Spline Model for Wake Effect Analysis: Characteristics of Single Wake and Its Impacts on Wind Turbine Power Generation.

Hoon Hwangbo[a], Andrew L. Johnson[a], and Yu Ding[a]

[a]Department of Industrial & Systems Engineering, Texas A&M University, TX, U.S.A

Abstract

Understanding and quantifying the wake effect plays an important role in improving wind turbine designs and operations as well as wind farm layout planning. The majority of the current wake effect models are physics-based, but these models have a number of shortcomings. Sophisticated models based on computational fluid dynamics suffer from computational limitations and are impractical for modeling commercial sized wind farms, whereas simplified physics-based models are generally inaccurate for wake effect quantification. Nowadays, data-driven wake effect models are gaining popularity as the data from commercially operating wind turbines become available, but this development is still in its early stages. This study contributes to the general category of the data-driven wake effect modeling that makes use of the actual wind turbine operational data. We propose a wake effect model based on splines with physical constraints incorporated, which sets out to estimate wake effect characteristics such as wake width and wake depth under single wake situations. Our model is one of the first data-driven models that provides a detailed account of the wake effect. Prediction accuracy of the proposed spline model, when compared with other alternatives, also confirms the benefit of incorporating the physical constraints in the statistical estimation.

KEY WORDS: Data-driven model, thin plate regression splines with non-negativity, wake model assessment, wake power loss, wind power generation.

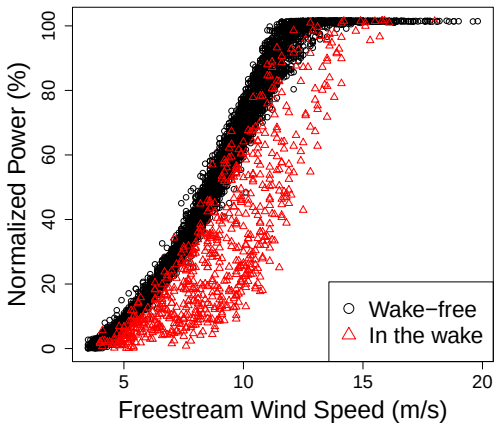

Figure 1: Power output in the wake versus that under a freestream condition. Freestream wind speed refers to wind speed measurements not affected by another turbine. Due to the confidentiality concern, power output is denoted throughout this paper by a percentage value normalized to its maximum.

1. Introduction

Wind energy is a promising renewable energy source. In 2015, wind energy supplied 4.7% of the total electricity generated in the U.S. (American Wind Energy Association, 2016). More impressively, the electricity generation from wind energy has increased tenfold in the past decade, from 18 million megawatt-hours (MWh) in 2005 to 190 million MWh in 2015 (Energy Information Administration, 2016). Department of Energy (DOE, 2015) envisions that wind energy could supply 10% of the nation's electricity by 2020, 20% by 2030, and 35% by 2050. Crucial to sustaining such a rapid growth is a deeper understanding of power generation performance of wind turbines, which translates to better designs and/or effective operations. In this study, we focus on an important aerodynamic feature affecting wind turbine performance, the so-called wake effect.

While a wind turbine is operating, the rotating blades disturb the natural flow of wind and create turbulence for the downstream turbines. During this process, the turbine absorbs kinetic energy in wind and converts the energy into electricity. As a result, the wind loses some of its original kinetic energy after the turbine rotor, exhibiting reduction in its speed. Such a phenomenon differentiating the after-rotor wind flow from the free stream one (before the rotor) is referred to as wake effect. As the amount of power output depends on ambient wind speed, the reduction in wind speed may substantially deteriorate power production at downstream wind turbines. Figure 1 illustrates power output of a wind turbine when it is wake free versus when it is in the wake of another turbine.

There have been significant efforts devoted to better understanding of wake effect and alleviation of its impact on power generation. Except for a few recent studies that are data-driven (which we will review in the next section), the majority of the current wake effect models are physics-based. A simplified parametric model, yet widely used in practice, is Jensen's model (Jensen, 1983). A primary shortcoming of the Jensen's model lies in its unsatisfactory accuracy in predicting turbine power loss under commercial operating conditions. The limitation of this simplified model has motivated researchers to resort to sophisticated, computational fluid dynamics (CFD) models that can achieve a higher accuracy (Laan et al., 2015). However, using the CFD models entails significant computational challenges. For example, running a large eddy simulation, one of the popular CFD methods, requires days or even weeks of computation on supercomputers for analyzing a single-wake situation (Sanderse et al., 2011).

