Contents lists available at ScienceDirect

International Journal of Forecasting

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

Spatiotemporal wind forecasting by learning a hierarchically sparse inverse covariance matrix using wind directions

Yin Liu [a], Sam Davanloo Tajbakhsh [a,*], Antonio J. Conejo [a,b]

[a] *Integrated Systems Engineering, The Ohio State University, Columbus, OH, USA*

[b] *Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA*

ARTICLE INFO

Keywords:
Spatio-temporal prediction
Hierarchical sparsity structure
Covariance selection
Nonsmooth optimization
Proximal map

ABSTRACT

Given the advances in online data acquisition systems, statistical learning models are increasingly used to forecast wind speed. In electricity markets, wind farm production forecasts are needed for the day-ahead, intra-day, and real-time markets. In this work, we use a spatiotemporal model that leverages wind dynamics to forecast wind speed. Using a priori knowledge of the wind direction, we propose a maximum likelihood estimate of the inverse covariance matrix regularized with a hierarchical sparsity-inducing penalty. The resulting inverse covariance estimate not only exhibits the benefits of a sparse estimator, but also enables meaningful sparse structures by considering wind direction. A proximal method is used to solve the underlying optimization problem. The proposed methodology is used to forecast six-hour-ahead wind speeds in 20-minute time intervals for a case study in Texas. We compare our method with a number of other statistical methods. Prediction performance measures and the Diebold–Mariano test show the potential of the proposed method, specifically when reasonably accurate estimates of the wind directions are available.

1. Introduction

As weather-dependent stochastic electricity power sources, such as wind- or solar-based production units, increasingly penetrate the generation mix, precise forecasting techniques for the production of these sources are needed. There is a range of methods for wind speed and power prediction. Although predicting wind speed and generated power are different in nature, with the goal of predicting power generation, the methods can be categorized as: physics-based mathematical methods, data-driven statistical methods, and hybrid methods that combine techniques from the first two categories.

There are a number of numerical weather prediction (NWP) methods that use physics-based mathematical models that rely on the current state of the ocean and atmosphere to predict the future speed and direction of the wind. For detailed studies on NWP methods, see Cassola and Burlando (2012), Lange and Focken (2006), Richardson (2007) and the references therein.

Advances in remote sensing devices have resulted in an abundance of weather-related data that are used as inputs to data-driven prediction methods. The easiest data-driven method is the *persistence* method, which is based on the assumption that wind does not change in the near future. The persistence method simply sets the predictions equal to the most current observed value for any time horizon. Moreover, given the autocorrelation in wind data, time series methods—e.g., autoregressive integrated moving average (ARIMA) methods (Box, Jenkins, Reinsel, & Ljung, 2015)—are common for wind speed or power output forecasting based on historical data (Erdem & Shi, 2011; Kavasseri & Seetharaman, 2009; Sanchez, 2006; Torres, García, De Blas, & De Francisco, 2005). Furthermore, advances in machine learning algorithms based on high-dimensional modeling of wind or wind power

[*] Corresponding author.
E-mail addresses: liu.6630@osu.edu (Y. Liu), davanloo.1@osu.edu (S. Davanloo Tajbakhsh), conejo.1@osu.edu (A.J. Conejo).

https://doi.org/10.1016/j.ijforecast.2020.09.009

data indexed by time and space have found applications in wind speed and power output prediction—see e.g., Chitsaz, Amjady, and Zareipour (2015), Damousis, Alexiadis, Theocharis, and Dokopoulos (2004), El-Fouly, El-Saadany, Salama, El-Fouly, El-Saadany, and Salama (2006), Kusiak, Zheng, and Song (2009), Landry, Erlinger, Patschke, and Varrichio (2016), Mandic, Javidi, Goh, Kuh, and Aihara (2009), Mangalova and Agafonov (2014), Mangalova and Shesterneva (2016), Mohandes, Halawani, Rehman, and Hussain (2004), Salcedo-Sanz, Ortiz-Garcı, Pérez-Bellido, Portilla-Figueras, and Prieto (2011), Zhang and Wang (2016), Zhou, Shi, and Li (2011). Spatiotemporal methods that are related to this research are discussed in more detail below.

Besides the two conventional categories, there are also hybrid methods that combine the outputs of NWP with data-driven methods for better prediction performance—see e.g., Chen, Qian, Nabney, and Meng (2014), Jursa and Rohrig (2008), Negnevitsky, Johnson, and Santoso (2007).

Considering a collection of wind farms, spatiotemporal information can be incorporated into a data-driven model to improve the prediction performance. For instance, Tastu, Pinson, Trombe, and Madsen (2014) include the spatial correlation information into a vector autoregressive model for reliable and economic power systems operations in a smart grid. Morales, Minguez, and Conejo (2010) propose a multivariate time series model that quantifies spatial correlations between the sites. Alexiadis, Dokopoulos, and Sahsamanoglou (1999), Barbounis and Theocharis (2007) use spatial information of neighboring sites to train neural network models to predict wind speed. Gneiting, Larson, Westrick, Genton, and Aldrich (2006) build regime-switching space–time models considering all salient features of wind speed including spatial correlation—see also (Hering & Genton, 2010; Zhu, Genton, Gu, & Xie, 2014). Khalid and Savkin (2012) predict wind speed and power from the observed data at nearby sites and use an NWP model for supplementary adjustments. Xie, Gu, Zhu, and Genton (2014) combine the spatiotemporal correlation of the wind speed and direction with a statistical model—see also (Dowell, Weiss, Hill, & Infield, 2014; Messner & Pinson, 2018; Tastu et al., 2014; Zhao, Ye, Pinson, Tang, & Lu, 2018). Gaussian processes have also been implemented for wind and power prediction in a number of studies. For instance, Yu, Chen, Mori, and Rashid (2013) use a localized Gaussian process model for long-term wind speed prediction based on a large data set.

The sparsification of the spatiotemporal model has also been deployed to enhance predictions by reducing variance. For instance, Dowell and Pinson (2016) represent spatial information by a sparse vector autoregressive process. Zhao et al. (2018) build a sparsity-controlled vector autoregressive model to include spatiotemporal dependence. Tascikaraoglu, Sanandaji, Poolla, and Varaiya (2016) exploit the sparsity of interconnections using wavelet transforms in a spatiotemporal model—see also (He, Yang, Zhang, & Vittal, 2014). Finally, Wytock and Kolter (2013) sparsify a high-dimensional conditional Gaussian process model for spatiotemporal wind power prediction and extend it to non-Gaussian data using a copula transform.

In these works, sparsification is mainly performed for variance reduction or to attain computational gain. Using the ℓ_1-norm to induce sparsity, the sparsity pattern of the optimal solution is determined by data. In this work, however, we want the sparsity pattern of the optimal solution (i.e., the elements of the inverse covariance matrix) to belong to a predetermined set of hierarchical structures known a priori from wind direction. Hence, to estimate the inverse covariance matrix, we incorporate a specific penalty function that induces such a hierarchical structure. Therefore, the optimal solution of the proposed learning problem is more interpretable and results in better prediction performance. Specifically, the paper discusses a new penalty function and how to solve the underlying learning optimization problem *efficiently*. The main idea of this work is described in Sections 3.1 and 3.2. The learning problem and a first-order optimization algorithm to solve it are discussed in Sections 3.3–3.5. In Section 4, we apply the proposed methodology to a case study in Texas. Finally, Section 5 provides our concluding remarks.

1.1. Notation

The wind speed at spatial location $s \in \mathcal{S}$ and time t is denoted by $w_t(s)$, where $s \in \mathcal{S} = \{1, \ldots, m\}$ is the index set of m wind sites. Furthermore, $\mathbf{w}_t^{t'}(\mathcal{S}) \in \mathbb{R}^{m(t'-t+1)}$ denotes a collection of wind speeds from time t to t' for all spatial sites, written as $(\mathbf{w}_t(\mathcal{S})^\top, \mathbf{w}_{t+1}(\mathcal{S})^\top, \ldots, \mathbf{w}_{t'}(\mathcal{S})^\top)^\top$, where $\mathbf{w}_t(\mathcal{S}) \in \mathbb{R}^m$ denotes the wind speeds at time t at all $\mathcal{S}$ spatial sites. Transforming the non-Gaussian wind data to a Gaussian field (see the transformation in 2.1), $y_t(s)$, $\mathbf{y}_t^{t'}(\mathcal{S}) \in \mathbb{R}^{m(t'-t+1)}$, and $\mathbf{y}_t(\mathcal{S}) \in \mathbb{R}^m$ are defined similarly. Here, $\mathbf{1}_w$ denotes the column vector of all ones of size w. Similarly, $\mathbf{0}_w$ denotes the column vector of all zeros of size w, and $\mathbb{S}^d_{++}$ denotes the cone of $d \times d$ positive definite matrices. The matrix inner product is defined as $\langle A, B\rangle = \mathbf{Tr}(AB^\top)$. Let $\mathcal{G}$ be a set, and then $|\mathcal{G}|$ denotes the set cardinality. Let $g \subseteq \mathcal{G}$ be an element of the set, and then g^c denotes its complement. Finally, vec$(\cdot)$ takes a matrix column-by-column and stacks the columns as a long vector.

