA schemes. Note that by the definition of the persistence method,

TABLE I
FORECASTING ERRORS AVERAGED OVER 48 HOURS. THE UNIT OF RMSE AND MAE IS MW.

| | Novel | Persistence | ARIMA |
|---|---|---|---|
| RMSE | **2.6517** | 2.9631 | 3.5103 |
| MAE | **2.232** | 2.3458 | 2.8806 |

its curve is simply shifted from the actual's by one hour. It can be seen that at the peaks and valleys of the actual outputs, the novel approach deviates from the persistent one, yielding improved average forecast accuracy [cf. Table I]. It is worth stressing that as the standard benchmark of the short-term forecasting, the persistent method is actually hard to very beat [10]. Two forecasting errors were evaluated: the root mean-square error (RMSE) $\|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2/\sqrt{N}$, and the mean-absolute errors (MAE) $\|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_1/N$. Clearly, the proposed novel forecast attains the lowest averaged RMSE and MAE, as listed in Table I.

## V. CONCLUSION

A novel spatio-temporal learning approach has been developed for wind forecasting in this paper. Leveraging the known locations of correlated wind farms, a topology-cognizant dictionary of the wind power generation was first learned using historical data. Missing measurements of the wind power outputs can be readily interpolated using nonnegative sparse coding. Based on the virtual observations accounting for temporal correlations, real-time forecasts can be efficiently obtained via the online dictionary learning. The novel approach yields lower forecasting errors than those of existing alternatives.

A number of interesting research directions open up towards extending the proposed framework. Atmospheric factors including wind speed and temperature should be incorporated as exogenous variables in the learning process to further enhance the prediction accuracy. Kernel dictionary learning can offer improved predictors with nonlinear structures.

## REFERENCES


[1] "20% wind energy by 2030: Increasing wind energy's contribution to U.S. electricity supply," July 2008, [Online]. Available: http://www1.eere.energy.gov/wind/pdfs/41869.pdf.

[2] 3TIER, "Western wind resources dataset," [Online]. Available: http://wind.nrel.gov/Web_nrel/.

[3] T. Barbounis, J. Theocharis, M. Alexiadis, and P. Dokopoulos, "Long-term wind speed and power forecasting using local recurrent neural network models," *IEEE Trans. Energy Convers.*, vol. 21, no. 1, pp. 273–284, Mar. 2006.

[4] P. Brockwell and R. Davis, *Time Series: Theory and Methods*, 2nd ed. New York, NY: Springer, 1991.

[5] A. Foley, P. Leahy, A. Marvuglia, and E. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renewable Energy*, vol. 37, pp. 1–8, Jan. 2012.

[6] P. Forero, K. Rajawat, and G. B. Giannakis, "Prediction of partially observed dynamical processes over networks via dictionary learning," *IEEE Trans. Signal Process.*, vol. 62, no. 13, pp. 3305–3320, Jul. 2014.

[7] GWEC, "Global wind statistics 2013," May 2014, [Online]. Available: http://www.gwec.net/wp-content/uploads/2014/02/GWEC-PRstats-2013_EN.pdf.

[8] A. Heemink and A. Segers, "Modeling and prediction of environmental data in space and time using Kalman filtering," *Stoch. Environ. Res. Risk Assess.*, vol. 16, no. 3, pp. 225–240, 2002.

[9] R. Hyndman, "forecast: Forecasting functions for time series and linear models," Feb. 2014. [Online]. Available: http://cran.r-project.org/web/packages/forecast/index.html

[10] G. Kariniotakis, P. Pinson, N. Siebert, G. Giebel, and R. Barthelmie, "The state of the art in short-term prediction of wind power - from an offshore perspective," in *Proc. of SeaTechWeek*, Brest, France, Oct. 2004, pp. 20–21.

[11] G. Kariniotakis, G. Stavrakakis, and E. Nogaret, "Wind power forecasting using advanced neural networks models," *IEEE Trans. Energy Convers.*, vol. 11, no. 4, pp. 762–767, Dec. 1996.

[12] R. Kavasseri and K. Seetharaman, "Day-ahead wind speed forecasting using $f$-ARIMA models," *Renewable Energy*, vol. 34, pp. 1388–1393, 2009.

[13] S.-J. Kim and G. B. Giannakis, "Cognitive radio spectrum prediction using dictionary learning," in *Proc. of Global Commun. Conf.*, Atlanta, GA, Dec. 2013, pp. 3206–3211.

[14] O. Kramer and F. Gieseke, "Short-term wind energy forecasting using support vector regression," in *Proc. of Intl. Conf. on Soft Computing Models in Industrial and Environ. Appl.*, Salamanca, Spain, Apr. 2011, pp. 271–280.

[15] L. Ma, S. Luan, C. Jiang, H. Liu, and Y. Zhang, "A review on the forecasting of wind speed and generated power," *Renew. Sust. Energy Rev.*, vol. 13, pp. 915–920, 2009.

[16] J. Mairal, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

[17] M. Milligan, M. Schwartz, and Y. Wan, "Statistical wind power forecasting models: Results for US wind farms," National Renewable Energy Laboratory, Tech. Rep., May 2003, [Online]. Available: http://www.nrel.gov/docs/fy03osti/33956.pdf.

[18] P. Pinson, "Estimation of the uncertainty in wind power forecasting," Ph.D. dissertation, Ecole des Mines de Paris, Mar. 2006.

[19] P. Pinson, C. Chevallier, and G. Kariniotakis, "Trading wind generation from short-term probabilistic forecasts of wind power," *IEEE Trans. on Power Syst.*, vol. 22, no. 3, pp. 1148–1156, Aug. 2007.

[20] P. Pinson and H. Madsen, "Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models," *J. of Forecasting*, vol. 31, no. 4, pp. 281–313, Jul. 2012.

[21] J. Poloczek, N. Treiber, and O. Kramer, "KNN regression as geo-imputation method for spatio-temporal wind data," in *Proc. of Intl. Joint Conf. SOCO'14-CISIS'14-ICEUTE'14*, Bilbao, Spain, June 2014, pp. 185–193.

[22] G. Sideratos and N. Hatziargyriou, "An advanced statistical method for wind power forecasting," *IEEE Trans. on Power Syst.*, vol. 22, no. 1, pp. 258–265, Feb. 2007.

[23] J. Zhou, J. Shi, and G. Li, "Fine tuning support vector machines for short-term wind speed forecasting," *Energy Convers. Manage.*, vol. 52, no. 4, pp. 1990–1998, Apr. 2011.

[24] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Statist. Soc. B*, vol. 67, pp. 301–320, 2005.