In this paper, we propose a data-driven alternative to the physics-based wake effect models. We consider single wakes arising between two turbines of which modeling assumptions are easier to justify. Single wake situations are of great interest in the wake studying literature (Prospathopoulos et al., 2011; Duckworth and Barthelmie, 2008), and the single wake behaviors provide valuable insights into various decisions for improving wind farm performance (more details in Section 2). To facilitate a successful transition from physics-based models to data-driven modeling, we incorporate certain physical understandings and considerations as constraints in the data model fitting procedure. Because of this, our resulting model is more than, and it does outperform, a purely data-driven model.

We highlight an additional difference between the physics-based models and the data-driven models in general, as illustrated in Figure 2. All the physics-based models do not directly estimate the power loss. Instead, they primarily focus on estimating the reduced wind speed due to wake. To quantify wake power loss, these models then require an additional layer of converting the wind speed estimates into a corresponding power output; such a conversion can be done by using a simple power curve model as recommended by the International Electrotechnical Commission (IEC12.1, 2005) or more complicated power curve models such as in Lee et al. (2015a). Having a competitive physics-based model thus requires both steps to be effective, but improving the first step encounters the problem of computational complexity, which is a hard problem to address in practice. By contrast, the data-driven models, including the one proposed in this paper, connect the wind data directly to the power output in a single step. In our numerical study, we compare the data-driven model with the two-step approach equipped with a simple physics-based model (like Jensen's model).

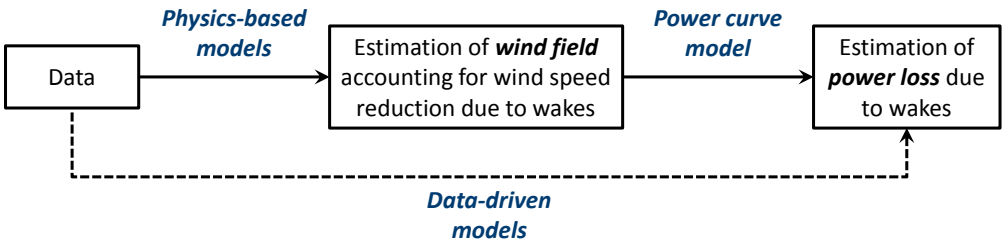

Figure 2: Wake power loss estimation procedures.

The remainder of the paper is organized as follows: in Section 2, we describe wake characteristics and their practical implications for improving power generation performance of wind energy systems. We also introduce the existing data-driven models and discuss their capability of deriving the wake characteristics. In Section 3, we present the proposed wake model based on a spline model structure incorporating physical constraints and describe the model-estimation procedure. We compare the prediction performance of our wake model with that of other alternatives in Section 4, and apply it to the analysis of wake effect under commercial operating conditions in Section 5. In Section 6, we conclude the paper.

2. Characteristics of wake effect and data-driven approaches

The wake of a turbine propagates with a certain range of angles (Gebraad et al., 2016), and its impact remains for a certain distance (Ammara et al., 2002). Figure 3(a) illustrates a snapshot of a single wake situation, in which we assume two operating turbines, for a given wind direction. In the figure, θ denotes an acute angle between the wind direction and the line connecting the two turbines, and it varies with wind direction. For the wind direction given in Figure 3(a), the wind passes through Turbine 1 along the center line and the wake caused by Turbine 1 affects the downstream region with a range of angles (the shaded area). The wind speed loss due to the wake is greater for locations closer to the upstream turbine (Turbine 1) and closer to the center line. Then, Turbine 2, given its fixed location, is subject to the greatest power loss when $\theta = 0$, and the amount of power loss decreases as θ deviates from zero. After θ exceeds a certain value, Turbine 2 is no longer in the wake of Turbine 1. The maximum power loss when $\theta = 0$ is referred to as the wake depth, whereas the range of θ for which a turbine is in the wake of another turbine (with positive power loss) is referred to as the wake width (see Figure 3(b)). Wake depth and width are expected to remain constant when the relative positions between two turbines are fixed, and their values will change when the turbines' relative positions are different.

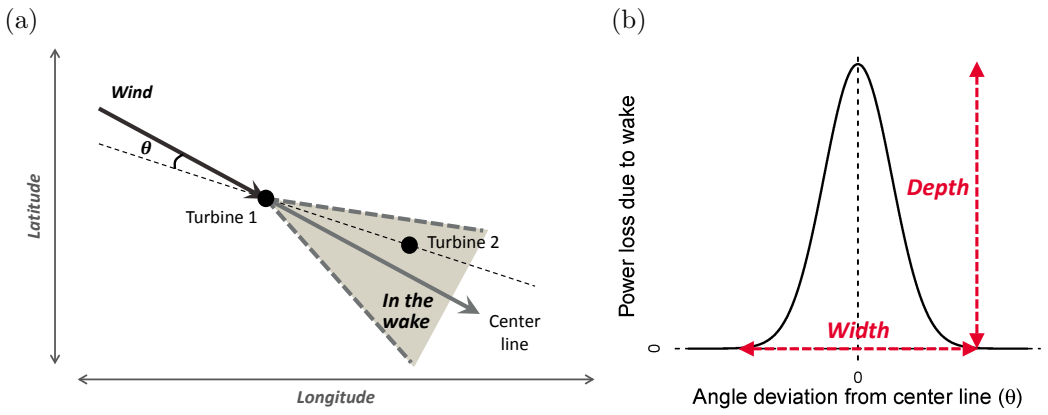

Figure 3: Characteristics of wind turbine wake and its effect: (a) wake region and θ; (b) wake depth and wake width. Wake power loss is expected to be a function of θ.