2. Some preliminaries and contributions of the work

2.1. Transformations between Gaussian and non-Gaussian fields

Since the predictions are performed in the Gaussian field based on the conditional Gaussian distribution, there is a need to transform the non-Gaussian wind data to the Gaussian field. Similarly, when predictions are obtained, there is a need to transform them back to the non-Gaussian field. In the forward transformation, given a wind speed $w_t(s)$, the empirical marginal distribution of wind at each spatial site in some time window is used to find the corresponding value of the cumulative distribution function, i.e., $\tilde{F}_s(w_t(s))$. The transformed data in the Gaussian field is then obtained by applying the inverse cumulative distribution function of the standard normal distribution, i.e., $y_t(s) = F^{-1}(\tilde{F}_s(w_t(s)))$. Similarly,

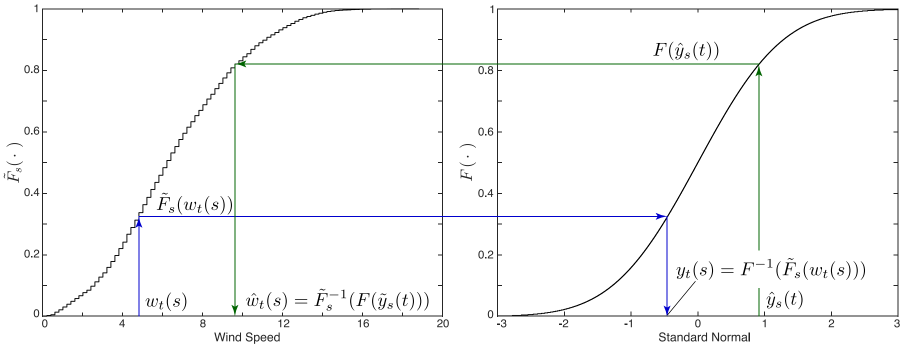


Fig. 1. True and normalized distributions of wind speed.

given a prediction in the Gaussian field $\hat{y}_t(s)$, the transformed value in the non-Gaussian field is obtained as $\hat{w}_t(s) = \tilde{F}_s^{-1}(F(\hat{y}_t(s)))$. These transformations are shown in Fig. 1—see also (Möller, Lenkoski, & Thorarinsdottir, 2013; Morales et al., 2010).

2.2. Gaussian random field (GRF) model on a discrete set

A Gaussian random field (GRF) is an indexed stochastic process such that for a fixed set of indices it results in a multivariate Gaussian distribution. Let $y_t(s) \in \mathbb{R}$ be an observation from a GRF at time t and location $s \in \mathcal{S}$, where $\mathcal{S}$ is a discrete set of spatial locations with $|\mathcal{S}| = m$ (i.e., the index $(s,t) \in \mathcal{S} \times \mathbb{R}_+$). Let the data (after the transformation to the Gaussian field) up to current time point t_c be $\mathbf{y}_1^{t_c}(\mathcal{S}) \in \mathbb{R}^{mt_c}$. To make predictions for w_f time points in the future over all sites $\mathcal{S}$, we use the data of all sites for the w_p previous time points to train the model, i.e., the training data at time t_c is $\mathbf{y}_{t_c-w_p+1}^{t_c}(\mathcal{S}) \in \mathbb{R}^{mw_p}$. Since any countable collection of observations from a GRF follows a multivariate normal distribution, the joint distribution of $(\mathbf{y}_{t_c-w_p+1}^{t_c}(\mathcal{S}), \mathbf{y}_{t_c+1}^{t_c+w_f}(\mathcal{S}))^\top$ is given by

$$\begin{pmatrix} \mathbf{y}_{t_c-w_p+1}^{t_c}(\mathcal{S}) \\ \mathbf{y}_{t_c+1}^{t_c+w_f}(\mathcal{S}) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}_{m(w_p+w_f)}, \Sigma \triangleq \begin{bmatrix} \Sigma_{pp} & \Sigma_{pf} \\ \Sigma_{pf}^\top & \Sigma_{ff} \end{bmatrix}\right). \tag{1}$$

where $\Sigma_{pp} \in \mathbb{S}_{++}^{mw_p}$ and $\Sigma_{ff} \in \mathbb{S}_{++}^{mw_f}$ are the covariance matrices for the past w_p and future w_f observations for all sites, respectively, and $\Sigma_{pf} \in \mathbb{R}^{mw_p \times mw_f}$ is the cross-covariance matrix between the past w_p and future w_f observations for all sites. Furthermore, the conditional probability distribution $p(\mathbf{y}_{t_c+1}^{t_c+w_f}(\mathcal{S})|\mathbf{y}_{t_c-w_p+1}^{t_c}(\mathcal{S}))$ is given by

$$\begin{aligned} &\mathbf{y}_{t_c+1}^{t_c+w_f}(\mathcal{S})|\mathbf{y}_{t_c-w_p+1}^{t_c}(\mathcal{S}) \\ &\quad \sim \mathcal{N}\left(\Sigma_{pf}^\top \Sigma_{pp}^{-1} \mathbf{y}_{t_c-w_p+1}^{t_c}(\mathcal{S}), \Sigma_{ff} - \Sigma_{pf}^\top \Sigma_{pp}^{-1} \Sigma_{pf}\right). \end{aligned} \tag{2}$$

The mean of the conditional distribution (2) is used to predict the wind speed for future w_f time points at all of the sites in $\mathcal{S}$. If the mean of the joint distribution (1) is *not zero*, then the conditional distribution is

$$\begin{aligned} &\mathbf{y}_{t_c+1}^{t_c+w_f}(\mathcal{S})|\mathbf{y}_{t_c-w_p+1}^{t_c}(\mathcal{S}) \\ &\quad \sim \mathcal{N}\Big(\boldsymbol{\mu}(\mathcal{S}, \mathbf{x}_{t_c+1}^{t_c+w_f}) \\ &\qquad + \Sigma_{pf}^\top \Sigma_{pp}^{-1} \left(\mathbf{y}_{t_c-w_p+1}^{t_c}(\mathcal{S}) - \boldsymbol{\mu}(\mathcal{S}, \mathbf{x}_{t_c-w_p+1}^{t_c})\right), \\ &\qquad \Sigma_{ff} - \Sigma_{pf}^\top \Sigma_{pp}^{-1} \Sigma_{pf}\Big), \end{aligned} \tag{3}$$

where $\boldsymbol{\mu}(\mathcal{S}, \mathbf{x}_{t_1}^{t_2})$ is the mean of $\mathbf{y}_{t_1}^{t_2}(\mathcal{S})$ from t_1 to t_2 over spatial locations in $\mathcal{S}$ given a set of r covariates $\mathbf{x}_{t_1}^{t_2} \in \mathbb{R}^{r(t_2-t_1+1)}$. A nonzero mean structure allows for the incorporation of covariates $\mathbf{x}$ into the prediction model, potentially increasing the prediction power of the model.

2.3. Contribution of the work

The predictive model in the proposed method is based on the conditional distributions (2) and (3), which require estimating the covariance matrix Σ in (1). Our work looks into estimating Σ through the sparse estimation of its inverse Σ^{-1}, known as the precision matrix. Such an approach reduces the number of parameters and, hence, reduces the prediction variance. Sparse estimation of the inverse covariance matrix is generally performed by introducing a convex ℓ_1-norm regularizer to the negative log-likelihood loss function, which results in the graphical lasso estimator (6) (called G-L in this work) (Friedman, Hastie, & Tibshirani, 2008). The ℓ_1-norm does not enforce particular sparsity structures to the Σ^{-1} estimate. For wind data, however, we usually have a priori knowledge of wind direction (see Fig. 4(a)), which if incorporated, allows for a better estimate of the inverse covariance matrix. By incorporating a hierarchical sparsity-inducing regularizer, we introduce another estimator for Σ^{-1}, namely GLOG-L, that allows a priori knowledge of wind direction to play a role. We then provide a computationally efficient algorithm to calculate the GLOG-L estimate, and demonstrate its performance in a wind prediction case study.

3. Methodology

To use the conditional distribution (2) for predictions, we need to estimate its parameters Σ. When GRF models are used to predict over a continuous index set, the covariance matrix is generally constructed based on a parametric covariance function (Rasmussen & Williams, 2006). Hence, there is a need to estimate the parameters

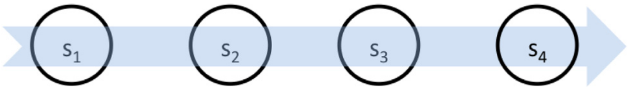


Fig. 2. Direction of the wind over four spatial locations $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$.

of the covariance function, and this requires solving a non-convex optimization problem to minimize the negative of a log-likelihood. Poor estimation of these covariance function parameters generally results in poor prediction performance, which is aggravated in extrapolation, e.g., forecasting over time (Davanloo Tajbakhsh, Serhat Aybat, & Del Castillo, 2014). In this work, however, given the fact that the covariance matrix is estimated over a fixed discrete set, we estimate the inverse of the covariance matrix by solving a convex optimization problem with theoretical guarantees for convergence.

To estimate $\Sigma \in \mathbb{S}_{++}^{m(w_p+w_f)}$, we first construct the data matrix $Y_{t_c-w_f-w_p-(N-1)\delta+1}^{t_c}(\mathcal{S}) \in \mathbb{R}^{N\times m(w_p+w_f)}$ as

$$Y_{t_c-w_f-w_p-(N-1)\delta+1}^{t_c}(\mathcal{S}) = \begin{pmatrix} \mathbf{y}_{t_c-w_f-w_p+1}^{t_c-w_f}(\mathcal{S})^\top & \mathbf{y}_{t_c-w_f+1}^{t_c}(\mathcal{S})^\top \\ \mathbf{y}_{t_c-w_f-w_p-\delta+1}^{t_c-w_f-\delta}(\mathcal{S})^\top & \mathbf{y}_{t_c-w_f-\delta+1}^{t_c-\delta}(\mathcal{S})^\top \\ \mathbf{y}_{t_c-w_f-w_p-2\delta+1}^{t_c-w_f-2\delta}(\mathcal{S})^\top & \mathbf{y}_{t_c-w_f-2\delta+1}^{t_c-2\delta}(\mathcal{S})^\top \\ \vdots & \vdots \\ \mathbf{y}_{t_c-w_f-w_p-(N-1)\delta+1}^{t_c-w_f-(N-1)\delta}(\mathcal{S})^\top & \mathbf{y}_{t_c-w_f-(N-1)\delta+1}^{t_c-(N-1)\delta}(\mathcal{S})^\top \end{pmatrix}, \quad (4)$$

where $w_p \geq w_f$, $\delta \geq 1$ is a parameter that determines the time shift from one row of Y to the next one. Given $Y_{t_c-w_f-w_p-(N-1)\delta+1}^{t_c}(\mathcal{S})$, the sample covariance is defined as

$$\bar{S} = \frac{1}{N}\sum_{i=1}^{N} Y_{t_c-w_f-w_p-(N-1)\delta+1}^{t_c}(\mathcal{S})_{i.}^{\ \top} Y_{t_c-w_f-w_p-(N-1)\delta+1}^{t_c}(\mathcal{S})_{i.} \quad (5)$$

where $Y_{t_c-w_f-w_p-(N-1)\delta+1}^{t_c}(\mathcal{S})_{i.}$ denotes the ith row of the $Y_{t_c-w_f-w_p-(N-1)\delta+1}^{t_c}(\mathcal{S})$ matrix. Given $\bar{S}$, the maximum likelihood (ML) estimate of the inverse covariance matrix penalized with the ℓ_1-norm requires solving

$$\hat{X}_{\text{G-L}} = \operatorname*{argmin}_{X\succeq 0} \left\langle X, \bar{S}\right\rangle - \log\det(X) + \lambda\|X\|_1 \quad (6)$$

which is the well-known G-L estimate (Friedman et al., 2008). The ℓ_1-norm induces sparsity that exists in the *inverse* of the covariance matrix due to potential conditional independence between the variables—see Davanloo Tajbakhsh et al. (2014). Note that $\lambda > 0$ is the sparsity tuning parameter. Wytock and Kolter (2013) use the estimator (6) to build their prediction model for wind power forecasting. In this paper, however, we use a *hierarchical sparsity-inducing penalty* as opposed to the unstructured sparsity-inducing penalty ℓ_1-norm. This is motivated by the wind forecasting application, as discussed below.

3.1. Hierarchical sparsity structures

In this section, we first discuss the idea behind the hierarchical sparsity structure for estimating Σ^{-1} for wind forecasting. Subsequently, we discuss the hierarchical sparsity structures and the penalty functions that generate such a structure.

Consider the four locations in the set $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$ depicted in Fig. 2, and assume that we have a priori knowledge that the wind blows from s_1 to s_2 to s_3 to s_4, as shown in Fig. 2. Let $\Sigma_t^{t'}(s_1, s_3)$ denote the covariance between the wind speeds at time t at location s_1 and time t' at location s_3. It is well known that each element of the *inverse* covariance matrix is proportional to the conditional covariance between the two points given all other points, i.e., $\Sigma_{ij}^{-1} \propto \text{cov}(y_i, y_j|y_k, k \neq i, j)$—see Davanloo Tajbakhsh et al. (2014), Whittaker (2009). Hence, ${\Sigma^{-1}}_t^{t'}(s_1, s_3)$ is proportional to the conditional covariance between the wind speed at time t at location s_1 and time t' at location s_3 given the wind speed at other time points and locations. The a priori knowledge about the direction of the wind generates some potential sparsity in the inverse covariance matrix. For instance, from Fig. 2, assuming $t' > t$, we get

$$\begin{aligned} &\text{if } {\Sigma^{-1}}_t^{t'}(s_1, s_2) = 0 \quad \Rightarrow \quad {\Sigma^{-1}}_t^{t'}(s_1, s_3) = 0, \quad \forall t, t', t' > t \\ &\text{if } {\Sigma^{-1}}_t^{t'}(s_1, s_3) = 0 \quad \Rightarrow \quad {\Sigma^{-1}}_t^{t'}(s_1, s_4) = 0, \quad \forall t, t', t' > t \\ &\text{if } {\Sigma^{-1}}_t^{t'}(s_2, s_3) = 0 \quad \Rightarrow \quad {\Sigma^{-1}}_t^{t'}(s_2, s_4) = 0, \quad \forall t, t', t' > t \end{aligned}$$

Put simply, the first statement means that if wind cannot make it from s_1 to s_2 in time $t' - t$, then it cannot make it from s_1 to s_3 in time $t' - t$. That is, $y_{t'}(s_3)$ is independent of $y_t(s_1)$ given $y_{t'}(s_2)$. Hence, assuming a priori knowledge regarding the direction of the wind, *the sparsity structure of the estimated inverse covariance matrix cannot solely be determined by data*. Hence, the ℓ_1-norm penalty is replaced by another penalty that respects the underlying hierarchical sparsity structures provided by the a priori knowledge pertaining to the wind direction. This knowledge can be used to achieve a better estimate of the covariance matrix, which may result in better prediction performance.

3.2. Latent overlapping group (LOG) lasso regularizer

First proposed by Jacob, Obozinski, and Vert (2009), the latent overlapping group lasso (LOG) penalty induces hierarchical sparsity following a directed acyclic graph (DAG) that represents the hierarchical sparsity structure of the variables. The DAG corresponding to the example discussed above is illustrated in Fig. 3.

Each arrow in the hierarchy above generates a parent–child relationship. Given this parent–child relationship, if the parent variable is zero, then the child variable should also be zero. Similarly, if the child variable is nonzero then the parent variable should also be nonzero. To simplify the notation, we denote the five variables in the above example as $\boldsymbol{\gamma} \in \mathbb{R}^5$.

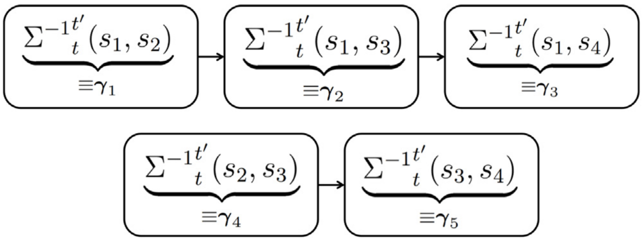

$\underbrace{\Sigma^{-1}{}_t^{t'}(s_1, s_2)}_{\equiv\gamma_1}$
$\underbrace{\Sigma^{-1}{}_t^{t'}(s_1, s_3)}_{\equiv\gamma_2}$
$\underbrace{\Sigma^{-1}{}_t^{t'}(s_1, s_4)}_{\equiv\gamma_3}$
$\underbrace{\Sigma^{-1}{}_t^{t'}(s_2, s_3)}_{\equiv\gamma_4}$
$\underbrace{\Sigma^{-1}{}_t^{t'}(s_3, s_4)}_{\equiv\gamma_5}$

Fig. 3. Directed acyclic graph (DAG) for the example in Fig. 2.

The LOG penalty defines an *ascending* grouping of variables based on a given DAG $\mathcal{D}$. The grouping of the variables is such that each node and all of its ancestors should be included as a subset into the set. For instance, for the graph $\mathcal{D}$ in Fig. 3, the set that contains all of the groups is $\mathcal{G} = \{\{1\}, \{1, 2\}, \{1, 2, 3\}, \{4\}, \{4, 5\}\}$, where $\mathcal{G}$ contains each node and all of its ancestors as a subset. Then, given the grouping of variables in the set $\mathcal{G}$, the LOG penalty is defined as

$$\Omega_{\mathrm{LOG}}(\boldsymbol{\gamma}) = \inf_{\boldsymbol{\nu}^{(g)},\ g\in\mathcal{G}} \left\{ \sum_{g\in\mathcal{G}} w_g \|\boldsymbol{\nu}^{(g)}\|_2 \quad \text{s.t.} \sum_{g\in\mathcal{G}} \boldsymbol{\nu}^{(g)} = \boldsymbol{\gamma},\ \boldsymbol{\nu}^{(g)}_{g^c} = 0 \right\} \tag{7}$$

where g denotes the groups inside the set $\mathcal{G}$, $\boldsymbol{\nu}^{(g)} \in \mathbb{R}^d$ denotes vectors indexed by g (d is equal to 5 in the above example), and $w_g > 0$ denotes weights. From the definition of $\Omega_{\mathrm{LOG}}(\boldsymbol{\gamma})$, $\boldsymbol{\gamma}$ is written as a summation of $|\mathcal{G}|$ latent variables $\boldsymbol{\nu}^{(g)}$. The ℓ_2-norms induce block sparsity; i.e., they seek to make the elements of a block simultaneously zero or nonzero. Hence, the penalty function indeed seeks to construct $\boldsymbol{\gamma}$ with only a few nonzero latent variables $\boldsymbol{\nu}^{(g)}$. The sparsity structure of latent variables based on their groups are such that the sparsity structure of their sum respects the hierarchy of the graph. For more details, see Jacob et al. (2009), Yan, Bien, et al. (2017).

3.3. GLOG-L estimate of the inverse covariance matrix

Given prior knowledge of the wind direction, a sparsity hierarchy for the elements of the inverse covariance matrix is built as a DAG, $\mathcal{D}$ (e.g., similar to Fig. 3). The set $\mathcal{G}$ that contains the groups can then be constructed from $\mathcal{D}$. Finally, the $\Omega_{\mathrm{LOG}}(\cdot)$ penalty is formed based on $\mathcal{G}$. Given the sparsity-inducing penalty $\Omega_{\mathrm{LOG}}(\cdot)$, we propose solving the following convex optimization problem for the inverse covariance matrix

$$\hat{X}_{\text{GLOG-L}} = \underset{X\succeq 0}{\operatorname{argmin}} \left\langle X, \bar{S} \right\rangle - \log\det(X) + \lambda\Omega_{\mathrm{LOG}}(X), \tag{8}$$

where $\bar{S}$ is the sample covariance matrix defined in (5). We call the solution to (8) the GLOG-L estimate of the inverse covariance matrix. We solve (8) using the alternating direction method of multipliers (ADMM) (Boyd, Parikh, Chu, Peleato, & Eckstein, 2011; Eckstein & Bertsekas, 1992; Gabay & Mercier, 1976; Glowinski & Marroco, 1975).