Knowledge of wake characteristics is crucial for improving power generation performance on wind farms. As wake width and depth strongly depend on the relative positions of turbines, characterizing the turbine specific wake effect facilitates the layout planning (Kusiak and Song, 2010; Emami and Noghreh, 2010), particularly when using the same turbine model in future wind projects. Understanding the wake characteristics also supports effective operational control of wind turbines through pitch and yaw controls (McKay et al., 2013; Gebraad et al., 2016). The pitch control can regulate the magnitude of wind speed loss in a downstream region by adjusting the energy absorption level of an upstream turbine, and the yaw control can also change the amount of the wind speed loss by tilting the downstream wake region; for instance, by controlling the yaw of Turbine 1, Turbine 2 can be as nearly wake free as possible for a given wind direction.

A common data-driven practice in industry to understand the wake effect is as follows. First, gather the power output data from two turbines. Second, choose a specific range of wind speed where the maximum power loss is expected, e.g., 8.0 $\pm$ 0.5 m/s (Barthelmie et al., 2010), or extend the coverage of wind speed to a wider range, e.g., 5.0–11.0 m/s (McKay et al., 2013). Then, plot the power difference between the two turbines under the above-specified wind speeds against the wind direction (0 degree means the north). Figure 4 shows a scatter plot of the power difference against wind direction. To smooth out the noise effect, people would apply the action of binning, namely, partition the wind direction by a unit, say 5°, and then average all the power difference data in a specific bin, and use the average as the representative of the original data. Applying the data binning to the raw data in Figure 4 produces the solid line passing through the data cloud.

The solid line is treated as the estimated curve representing the wake effect. The wake depth can be read from the plot by observing the two peaks around 120° and 300°, respectively. As we move with the wind direction from 0 to 360 degrees, the roles of the two turbines, namely that which one is wake free and which one is in the wake, are reversed. That is why we observe that one of the peaks is downward. The wake width is not immediately obvious, so that researchers usually impose a large enough angle coverage, say, $\theta \in (-25°, 25°)$, and then verify with the estimated curve if the angle range is broad enough to represent the wake width (McKay et al., 2013). Sometimes, one finds a wind direction value on each side of the center line at which the power loss estimate is within a certain level, for example, $\pm 5\%$ of the freestream power, and calculates the angle coverage formed by these wind direction values (Barthelmie et al., 2010). When using this data binning approach, a purely data-driven method, the estimate of the power difference from the wake-free turbine to the in-wake turbine (that estimates wake power loss) is not guaranteed to be positive. As a matter of fact, previous studies (Prospathopoulos et al., 2011; Troldborg et al., 2011) often show that some of the bin-wise estimate of this power difference is negative even after θ moves beyond the obvious wake width region; this phenomenon is in fact evident in Figure 4.

Other than the data binning approach, there are few other data-driven wake models. One exception is a sophisticated statistical model recently developed (You et al., 2017), which is a Gaussian Markov random field (GMRF) model with a Bayesian hierarchical structure that accounts for spatial correlation of turbines' power output at different locations. To be clear, this GMRF model was not particularly developed for studying the single

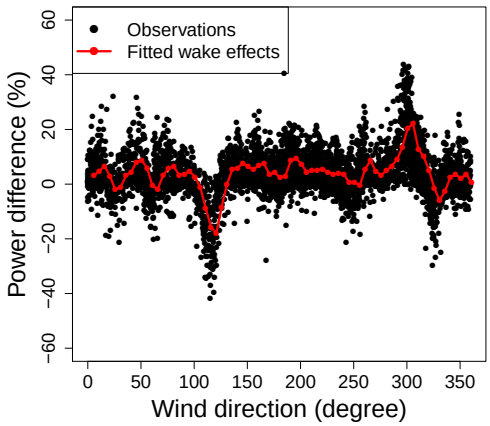

Figure 4: Estimation of wake effects between a pair of turbines. The between-turbine distance is four times the rotor diameter d. The distance from this pair to other turbines is more than $10d$.

wake situations. You et al. (2017) make use of the spatial correlations among turbines located close to one another and simultaneously estimates the heterogeneous power outputs from multiple turbines by modeling the wake interactions with GMRF. Their model is useful for analyzing wake effects in mid- to large-size wind farms, but it can lose estimation accuracy for two turbine settings. Moreover, their model does not impose the constraint that the wake power loss is positive. Nevertheless, due to its consideration of turbines in their relative spatial locations, the GMRF model could estimate the wake power loss indirectly by taking the difference of the maximum fitted value among all turbines and the power output fitted to a specific turbine.

You et al. (2017) review a number of other possible data-driven alternatives, such as global model (GLB), individual model (IND), and random effect model (RND), but conclude that those other alternatives are not competitive with the GMRF model. GLB and RND are in fact special cases of GMRF, all of which utilize the data from multiple turbines. We decide not to consider them further. On the other hand, IND predicts individual turbines' power output separately without using data from its neighbors, so that IND may have particular benefits in single wake setting. In fact, You et al. (2017) report IND is the second best model, just behind GMRF. As such, we use the data binning approach, GMRF, and IND as our benchmarks.

3. Spline-based wake effect model and its estimation

3.1. Baseline power production model

In this study, we borrow a model structure from the field of production economics for representing wind power production (Hwangbo et al., 2015), which reads as:

$$y_t(\mathbf{x}) = f(\mathbf{x}) - \eta_t(\mathbf{x}) - \omega_t(\mathbf{x}) + \epsilon_t, \quad t = 1, \ldots, T \tag{1}$$

where y is the power output, $\mathbf{x}$ is a vector of predictors that include wind related (and possibly other environmental) variables, and t denotes each of T turbines. $f(\cdot)$ represents the best achievable power output characterizing the full power production potential of a specific type of wind turbines given wind resources (also known as the production frontier in the production economics literature; Aigner et al. 1977). $\eta_t(\cdot)$ and $\omega_t(\cdot)$ are the inefficiency terms, namely the power losses relative to the best achievable production, and ϵ_t is the i.i.d. random noise. In Hwangbo et al. (2015), $\eta_t(\cdot)$ and $\omega_t(\cdot)$ are lumped into one single inefficiency term. Here, we split them into two so that $\eta_t(\cdot)$ represents a turbine's inherent

inefficiency independent of wake, whereas $\omega_t(\cdot)$ represents the turbine's power loss due to wake. In this model, we need both power loss terms to be non-negative, i.e., $\eta_t(\cdot) \geq 0$ and $\omega_t(\cdot) \geq 0$ for $\forall t = 1, \ldots, T$, to be consistent with the physical understanding of the phenomenon.

The existence of $f(\cdot)$ may sound strange to some readers, but including $f(\cdot)$ does not require any restrictive model assumptions. For a pair of turbines, one can pool the two turbines' power production data together and estimate a common production frontier; the full detail of such estimation process is described in Hwangbo et al. (2015). For our study of the single wake situations, however, this $f(\cdot)$ does not even need to be estimated explicitly. As one will see shortly, we will establish a power difference model in the subsequent section, which takes the power difference between a pair of turbines. By doing so, the common frontier cancels in the resulting model.

3.2. Power difference model for two-turbine cases

Recall that our focus is a single wake situation with two turbines. We introduce two angle variables, θ_1 and θ_2, to be associated with the two turbines. Specifically, θ_1 is related to the wind direction causing power loss on Turbine 1 and θ_2 is with the wind direction under which Turbine 2 endures power loss. As illustrated in Figure 5, the wind directions associated with θ_1 and θ_2 can take any value in the sets $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, given the definition of these sets stated below. For the purpose of analyzing the wake effect, θ_1 and θ_2 only need to vary in the 180° outer hemisphere surrounding their respective turbine. Note that θ_1 is actually on the side of Turbine 2, whereas θ_2 is on the side of Turbine 1. If we position the zero degree of θ_1 and θ_2 at the line connecting the two turbines, then $\theta_1, \theta_2 \in (-90°, 90°)$. We denote by $\mathcal{D}_1$ the set of wind directions corresponding to the support of θ_1, and likewise, by $\mathcal{D}_2$ the set of directions in which θ_2 is defined.