3.4. Finding the GLOG-L estimate

Problem (8) is first written as a problem with an equality constraint in two blocks as

$$\min_{X\succeq 0}\left\{ \left\langle X^1, \bar{S} \right\rangle - \log\det(X^1) + \lambda\Omega_{\mathrm{LOG}}(X^2),\ \text{ s.t. } X^1 = X^2 \right\}$$

where X^1 is the block of variables for the smooth part, and X^2 is the block of variables for the nonsmooth part. Dualizing the equality constraint, the *augmented* Lagrangian function is

$$\begin{aligned} L_\rho(X^1, X^2, W) = {} & \left\langle X^1, \bar{S} \right\rangle - \log\det(X^1) + \lambda\Omega_{\mathrm{LOG}}(X^2) \\ & + \left\langle X^1 - X^2, W \right\rangle + \frac{\rho}{2}\|X^1 - X^2\|_F^2, \end{aligned}$$

where $W \in \mathbb{R}^{(w_p+w_f)\times(w_p+w_f)}$ is the matrix of dual variables, and $\rho > 0$ is a parameter. Taking $U = W/\rho$ and following a Gauss–Seidel update, the ADMM subproblems can then be written as

$$\begin{aligned} X^{1,k+1} &= \underset{X\succeq 0}{\operatorname{argmin}} \left\langle X^1, \bar{S} \right\rangle - \log\det(X^1) \\ &\quad + \frac{\rho}{2}\|X^1 - X^{2,k} + U^k\|_F^2, \qquad \text{(9a)} \\ X^{2,k+1} &= \underset{X^2}{\operatorname{argmin}}\, \lambda\Omega_{\mathrm{LOG}}(\mathrm{vec}(X^2)) \\ &\quad + \frac{\rho}{2}\|\mathrm{vec}(X^2) - \mathrm{vec}(X^{1,k+1} + U^k)\|_2^2, \qquad \text{(9b)} \\ U^{k+1} &= U^k + X^{1,k+1} - X^{2,k+1}. \qquad \text{(9c)} \end{aligned}$$

The solution to subproblem (9a) is similar to that of the graphical lasso. In Appendix A, the solution to subproblem (9a) is discussed; for more detailed discussion, we refer the reader to Section 6.5 of Boyd et al. (2011). The solution to subproblem (9b) is equivalent to finding the proximal mapping of the $\lambda\Omega_{\mathrm{LOG}}(\cdot)$ penalty, i.e., $\mathbf{prox}_{(\lambda/\rho)\Omega_{\mathrm{LOG}}}(X^{1,k+1} + U^k)$, where the proximal map of a generic function $f(\cdot)$ at Q is defined as

$$\mathbf{prox}_{\lambda f}(V) \triangleq \underset{X}{\operatorname{argmin}}\ f(X) + \frac{1}{2}\|X - Q\|_F^2. \tag{10}$$

In Section 3.5 below, we propose an algorithm to solve the proximal map of the $\Omega_{\mathrm{LOG}}(\cdot)$ penalty.

3.5. Finding the proximal map of the $\Omega_{\mathrm{LOG}}(\cdot)$ penalty

Evaluating the proximal mapping of the $\Omega_{\mathrm{LOG}}(\cdot)$ penalty is not straightforward. Recall that inducing the hierarchical sparsity structure is performed through utilizing the overlaps between the groups. These overlaps result in nonsmooth ℓ_2-norms that share variables, complicating the optimization (9b). Let $\boldsymbol{\gamma} \triangleq \mathrm{vec}(X^2) \in \mathbb{R}^{(w_p+w_f)^2}$ and $\mathbf{q} \triangleq \mathrm{vec}(X^{1,k+1} + U^k)$, and then (9b) can be written as

$$\min_{\boldsymbol{\gamma}\in\mathbb{R}^{(w_p+w_f)^2}} \lambda\Omega_{\mathrm{LOG}}(\boldsymbol{\gamma}) + \frac{\rho}{2}\|\boldsymbol{\gamma}-\mathbf{q}\|_2^2, \tag{11}$$

where the LOG penalty is with respect to a known grouping of variables $\mathcal{G}$. Using the definition of the LOG penalty in (7), (11) is equivalent to

$$\min_{Z\in\mathbb{R}^{(w_p+w_f)^2\times|\mathcal{G}|}} \left\{\lambda\sum_{g\in\mathcal{G}} w_g\|Z_{.g}\|_2 + \frac{1}{2}\|\sum_{g\in\mathcal{G}} Z_{.g} - \mathbf{q}\|_2^2 \text{ s.t. } (Z_{.g})_{g^c} = 0,\ \forall g\in\mathcal{G}\right\}, \tag{12}$$

where $Z_{.g}$ denotes the column of Z indexed by g. We solve (12) using another ADMM algorithm with a sharing scheme. Splitting (12) into two blocks, the problem is equivalent to

$$\min_{Z^1,Z^2\in\mathbb{R}^{(w_p+w_f)^2\times|\mathcal{G}|}} \left\{\lambda\sum_{g\in\mathcal{G}} w_g\|Z_{.g}^1\|_2 + \frac{1}{2}\|\sum_{g\in\mathcal{G}} Z_{.g}^2 - \mathbf{b}\|_2^2, \text{ s.t. } Z^1 = Z^2,\ (Z_{.g}^1)_{g^c} = \mathbf{0}\ \forall g\in\mathcal{G}\right\}. \tag{13}$$

The ADMM iterates in scaled form (Boyd et al., 2011) to solve (13):

$$Z_{.g}^{1,k+1} \leftarrow \operatorname*{argmin}_{Z_{.g}^1\in\mathbb{R}^{(w_p+w_f)^2}} \left\{\lambda w_g\|Z_{.g}^1\|_2 + \frac{\rho}{2}\|Z_{.g}^1 - Z_{.g}^{2,k} + W_{.g}^k\|_2^2 \text{ s.t. } (Z_{.g}^1)_{g^c} = \mathbf{0}\right\},\ \forall g\in\mathcal{G}, \tag{14}$$

$$Z^{2,k+1} \leftarrow \operatorname*{argmin}_{Z^2\in\mathbb{R}^{(w_p+w_f)^2\times|\mathcal{G}|}} \frac{1}{2}\|\sum_{g\in\mathcal{G}} Z_{.g}^2 - \mathbf{q}\|_2^2 + \frac{\rho}{2}\sum_{g\in\mathcal{G}}\|Z_{.g}^2 - Z_{.g}^{1,k+1} - W_{.g}^k\|_2^2, \tag{15}$$

$$W_{.g}^{k+1} \leftarrow W_{.g}^k + \alpha(Z_{.g}^{1,k+1} - Z_{.g}^{2,k+1}),\ \forall g\in\mathcal{G}. \tag{16}$$

The solution to subproblem (14) is provided by the proximal map of the ℓ_2-norm (Parikh & Boyd, 2014) and can be parallelized across groups.

Subproblem (15) is potentially a large-scale problem with $(w_p + w_f)^2|\mathcal{G}|$ variables; however, it is possible to decrease its size to $(w_p+w_f)^2$ variables. (15) is equivalent to

$$\min_{Z^2\in\mathbb{R}^{(w_p+w_f)^2\times|\mathcal{G}|},\ \mathbf{z}\in\mathbb{R}^{(w_p+w_f)^2}} \left\{\frac{1}{2}\|\ |\mathcal{G}|\mathbf{z} - \mathbf{q}\|_2^2 + \frac{\rho}{2}\sum_{g\in\mathcal{G}}\|Z_{.g}^2 - Z_{.g}^{1,k+1} - W_{.g}^k\|_2^2 \text{ s.t. } \mathbf{z} = (1/|\mathcal{G}|)\sum_{g\in\mathcal{G}} Z_{.g}^2\right\}. \tag{17}$$

Minimizing over $Z_{.g}^2$ with $\mathbf{z}$ fixed and using first-order optimality conditions, we get

$$Z_{.g}^2 = \mathbf{z} + Z_{.g}^{1,k+1} + W_{.g}^k - (1/|\mathcal{G}|)\sum_{g\in\mathcal{G}}(Z_{.g}^{1,k+1} + W_{.g}^k),\ \forall g\in\mathcal{G}. \tag{18}$$

Using (18) in (17), we get

$$\mathbf{z}^{k+1} = \frac{1}{|\mathcal{G}|+\rho}\left(\mathbf{q} + \frac{\rho}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(Z_{.g}^{1,k+1} + W_{.g}^k)\right). \tag{19}$$

Furthermore, using (18) in (16), we finally get

$$W_{.g}^{k+1} = (1/|\mathcal{G}|)\sum_{g\in\mathcal{G}}(Z_{.g}^{1,k+1} + W_{.g}^k) - \mathbf{z}^{k+1},\ \forall g\in\mathcal{G}. \tag{20}$$

Note that the dual variables are independent of g; i.e., dual variables are equal for all groups. The sharing implementation of the proposed ADMM algorithm is illustrated in Algorithm 1. The convergence properties of the proposed ADMM algorithm on the general class of problems are discussed in Zhang, Liu, and Tajbakhsh (2020).