With this notation and by following the baseline power production model in (1), we can write the individual power production functions for the two turbines, respectively:

$$\begin{aligned} y_1(\mathbf{x}) &= f(\mathbf{x}) - \eta_1(\mathbf{x}) - \omega_1(\mathbf{x}) \cdot \mathbb{1}_{\mathcal{D}_1}(\mathbf{x}) + \epsilon_1, \\ y_2(\mathbf{x}) &= f(\mathbf{x}) - \eta_2(\mathbf{x}) - \omega_2(\mathbf{x}) \cdot \mathbb{1}_{\mathcal{D}_2}(\mathbf{x}) + \epsilon_2, \end{aligned} \tag{2}$$

where $\mathbb{1}_{\mathcal{D}_t}(\mathbf{x})$ is an indicator function taking the values of 1 if the wind direction belongs to $\mathcal{D}_t$ or 0 otherwise. Again, the production frontier $f(\mathbf{x})$ is assumed to be common to the same type of turbines, so there is no differentiating subscript used on it. Let us take the

difference between the two equations in (2). Then, we have

$$\tilde{y}_{1-2}(\mathbf{x}) = \tilde{\eta}_{2-1}(\mathbf{x}) - \omega_1(\mathbf{x}) \cdot \mathbb{1}_{\mathcal{D}_1}(\mathbf{x}) + \omega_2(\mathbf{x}) \cdot \mathbb{1}_{\mathcal{D}_2}(\mathbf{x}) + \tilde{\epsilon}, \tag{3}$$

where the tilde indicates a turbine difference term and the subscripts $1-2$ and $2-1$ signify the specific order of the difference. The above model is interpreted as follows: the power difference of Turbine 1 over Turbine 2 is due to the inherent production difference between the two turbines, $\tilde{\eta}_{2-1}(\cdot)$, and the power loss caused by the wake effect, characterized by either $\omega_1(\cdot)$ or $\omega_2(\cdot)$, both depending on specific wind conditions. Because the sets $\mathcal{D}_1$ and $\mathcal{D}_2$ are mutually exclusive, $\omega_1(\cdot)$ and $\omega_2(\cdot)$ will not appear at the same time.

To specify the above model, we further clarify what should be included in the input vector $\mathbf{x}$. In general, it is well-known that the dominating input factors for wind power production are wind speed, V, and wind direction, D. However, Lee et al. (2015a) have shown that environmental factors other than wind speed and direction, such as air density and humidity, may also have an impact on wind power output. One advantage of using the power difference model (3) is that we no longer need to consider other environmental factors because once we take the power difference between the two turbines, the impact of the environmental factors other than that of the wind is neutralized. Still, to be consistent with the IEC standard procedure and to further neutralize the effect of air density, ρ, we decide to use the normalized wind speed, following the industry standard (IEC12.1, 2005):

$$V = V' \left(\frac{\rho}{\rho_0} \right)^{1/3},$$

where V' is the raw measurement of wind speed and $\rho_0 = 1.225\,\mathrm{kg/m^2}$ is the air density

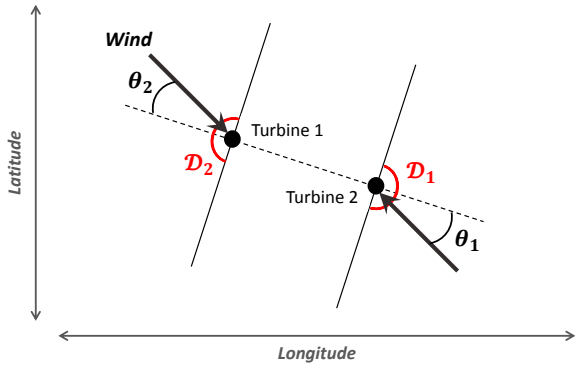

Figure 5: Two subsets of wind direction, $\mathcal{D}_1$ and $\mathcal{D}_2$. The union of the two subsets covers the entire 360° wind direction.

of the international standard atmosphere at sea level and at 15 °C. We hereinafter refer to this normalized wind speed V simply as the wind speed, unless otherwise stated.

For the turbine difference term and the indicator function terms in equation (3), either wind speed or wind direction, but not both, is needed as an input. The input to the indicator function $\mathbb{1}_{\mathcal{D}_1}(\cdot)$ or $\mathbb{1}_{\mathcal{D}_2}(\cdot)$ is wind direction. The turbine difference term, $\tilde{\eta}_{2-1}(\cdot)$, represents the between-turbine production difference independent of wake, and we assume it is only a function of wind speed, not of wind direction; the portion of the power difference, $\tilde{y}_{1-2}$ or $\tilde{y}_{2-1}$, related to wind direction should be included in the wake related term ω.

As such, the power difference model can be expressed as:

$$\tilde{y}_{1-2}(V, D) = \tilde{\eta}_{2-1}(V) - \omega_1(V, D) \cdot \mathbb{1}_{\mathcal{D}_1}(D) + \omega_2(V, D) \cdot \mathbb{1}_{\mathcal{D}_2}(D) + \tilde{\epsilon}. \tag{4}$$

Here, $\tilde{\epsilon}$ is still an i.i.d. noise. We further assume it follows a normal distribution.