Algorithm 1 ADMM with sharing scheme to solve $\mathbf{prox}_{\lambda\Omega_{\mathrm{LOG}}}(\mathbf{q})$

```
Require: $\mathbf{q}$, $\lambda$, $w_g$ $\forall g\in\mathcal{G}$
$k = 0,\ W_{.g}^0 = \mathbf{0},\ Z_{.g}^{2,0} = \mathbf{0}\ \forall g\in\mathcal{G}$
while stopping criterion not met do
    $k \leftarrow k+1$
    $Z_{gg}^{1,k+1} \leftarrow \mathbf{prox}_{\lambda w_g\|\cdot\|_2}(Z_{gg}^{2,k} - W_{gg}^k),\quad \forall g\in\mathcal{G}$
    $Z_{g^cg}^{1,k+1} \leftarrow \mathbf{0},\quad \forall g\in\mathcal{G}$
    $\mathbf{z}^{k+1} \leftarrow \frac{1}{|\mathcal{G}|+\rho}\left(\mathbf{q} + \frac{\rho}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(Z_{.g}^{1,k+1} + W_{.g}^k)\right)$
    $Z_{.g}^{2,k+1} \leftarrow \mathbf{z}^{k+1} + Z_{.g}^{1,k+1} + W_{.g}^k - (1/|\mathcal{G}|)\sum_{g\in\mathcal{G}}(Z_{.g}^{1,k+1} + W_{.g}^k),\ \forall g\in\mathcal{G}$
    $W_{.g}^{k+1} = (1/|\mathcal{G}|)\sum_{g\in\mathcal{G}}(Z_{.g}^{1,k+1} + W_{.g}^k) - \mathbf{z}^{k+1},\ \forall g\in\mathcal{G}$
end while
Output: $\sum_{g\in\mathcal{G}} Z_{.g}^{1,k+1}$
```

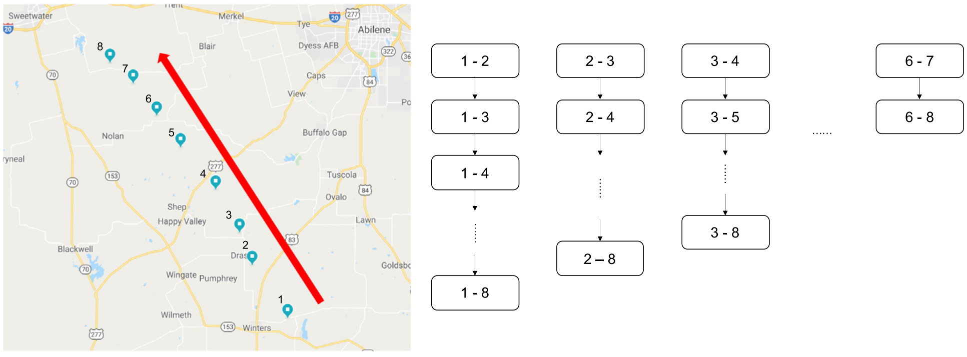
4. Application

In this section, the proposed methodology is used to predict wind speed in a case study in Texas. In the following numerical experiments, the prediction performance of the proposed GLOG-L method is compared with the following methods: (i) ARIMA; (ii) Vector AR (VAR); (iii) Sparse-VAR (SVAR), see Cavalcante, Bessa, Reis, and Browell (2017) (where all three methods are implemented as R packages); (iv) G-L (our implementation by solving the graphical lasso estimate in (6)); and (v) GLOG-L with mean structure, namely GLOG-L$_\mu$ (using weather-related covariates to model the mean).

For GLOG-L and G-L, the data are transformed into the Gaussian field, and the predictions are calculated based on the mean and variance of the conditional distribution (2). Then, the results are transformed back to the non-Gaussian field (see Fig. 1). On the other hand, for ARIMA, VAR, and SVAR, the predictions are obtained directly in the original (non-Gaussian) field.

4.1. Data description and corresponding hierarchical structures

The data for this study pertain to the Wind Integration National Dataset (Draxl, Clifton, Hodge, & McCaa, 2015)



(a)

(b)

Fig. 4. (a) Selected site locations and the wind direction. (b) Hierarchical sparsity structures obtained from the site locations and wind direction.

obtained from the NREL website.[1] We used data from June to August from 2007 to 2012 over $m = 8$ sites located along a line that is closely aligned with the summer prevailing wind direction from southeast to northwest. The site locations and wind direction are illustrated in Fig. 4(a). As stated in Sections 3.1 and 3.2, the derived wind direction pattern is used to generate the hierarchical sparsity structure over the elements of the inverse covariance matrix that pertain to the between-site wind directions. The corresponding hierarchies are presented in Fig. 4(b), and the site coordinates are provided in Table B.5 in Appendix A.

We used the first five years to build the models and predict the wind speed in the sixth year. The prediction horizon was set to $w_f = 18$, which included 18 consecutive time points in 20-minute intervals from 12:00 p.m. to 6:00 p.m.

4.2. Performance measures

The two performance measures based on point predictions are the root mean square error (RMSE) and the mean absolute error (MAE). Let $\hat{w}_T(s)$ denote the wind prediction at time T and location $s \in \mathcal{S}$ by a given model, and let $w_T(s)$ be the corresponding observation. Define $t_i \triangleq t_c + 72(i-1) + t$, where $t_c = 108$, calculated as $108 = (24+12) \times 3$, which corresponds to 11:40 a.m. the next day. The RMSE at time point t and site s is calculated as

$$\mathrm{RMSE}_t(s) = \left(\frac{1}{n} \sum_{i=1}^{n} \left(\hat{w}_{t_i}(s) - w_{t_i}(s) \right)^2 \right)^{1/2}, \quad t = 1, \ldots, w_f, \; s \in \mathcal{S},$$

and the MAE is calculated as

$$\mathrm{MAE}_t(s) = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{w}_{t_i}(s) - w_{t_i}(s) \right|, \quad t = 1, \ldots, w_f, \; s \in \mathcal{S},$$

where n is the number of prediction days (in the sixth year) and is equal to 90.

The performance measure based on the prediction interval is the Winkler score (Winkler, 1972). Let $[\hat{l}_T(s), \hat{u}_T(s)]$ denote the $100(1-\alpha)\%$ prediction interval for site s at time T. The Winkler score is calculated as

$$W_t(s) = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} \left(\hat{u}_{t_i}(s) - \hat{l}_{t_i}(s) \right), \\ \quad \text{if} \quad \hat{l}_{t_i}(s) \le w_{t_i}(s) \le \hat{u}_{t_i}(s) \\ \frac{1}{n} \sum_{i=1}^{n} \left(\hat{u}_{t_i}(s) - \hat{l}_{t_i}(s) + \frac{2}{\alpha}(\hat{l}_{t_i}(s) - w_{t_i}(s)) \right), \\ \quad \text{if} \quad w_{t_i}(s) < \hat{l}_{t_i}(s) \\ \frac{1}{n} \sum_{i=1}^{n} \left(\hat{u}_{t_i}(s) - \hat{l}_{t_i}(s) + \frac{2}{\alpha}(w_{t_i}(s) - \hat{u}_{t_i}(s)) \right), \\ \quad \text{if} \quad w_{t_i}(s) > \hat{u}_{t_i}(s). \end{cases}$$

Wide intervals and non-coverage of the true observations will increase the value of this performance measure. Hence, lower values indicate better prediction performance.

4.3. Parameter settings

For each site, the order of ARIMA(P,D,Q) was determined individually by AIC, as listed in Table C.6 in Appendix B. The order of the VAR model was also identified by AIC, where VAR(14) had the lowest AIC among VAR(1) to VAR(15). Accordingly, the highest order for SVAR was set to 15. The tuning parameter λ of SVAR, as well as for GLOG-L and G-L, was determined by cross-validation. Furthermore, w_g in (7) was set to $w_g = w_0|g|^{1/k}$, where w_0 and $k > 1$ were also determined by cross-validation.

We also implemented GLOG-L with a mean structure, namely GLOG-L$_\mu$, where a linear regression model was used to model the mean in (3) with temperature, air pressure, and air density as covariates. To fit GLOG-L$_\mu$, wind speeds were first transformed to the Gaussian field. A linear regression model was then fitted to the transformed data. Finally, the GLOG-L covariance matrix was estimated from the residual process. Predictions were based on (3), which were then transformed back to the non-Gaussian field.

[1] https://www.nrel.gov/grid/wind-toolkit.html.

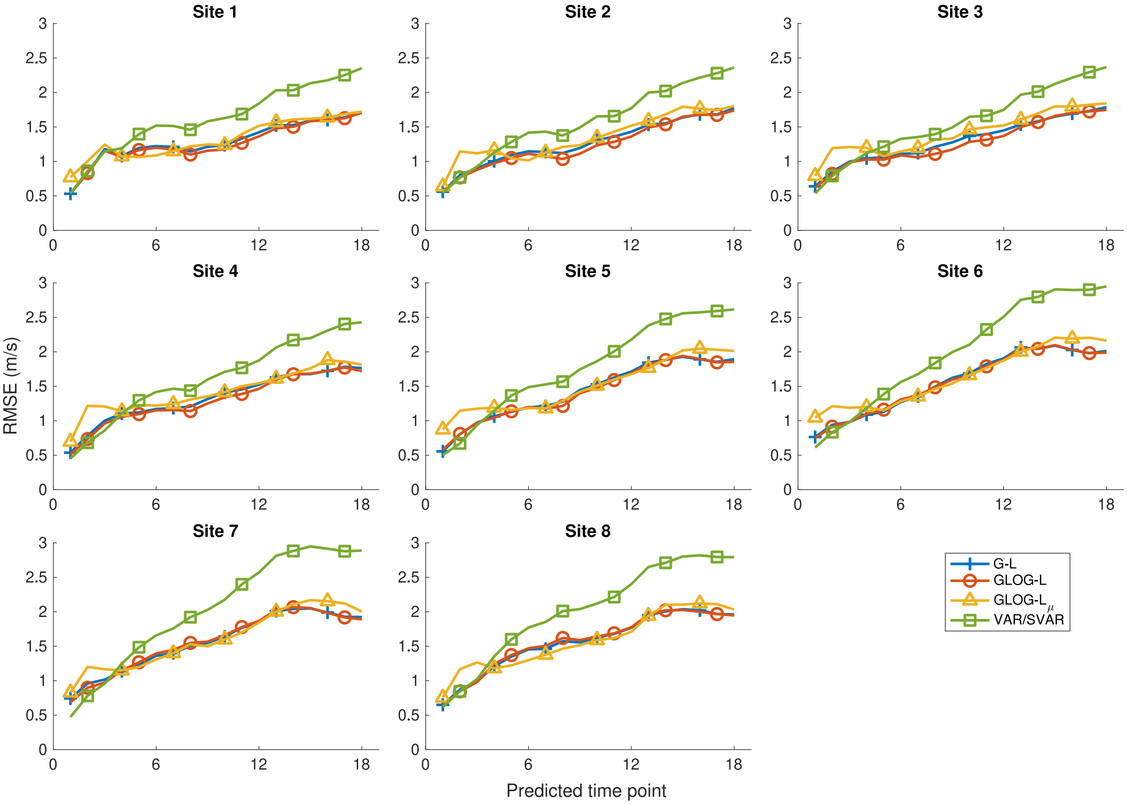


Fig. 5. Average RMSE for the next $w_f = 18$ time points at eight different sites.

Table 1
Comparison of prediction errors with different w_p and δ; boldface numbers show optimal combinations.

$w_p\backslash\delta$	RMSE				MAE			
	1	6	12	18	1	6	12	18
18	1.7225	1.7326	1.8042	1.7794	1.3746	1.3761	1.4434	1.3811
36	1.6028	1.5995	1.5946	1.6850	1.2711	1.2719	1.2650	1.2941
54	1.4881	1.4836	1.5552	1.5959	1.1848	1.1809	1.2369	1.2296
72	1.4306	1.4238	**1.3977**	1.5444	1.1277	1.1193	**1.0992**	1.1837
90	1.4516	1.4584	1.5309	1.5722	1.1537	1.1593	1.2159	1.2290

4.4. Setting w_p and δ to construct the data matrix for GLOG-L

Note that to construct the data matrix for GLOG-L, there is a trade-off in tuning δ and w_p in (4). Increasing δ results in a data matrix with a smaller N, but it increases the independence between consecutive data points. Increasing w_p increases the dimension of the covariance matrix Σ, which allows longer temporal correlations to play a role in predictions, but at the cost of making the model more complex. This is the tradeoff between bias and variance. Furthermore, we note that increasing w_p and hence the covariance matrix size, significantly increases the overall computational complexity, as the per-iteration complexity of the algorithm is $\mathcal{O}((w_p+w_f)^3)$. The effects of w_p and δ on prediction errors are illustrated in Table 1.

Additionally, we present the CPU times required for obtaining the GLOG-L estimate in Table 2. Calculations were implemented on a MacBook Pro with a 2.40 GHz 8-core CPU and 32 GB of RAM.

The parameters were set to $w_p = 72$ (of the previous data points, which were in 20-minute resolution) and $\delta = 12$, which were determined by the lowest prediction error over all combinations of $w_p \in \{18, 36, 54, 72, 90\}$ and $\delta \in \{1, 6, 12, 18\}$, resulting in $N = 2725$.

4.5. Numerical comparisons

4.5.1. Comparison based on point prediction performance measures

In Table 3, we report the RMSE and MAE of the predictions for each site (averaged over time). GLOG-L performed better than all other methods at six sites in terms of the RMSE and seven sites in terms of the MAE. On average over all sites, GLOG-L performed relatively better than the other methods on both performance measures.

Table 2
GLOG-L estimation times in seconds.

w_p	δ			
	1	6	12	18
18	77	78	77	79
36	208	206	201	210
54	410	410	411	422
72	736	696	722	764
90	1131	1179	1190	1147

Table 3
Comparison of RMSE (m/s) and MAE (m/s) for different methods. Boldface numbers indicate the methods with the least prediction errors.

Site	RMSE (m/s)					
	GLOG-L	G-L	GLOG-L$_\mu$	VAR	SVAR	ARIMA
1	**1.260**	1.289	1.325	1.630	1.629	1.907
2	**1.234**	1.273	1.354	1.581	1.575	1.853
3	**1.265**	1.306	1.416	1.569	1.571	1.840
4	**1.301**	1.337	1.426	1.624	1.631	2.422
5	**1.418**	1.435	1.509	1.789	1.799	2.621
6	**1.566**	1.571	1.643	2.012	2.040	3.215
7	1.566	**1.561**	1.610	2.043	2.079	3.462
8	1.572	**1.563**	1.588	2.024	2.015	3.197
Mean	**1.398**	1.417	1.484	1.784	1.792	2.565
	MAE (m/s)					
1	**0.978**	0.988	1.048	1.304	1.313	1.553
2	**0.976**	0.998	1.104	1.264	1.265	1.489
3	**1.001**	1.026	1.147	1.242	1.252	1.475
4	**1.019**	1.041	1.126	1.305	1.310	2.022
5	**1.116**	1.133	1.193	1.435	1.444	2.200
6	**1.229**	1.239	1.297	1.622	1.641	2.699
7	**1.225**	1.226	1.272	1.675	1.699	2.962
8	1.249	**1.235**	1.252	1.627	1.627	2.710
Mean	**1.099**	1.111	1.180	1.434	1.444	2.139

Table 4
Winkler score for different methods. Boldface numbers indicate the methods with the lowest scores.

α	Site	GLOG-L	G-L	GLOG-L$_\mu$	VAR	SVAR	ARIMA
0.1	1	**5.132**	5.166	5.659	5.949	5.505	6.804
	2	**5.051**	5.113	5.586	5.882	5.412	6.732
	3	**5.051**	5.097	5.545	5.887	5.445	6.646
	4	**5.190**	5.270	5.716	5.862	5.458	6.903
	5	**5.278**	5.387	5.772	5.997	5.608	7.012
	6	**5.872**	6.026	6.354	6.631	6.216	7.913
	7	**5.992**	6.157	6.514	6.799	6.381	8.116
	8	**5.798**	5.958	6.335	6.657	6.256	7.749
	Mean	**5.421**	5.522	5.935	6.208	5.785	7.235
0.05	1	**5.983**	6.016	6.539	6.973	6.442	8.031
	2	**5.892**	5.954	6.448	6.916	6.345	7.959
	3	**5.890**	5.936	6.398	6.923	6.389	7.870
	4	**6.043**	6.126	6.588	6.886	6.389	8.217
	5	**6.137**	6.262	6.652	7.015	6.523	8.361
	6	**6.843**	7.024	7.357	7.732	7.204	9.394
	7	**7.000**	7.193	7.565	7.936	7.398	9.643
	8	**6.767**	6.955	7.361	7.752	7.247	9.210
	Mean	**6.319**	6.433	6.864	7.267	6.742	8.586
0.01	1	**7.548**	7.583	8.203	8.928	8.270	10.248
	2	**7.439**	7.505	8.098	8.887	8.171	10.200
	3	**7.455**	7.503	8.063	8.908	8.235	10.099
	4	**7.622**	7.721	8.240	8.894	8.238	10.658
	5	**7.761**	7.913	8.356	9.038	8.376	10.924
	6	**8.709**	8.935	9.319	9.935	9.226	12.265
	7	**8.960**	9.211	9.637	10.192	9.481	12.589
	8	**8.660**	8.897	9.399	9.931	9.272	12.025
	Mean	**8.019**	8.158	8.664	9.339	8.659	11.126

However, differences among GLOG-L, G-L, and $GLOG - L_\mu$ were not large on these two measures.

In Fig. 5, we plot the RMSE at each predicted future time point for all eight sites to compare the prediction performance of GLOG-L, G-L, GLOG-L$_\mu$, VAR, and SVAR. The ARIMA model is omitted from the plot due to its poor performance across the prediction horizon. Furthermore, VAR and SVAR are plotted as one curve, as their values were almost identical. At earlier time points, VAR had a lower RMSE, while all other methods performed better at later time points. GLOG-L, GLOG-L$_\mu$, and G-L had relatively similar performance, and GLOG-L performed better than the other methods, especially at midrange time points. Furthermore, GLOG-L performed better than GLOG-L$_\mu$ at earlier time points, which shows that the mean regression model is unhelpful at short time horizons.

To further compare the prediction results, we implemented the Diebold–Mariano (DM) test (Diebold & Mariano, 2002) to identify whether GLOG-L has better predictive accuracy. Fig. 6 illustrates the DM test statistic that compares the performance of G-L, GLOG-L$_\mu$, and VAR relative to the GLOG-L method, along with $\alpha = 0.1,\ 0.05,\ 0.01$ significance intervals. Note that a negative value of the DM statistic means that the prediction of GLOG-L has a lower error at that time point. Needless to say, the differences are significant if the statistic lies outside of the insignificance intervals. Except at earlier time points, GLOG-L performed better than VAR at all significance levels. Furthermore, at significance levels 0.1 and 0.05, GLOG-L performed better than G-L at midrange time points, whereas they performed similarly at other times. GLOG-L and GLOG-L$_\mu$ performed similarly in relative terms except at earlier time points, where GLOG-L performed better.

4.5.2. Comparison based on prediction interval performance measure

In addition to the two point prediction performance measures, we compared different methods based on the Winkler score prediction interval performance measures. Table 4 presents the averaged Winkler score for each site at three different significance levels. As discussed above, a smaller value of the Winkler score indicates a better coverage of the true observation by the prediction interval or a tighter interval. At all three significance levels, GLOG-L had the lowest Winkler scores at all sites, which indicates that GLOG-L achieves relatively better prediction intervals compared to the other methods.