3.3. Spline model with non-negativity and model estimation

In order to empirically estimate the power difference model in (4), we need to assume an underlying model structure for the three functional terms, $\tilde{\eta}$, ω_1, and ω_2. For $\tilde{\eta}$ that has a single input, we choose to use a cubic smoothing spline, whereas for the two wake power loss terms, ω_1 and ω_2, that have two inputs, we choose to use thin plate splines (Duchon, 1977), the multidimensional generalization of the smoothing splines. As such, the power difference model (4) becomes a spline model with the non-negativity constraint imposed on ω_1 and ω_2. We refer to the resulting model as the thin plate regression spline model with non-negativity (TPRS-N). Note that, although the inefficiency term $\eta(\cdot)$ in the baseline power production model needs to be non-negative, $\tilde{\eta}(\cdot)$, which is the difference between two individual $\eta(\cdot)$'s, can take any value.

There are alternative models other than the spline models for fitting the power difference model in (4), such as Gaussian process regression (Rasmussen and Williams, 2006) or kernel regression (Nadaraya, 1964; Watson, 1964). Without the non-negativity constraints, all these models produce outcome comparable with one another, thus the particular modeling option chosen matters less. With the non-negativity constraints, however, we find it easier to work with the spline models. The traditional response surface models, represented by second order regression models, are not considered competitive options here, because of their parametric nature as well as the difficulty of imposing the constraints.

To estimate the spline-based power difference model, we follow the generalized additive

model (GAM) scheme (Hastie and Tibshirani, 1990). GAMs represent a univariate response as an additive sum of multiple smooth functions, each having its own predictor variables. Estimation of GAMs can be performed by implementing the backfitting algorithm for which each smooth function is fitted for the residuals of all the others, iteratively one at a time until the fitted functions converge.

Consider n data pairs for which a residual r is paired with covariates $\mathbf{x}$, i.e., $(\mathbf{x}_i, r_i)$ for $i = 1, \ldots, n$. Then, we estimate a smooth function by finding h that minimizes

$$\sum_{i=1}^{n} \{r_i - h(\mathbf{x}_i)\}^2 + \lambda J[h], \tag{5}$$

where λ is a penalty parameter controlling the trade-off between data fitting and smoothness of h. Both smoothing splines and thin plate splines solve the penalized least squares problem stated in (5), but the measure of the smoothness, denoted by $J[h]$, would be different as they consider different dimensional functional spaces. For smoothing splines with a univariate predictor x,

$$J[h] = \int_{\Re} h''(x)^2 dx,$$

whereas for thin plate splines with two predictors, x_1 and x_2,

$$J[h] = \int\int_{\Re^2} \Big[\left(\frac{\partial^2 h(\mathbf{x})}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 h(\mathbf{x})}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 h(\mathbf{x})}{\partial x_2^2} \right)^2 \Big] \, dx_1 \, dx_2.$$

The minimizer of (5) for a cubic smoothing spline corresponds to a natural cubic spline with $n-2$ interior knots. If we use the B-spline basis to represent it, i.e., $h(x) = \sum_{j=1}^{n+2} \gamma_j B_j(x)$, equation (5) can be rewritten in a matrix format as

$$(\boldsymbol{r} - \mathbf{B}\boldsymbol{\gamma})^T (\boldsymbol{r} - \mathbf{B}\boldsymbol{\gamma}) + \lambda \boldsymbol{\gamma}^T \boldsymbol{\Omega} \boldsymbol{\gamma}, \tag{6}$$

where $\boldsymbol{r} = (r_1, \ldots, r_n)^T$, $\mathbf{B}_{ij} = B_j(x_i)$ and $\boldsymbol{\Omega}_{jl} = \int B_j''(x) B_l''(x) dx$, and $\boldsymbol{\gamma}$ is the coefficient vector of the $n+2$ basis functions, to be estimated. After taking the derivative of (6) with respect to $\boldsymbol{\gamma}$ and setting the derivative equal to zero, we have $(\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega})\hat{\boldsymbol{\gamma}} = \mathbf{B}^T\boldsymbol{r}$. Let $\mathbf{M} = (\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega})$, and calculate its Cholesky decomposition $\mathbf{M} = \mathbf{L}\mathbf{L}^T$. Then, solving $\mathbf{L}\mathbf{L}^T\hat{\boldsymbol{\gamma}} = \mathbf{B}^T\boldsymbol{r}$ by back-substitution provides $\hat{\boldsymbol{\gamma}}$ and thereby $\hat{h}(x) = \sum_{j=1}^{n+2} \hat{\gamma}_j B_j(x)$ in $O(n)$ operations (Hastie and Tibshirani, 1990).

When using thin plate splines, the solution of (5) is equivalent to that of

$$\min ||\boldsymbol{r} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\Phi}\boldsymbol{\delta}||^2 + \lambda \boldsymbol{\delta}^T \boldsymbol{\Phi} \boldsymbol{\delta}, \quad \text{subject to } \mathbf{X}^T \boldsymbol{\delta} = \mathbf{0}. \tag{7}$$

where the $n \times 3$ matrix $\mathbf{X} = [\mathbf{1}_n; \boldsymbol{x}_1; \boldsymbol{x}_2]$ includes the unit vector of size n as its first column and the n observations for the two covariates as its second and third columns. The radial basis matrix $\boldsymbol{\Phi}$ is defined by $\boldsymbol{\Phi}_{ji} = \phi(||\mathbf{x}_j - \mathbf{x}_i||) = ||\mathbf{x}_j - \mathbf{x}_i||^2 \log||\mathbf{x}_j - \mathbf{x}_i||$ for $i, j = 1, \ldots, n$. The 3-dimensional vector $\boldsymbol{\beta}$ and the n-dimensional vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$ are, respectively, the coefficients associated with $\mathbf{X}$ and those associated with the radial basis functions, and both sets of coefficients need to be estimated.

Using as many basis functions as the number of data points could be computationally challenging when the data size n is considerably large. Different from the univariate spline problem that can be solved by $O(n)$ operations (as discussed above), the computations for the thin plate splines require $O(n^3)$ operations (Hastie and Tibshirani, 1990). To overcome the computational problem, Wood (2003) proposed the thin plate regression splines (TPRS). Although the name includes the term 'regression splines,' unlike other regression splines, TPRS does not require the selection of the knots. To improve the computational efficiency, TPRS uses only k eigenbasis functions ($k \ll n$) corresponding to the largest k eigenvalues of the basis matrix $\boldsymbol{\Phi}$, and consequently reduces the rank of the basis matrix significantly.

TPRS can be fitted as follows. First, by applying the eigen decomposition of $\boldsymbol{\Phi}$, we have $\boldsymbol{\Phi} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ where $\mathbf{D}$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\boldsymbol{\Phi}$ and arranged in a non-increasing order, i.e., $\mathbf{D}_{i,i} \geq \mathbf{D}_{i+1,i+1}$ for $i = 1, \ldots, n-1$. $\mathbf{U}$ is an orthogonal matrix whose columns are the eigenvectors ordered accordingly. Then, TPRS considers the first k columns of the matrix $\mathbf{U}$, denoted by $\mathbf{U}_k$, resulting in a rank k eigenbasis matrix $\boldsymbol{\Phi}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{U}_k^T$ where $\mathbf{D}_k$ is a $k \times k$ diagonal matrix taking the first k rows and columns of $\mathbf{D}$. Subsequently, the constraint in (7) can be dropped by using QR decomposition on $\mathbf{U}_k^T \mathbf{X}$, i.e., $\mathbf{U}_k^T \mathbf{X} = \mathbf{Q}\mathbf{R}$ where $\mathbf{Q}$ is a $k \times k$ orthogonal matrix and $\mathbf{R}$ is a $k \times 3$ upper triangular matrix. Let $\mathbf{Z}_k$ take the last $k - 3$ columns of $\mathbf{Q}$. By restricting $\boldsymbol{\delta} = \mathbf{U}_k \mathbf{Z}_k \boldsymbol{\delta}_k$ with $(k-3)$-dimensional coefficient vector $\boldsymbol{\delta}_k$, the rank k approximation then can be used to fit TPRS by solving

$$\min ||\boldsymbol{r} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}_k \mathbf{D}_k \mathbf{Z}_k \boldsymbol{\delta}_k||^2 + \lambda \boldsymbol{\delta}_k^T \mathbf{Z}_k^T \mathbf{D}_k \mathbf{Z}_k \boldsymbol{\delta}_k, \tag{8}$$

for the unknown $\boldsymbol{\beta}$ and $\boldsymbol{\delta}_k$. Then, prediction for any given $\mathbf{x}$ can be achieved by calculating

$\hat{\boldsymbol{\delta}} = \mathbf{U}_k \mathbf{Z}_k \hat{\boldsymbol{\delta}}_k$ and plugging $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\beta}}$ into

$$\hat{h}(\mathbf{x}) = \mathbf{X}\hat{\boldsymbol{\beta}} + \sum_{i=1}^{n} \hat{\delta}_i \phi(||\mathbf{x} - \mathbf{x}_i||). \tag{9}$$

Recall that we assume the wake power loss term ω_t is non-negative so that our model can be consistent with the physical understanding of the wake effect, but the modeling procedure of TPRS does not guarantee non-negativity. In order to make sure the wake power loss is indeed non-negative, we apply an exponential transformation on top of the conventional TPRS estimation in (9), so we have

$$\hat{\omega}(\mathbf{x}) = \exp\{\mathbf{X}\hat{\boldsymbol{\beta}} + \sum_{i=1}^{n} \hat{\delta}_i \phi(||\mathbf{x} - \mathbf{x}_i||)\}. \tag{10}$$

Because of this change, instead of (8), we solve

$$\min \; ||\boldsymbol{r} - \exp\{\mathbf{X}\boldsymbol{\beta} + \mathbf{U}_k \mathbf{D}_k \mathbf{Z}_k \boldsymbol{\delta}_k\}||^2 + \lambda \boldsymbol{\delta}_k^T \mathbf{Z}_k^T \mathbf{D}_k \mathbf{Z}_k \boldsymbol{\delta}_k, \tag{11}$$

with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\delta}_k$.