4.5.3. Impact of the wind direction on the GLOG-L method

Finally, to check the significance of the considered wind direction in GLOG-L, we tested a new wind direction, opposite to the prevailing wind direction, namely GLOG-L(opp). To compare the prediction accuracy, the DM test for G-L and GLOG-L(opp) based on GLOG-L is plotted

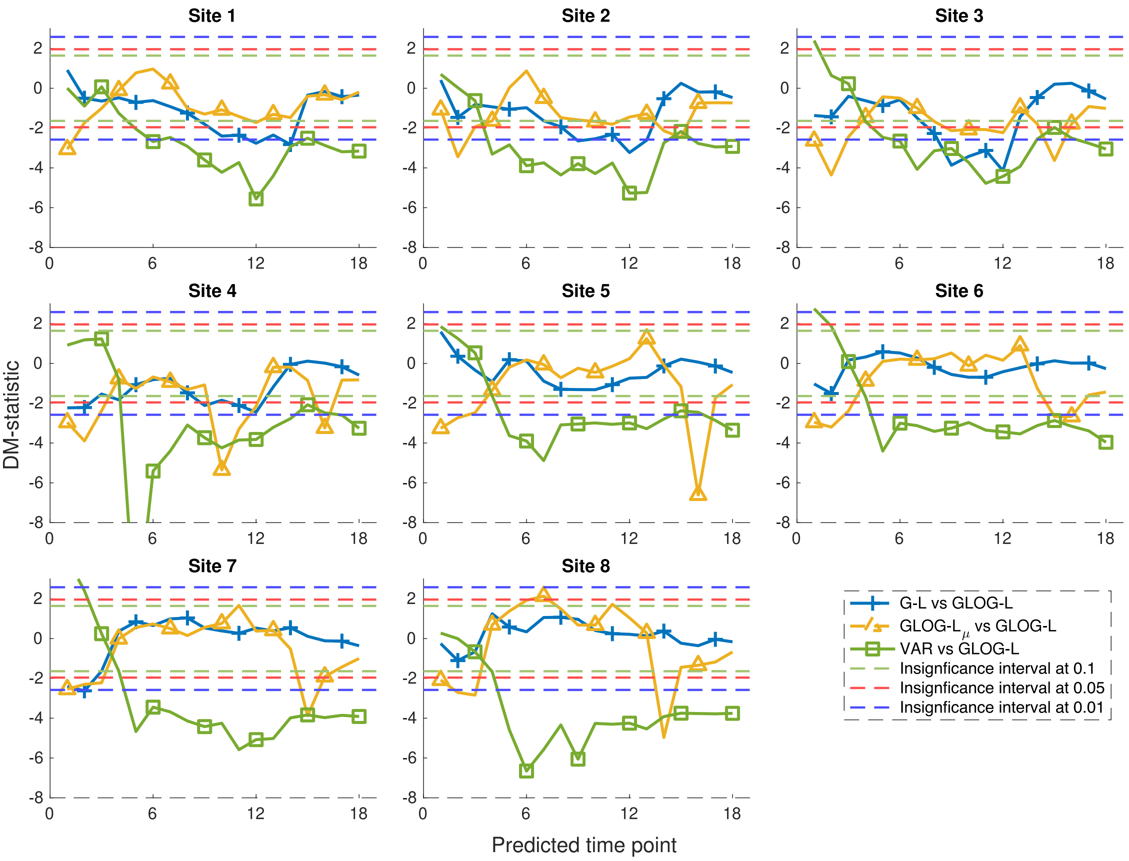


Fig. 6. Diebold–Mariano (DM) test for the next $w_f = 18$ time points at eight sites.

in Fig. 7. At significance levels 0.1 and 0.05, the prediction of GLOG-L was better than GLOG-L(opp), whereas at some periods, GLOG-L was not significantly better than G-L. Therefore, misidentified wind directions generally degenerate the performance of the GLOG-L method.

Furthermore, we compared the sparsity pattern of the estimated inverse covariance matrix by GLOG-L, G-L, and GLOG-L(opp) (see Fig. 8). As expected, different learning methods (GLOG-L vs. G-L) and different hierarchical structures (along with or opposite to the wind direction) resulted in different sparsity patterns in the inverse covariance matrices.

5. Conclusions

In the context of wind forecasting by conditional distribution based on spatiotemporal data, we proposed a new method to estimate the inverse covariance matrix of the Gaussian distribution for the transformed data, namely, the GLOG-L estimate. This method requires minimizing the negative log-likelihood function regularized with the LOG penalty. The LOG penalty forces the sparsity pattern of the optimal solution to follow a hierarchical structure, which conforms to wind dynamics known a priori. Further, we proposed an ADMM to efficiently evaluate the proximal mapping of the LOG penalty.

The proposed methodology was implemented and tested in a case study pertaining to Texas using data from eight different sites. Hierarchical sparsity structures were constructed based on wind direction inferred from wind roses obtained from these sites.

The prediction performance of the proposed method was benchmarked against a number of other statistical methods over different performance measures. In general, the results showed the potential of the proposed methodology as it takes into account wind dynamics.

While the proposed method may improve wind predictions when reasonably accurate estimates of the wind directions are available, inconsistent wind directions generally deteriorate its performance. Hence, a robust extension of the underlying optimization problem for regions with unstable wind directions could be a relevant future research direction.

Acknowledgments

The second author (Dr. Davanloo Tajbakhsh) would like to acknowledge National Science Foundation for the Award #1634867.

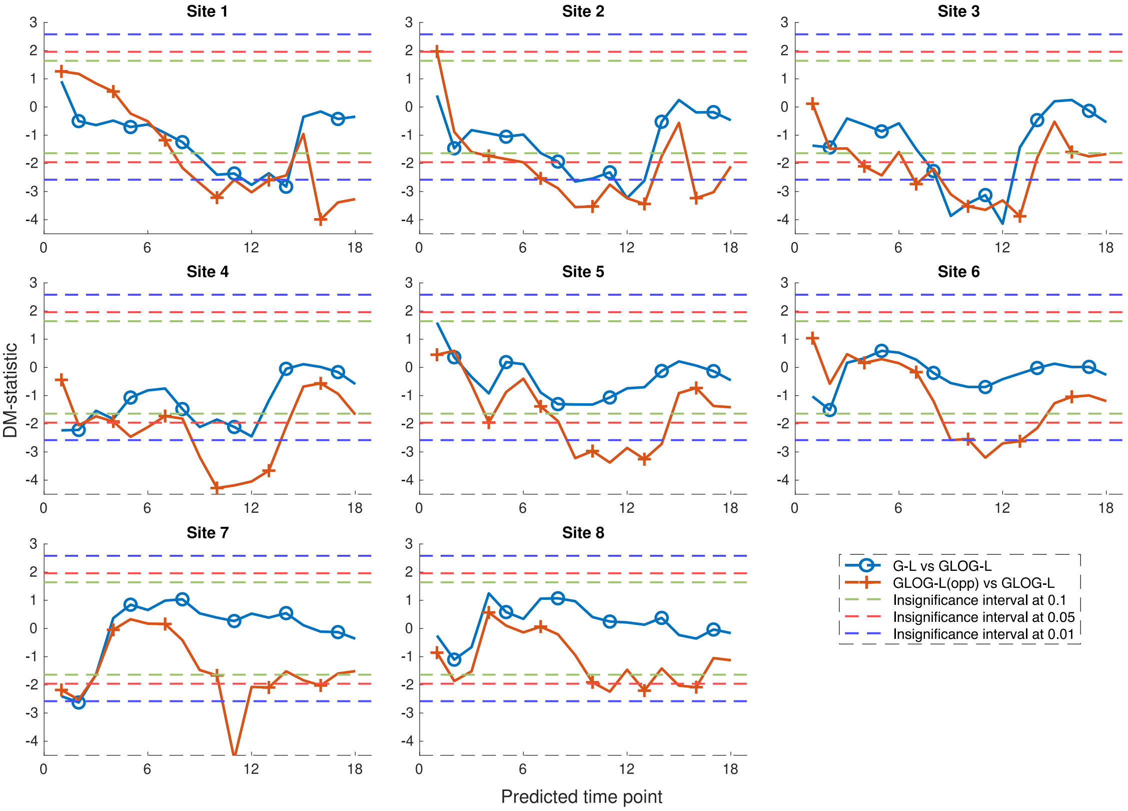


Fig. 7. Diebold–Mariano (DM) test for G-L and GLOG-L(opp) based on GLOG-L.

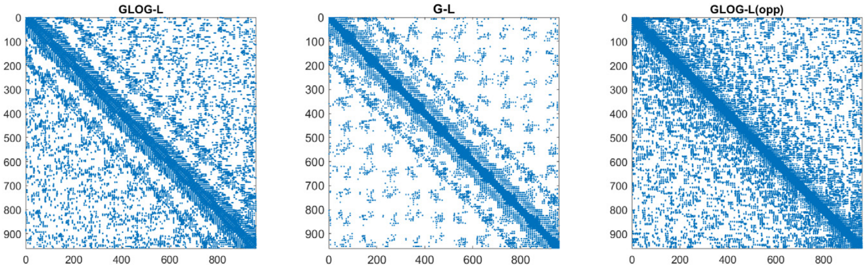


Fig. 8. Sparsity patterns of the inverse covariance matrices estimated by GLOG-L (along with and opposite to the wind direction) and G-L methods.

Appendix A. Solution to subproblem (9a) in the GLOG-L problem

Consider problem (9a) with X instead of X^1 as the variable for simplicity of notation. From the first-order optimality condition, we have

$$\rho X - X^{-1} = \rho(X^{2,k} - U^k) - \bar{S}, \tag{A.1}$$

with an implicit constraint $X \succ 0$. The idea is to find a solution that satisfies the above optimality condition and is also positive definite. Calculating the eigenvalue decomposition of the right-hand side, we get $\rho(X^{2,k} - U^k) - \bar{S} = Q\mathbf{diag}(\boldsymbol{\lambda})Q^\top$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{w_p+w_f})^\top$, and $QQ^\top = Q^\top Q = \mathbf{I}$. Multiplying (A.1) by $Q^\top$ from the left and Q from the right, we get

$$\rho\tilde{X} - \tilde{X}^{-1} = \mathbf{diag}(\boldsymbol{\lambda}),$$

where $\tilde{X} = Q^\top X Q$. Given that the right-hand side in the above equation is a diagonal matrix, we need to find $\tilde{X}_{ii}$ that satisfy $\rho\tilde{X}_{ii} - 1/\tilde{X} = \lambda_i$. The solution is

$$\tilde{X}_{ii} = \frac{\lambda_i + \sqrt{\lambda_i^2 + 4\rho}}{2\rho}$$

It follows that $X = Q\tilde{X}Q^\top$ satisfies the optimality condition (A.1) and is positive definite; hence, it is the solution to (9a).

Table B.5
Wind sites in Texas and their coordinates.

Site number	Latitude	Longitude
1	32.002563	−99.926147
2	32.055130	−99.972046
3	32.108597	−99.996613
4	32.179260	−100.043732
5	32.249004	−100.112366
6	32.301521	−100.158600
7	32.354027	−100.204895
8	32.388382	−100.250092

Table C.6
Selected orders for ARMA models in the Texas case study.

Site	(P, D, Q)
1	(0,1,4)
2	(0,1,5)
3	(1,1,2)
4	(0,1,4)
5	(2,1,0)
6	(3,1,2)
7	(1,1,4)
8	(0,1,5)

Appendix B. Wind sites in Texas

See Table B.5.

Appendix C. Selected ARIMA(P,D,Q) model orders for the wind sites in Texas

See Table C.6.

References

Alexiadis, M. C., Dokopoulos, P. S., & Sahsamanoglou, H. S. (1999). Wind speed and power forecasting based on spatial correlation models. *IEEE Transactions on Energy Conversion*, *14*(3), 836–842.

Barbounis, T., & Theocharis, J. (2007). Locally recurrent neural networks for wind speed prediction using spatial correlation. *Information Sciences*, *177*(24), 5775–5797.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, *3*(1), 1–122.

Cassola, F., & Burlando, M. (2012). Wind speed and wind energy forecast through Kalman filtering of numerical weather prediction model output. *Applied Energy*, *99*, 154–166.

Cavalcante, L., Bessa, R. J., Reis, M., & Browell, J. (2017). LASSO Vector autoregression structures for very short-term wind power forecasting. *Wind Energy*, *20*(4), 657–675.

Chen, N., Qian, Z., Nabney, I. T., & Meng, X. (2014). Wind power forecasts using Gaussian processes and numerical weather prediction. *IEEE Transactions on Power Systems*, *29*(2), 656–665.

Chitsaz, H., Amjady, N., & Zareipour, H. (2015). Wind power forecast using wavelet neural network trained by improved clonal selection algorithm. *Energy Conversion and Management*, *89*, 588–598.

Damousis, I., Alexiadis, M., Theocharis, J., & Dokopoulos, P. (2004). A fuzzy model for wind speed prediction and power generation in wind parks using spatial correlation. *IEEE Transactions on Energy Conversion*, *19*(2), 352–361.

Davanloo Tajbakhsh, S., Serhat Aybat, N., & Del Castillo, E. (2014). On the theoretical guarantees for parameter estimation of Gaussian random field models: A sparse precision matrix approach (pp. arXiv–1405).

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Dowell, J., & Pinson, P. (2016). Very-short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Transactions on Smart Grid*, *7*(2), 763–770.

Dowell, J., Weiss, S., Hill, D., & Infield, D. (2014). Short-term spatio-temporal prediction of wind speed and direction. *Wind Energy*, *17*(12), 1945–1955.

Draxl, C., Clifton, A., Hodge, B.-M., & McCaa, J. (2015). The wind integration national dataset (wind) toolkit. *Applied Energy*, *151*, 355–366.

Eckstein, J., & Bertsekas, D. P. (1992). On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, *55*(1–3), 293–318.

El-Fouly, T., El-Saadany, E., Salama, M., El-Fouly, T., El-Saadany, E., & Salama, M. (2006). Grey predictor for wind energy conversion systems output power prediction. *IEEE Transactions on Power Systems*, *21*(3), 1450–1452.

Erdem, E., & Shi, J. (2011). ARMA based approaches for forecasting the tuple of wind speed and direction. *Applied Energy*, *88*(4), 1405–1414.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Gabay, D., & Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, *2*(1), 17–40.

Glowinski, R., & Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue Française d'automatique, Informatique, Recherche Opérationnelle. Analyse Numérique*, *9*(R2), 41–76.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., & Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center. *Journal of the American Statistical Association*, *101*(475), 968–979.

He, M., Yang, L., Zhang, J., & Vittal, V. (2014). A spatio-temporal analysis approach for short-term forecast of wind farm generation. *IEEE Transactions on Power Systems*, *29*(4), 1611–1622.

Hering, A. S., & Genton, M. G. (2010). Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, *105*(489), 92–104.

Jacob, L., Obozinski, G., & Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning* (pp. 433–440). ACM.

Jursa, R., & Rohrig, K. (2008). Short-term wind power forecasting using evolutionary algorithms for the automated specification of artificial intelligence models. *International Journal of Forecasting*, *24*(4), 694–709.

Kavasseri, R. G., & Seetharaman, K. (2009). Day-ahead wind speed forecasting using f-ARIMA models. *Renewable Energy*, *34*(5), 1388–1393.

Khalid, M., & Savkin, A. V. (2012). A method for short-term wind power prediction with multiple observation points. *IEEE Transactions on Power Systems*, *27*(2), 579–586.

Kusiak, A., Zheng, H., & Song, Z. (2009). Short-term prediction of wind farm power: a data mining approach. *IEEE Transactions on Energy Conversion*, *24*(1), 125–136.

Landry, M., Erlinger, T. P., Patschke, D., & Varrichio, C. (2016). Probabilistic gradient boosting machines for GEFCom2014 wind forecasting. *International Journal of Forecasting*, *32*(3), 1061–1066.

Lange, M., & Focken, U. (2006). *Physical approach to short-term wind power prediction*. Springer.

Mandic, D., Javidi, S., Goh, S., Kuh, A., & Aihara, K. (2009). Complex-valued prediction of wind profile using augmented complex statistics. *Renewable Energy*, *34*(1), 196–201.

Mangalova, E., & Agafonov, E. (2014). Wind power forecasting using the k-nearest neighbors algorithm. *International Journal of Forecasting*, *30*(2), 402–406.

Mangalova, E., & Shesterneva, O. (2016). K-nearest neighbors for GEFCom2014 probabilistic wind power forecasting. *International Journal of Forecasting*, *32*(3), 1067–1073.

Messner, J. W., & Pinson, P. (2018). Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting. *International Journal of Forecasting*.

Mohandes, M. A., Halawani, T. O., Rehman, S., & Hussain, A. A. (2004). Support vector machines for wind speed prediction. *Renewable Energy*, *29*(6), 939–947.

Möller, A., Lenkoski, A., & Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, *139*(673), 982–991.

Morales, J. M., Minguez, R., & Conejo, A. J. (2010). A methodology to generate statistically dependent wind speed scenarios. *Applied Energy*, *87*(3), 843–855.

Negnevitsky, M., Johnson, P., & Santoso, S. (2007). Short term wind power forecasting using hybrid intelligent systems. In *2007 IEEE power engineering society general meeting* (pp. 1–4). Tampa, FL, USA: IEEE.

Parikh, N., & Boyd, S. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, *1*(3), 127–239.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Richardson, L. F. (2007). *Weather prediction by numerical process*. Cambridge University Press.

Salcedo-Sanz, S., Ortiz-Garcı, E. G., Pérez-Bellido, Á. M., Portilla-Figueras, A., & Prieto, L. (2011). Short term wind speed prediction based on evolutionary support vector regression algorithms. *Expert Systems with Applications*, *38*(4), 4052–4057.

Sanchez, I. (2006). Short-term prediction of wind energy production. *International Journal of Forecasting*, *22*(1), 43–56.

Tascikaraoglu, A., Sanandaji, B. M., Poolla, K., & Varaiya, P. (2016). Exploiting sparsity of interconnections in spatio-temporal wind speed forecasting using wavelet transform. *Applied Energy*, *165*, 735–747.

Tastu, J., Pinson, P., Trombe, P., & Madsen, H. (2014). Probabilistic forecasts of wind power generation accounting for geographically dispersed information. *IEEE Transactions on Smart Grid*, *5*(1), 480–489.

Torres, J., García, A., De Blas, M., & De Francisco, A. (2005). Forecast of hourly average wind speed with ARMA models in Navarre (Spain). *Solar Energy*, *79*(1), 65–77.

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, *67*(337), 187–191.

Wytock, M., & Kolter, J. Z. (2013). Large-scale probabilistic forecasting in energy systems using sparse Gaussian conditional random fields. In *52nd IEEE conference on decision and control* (pp. 1019–1024). Firenze: IEEE.

Xie, L., Gu, Y., Zhu, X., & Genton, M. G. (2014). Short-term spatio-temporal wind power forecast in robust look-ahead power system dispatch. *IEEE Transactions on Smart Grid*, *5*(1), 511–520.

Yan, X., Bien, J., et al. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, *32*(4), 531–560.

Yu, J., Chen, K., Mori, J., & Rashid, M. M. (2013). A Gaussian mixture copula model based localized Gaussian process regression approach for long-term wind speed prediction. *Energy*, *61*, 673–686.

Zhang, D., Liu, Y., & Tajbakhsh, S. D. (2020). A first-order optimization algorithm for statistical learning with hierarchical sparsity structure. arXiv:2001.03322.

Zhang, Y., & Wang, J. (2016). K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting. *International Journal of Forecasting*, *32*(3), 1074–1080.

Zhao, Y., Ye, L., Pinson, P., Tang, Y., & Lu, P. (2018). Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting. *IEEE Transactions on Power Systems*, *33*(5), 5029–5040.

Zhou, J., Shi, J., & Li, G. (2011). Fine tuning support vector machines for short-term wind speed forecasting. *Energy Conversion and Management*, *52*(4), 1990–1998.

Zhu, X., Genton, M. G., Gu, Y., & Xie, L. (2014). Space-time wind speed forecasting for improved power system dispatch. *TEST*, *23*(1), 1–25.