One may argue that, after the exponential transformation, the penalty term in (11) needs to be re-derived by following the formulation of (5) and calculating the second derivatives of the new exponential term in (10). Doing so would make the regularized learning formulation in (11) messier but the benefit is marginal (explained below). Recall that the role of the penalty term is to impose a certain degree of smoothness on the final estimate for better prediction. By using the same penalty term as in (8), we understand that the balance between the goodness-of-fit and the smoothness may not be optimal. Regardless of that, however, the optimality would not be attained anyway since we use the basis truncation to speed up the computation (i.e., the use of k eigenbasis functions). We may not be worse off with all these approximations, as we choose the smoothing parameter that provides the best prediction via cross validation and the continuous transformation through the exponentiation only makes the final fit smoother. Due to these considerations, the treatment we apply here was in fact advocated previously in smoothing spline research (Ramsay and Silverman, 2005) as an effective way to handle non-negativity.

In general, when estimating a GAM, a constant term precedes functional terms, and it is estimated by the global mean. Specifically, the global mean is calculated and subtracted from the response in advance, before implementing the backfitting algorithm that will

estimate the rest of the functional terms. In the power difference model (4), this constant term should be part of the turbine-difference term, $\tilde{\eta}(\cdot)$, meaning that a portion of the turbine difference is constant regardless of the wind conditions, while the other portion may change with the wind speed. For the implementation of the backfitting algorithm, we can re-write equation (4) as

$$\tilde{y} = \alpha + [\tilde{\eta}(V) - \alpha] - \omega_1(V, D) \cdot \mathbb{1}_{\mathcal{D}_1}(D) + \omega_2(V, D) \cdot \mathbb{1}_{\mathcal{D}_2}(D) + \tilde{\epsilon}, \tag{12}$$

and estimate α using the global mean and $[\tilde{\eta}(V) - \alpha]$ using a cubic smoothing spline (and the wake loss terms using TPRS-N). Once all the functional terms are estimated, $\tilde{\eta}(V)$ is restored by $\hat{\alpha} + \hat{\eta}(V)$ where $\hat{\eta}(V)$ is the estimate of $[\tilde{\eta}(V) - \alpha]$.

Before implementing the backfitting algorithm, some tuning parameters need to be set, including the smoothing parameter λ and the value of the reduced rank k used for improving the computational efficiency of TPRS-N. There are in fact three smoothing parameters λ, one for each smooth function estimation, associated with $\tilde{\eta}(\cdot)$ and the two $\omega(\cdot)$'s, respectively. They are chosen based on a 10-fold cross validation while applying grid search. For the reduced rank k, Wood (2003) stated that the choice of k is not so critical as long as it is larger than the degrees of freedom required for the estimation. In the subsequent analysis sections, we set $k = 30$ which, we believe, is large enough for the wake effect analysis application based on our graphical inspection of the estimation results (see Appendix A for further discussion). Finally, we set a threshold τ that determines the convergence of the model fitting to 0.1 which is a sufficiently small number considering the magnitude of the functional estimates changing exponentially due to the imposition of non-negativity.

The backfitting algorithm for the power difference model is summarized in Algorithm 1.

4. Performance comparison of different wake models

In this section, we compare the proposed spline-based wake model with other wake models in terms of the prediction error of the power difference. Because measuring the actual wake power loss directly is extremely difficult, if not impossible, the prediction or estimation of the power difference becomes an important proxy alluding to the model capability of accounting for wake effects in wind power production. The same proxy was used by You et al. (2017) for evaluating their wake model. Furthermore, power difference prediction is in and by itself useful in a number of wind energy applications, for instance,

Algorithm 1 Backfitting algorithm for wake power loss estimation.

Initialize:
$m \leftarrow 0$; $\hat{\alpha} \leftarrow \sum_{i=1}^{n} y_i/n$; $\hat{\boldsymbol{\eta}}^m \leftarrow \mathbf{0}$; $\hat{\boldsymbol{\omega}}_1^m \leftarrow \mathbf{0}$; $\hat{\boldsymbol{\omega}}_2^m \leftarrow \mathbf{0}$.
repeat
  Set $m \leftarrow m + 1$.
  Estimation of $\hat{\boldsymbol{\eta}}$
    Calculate partial residuals: $\boldsymbol{r}_\eta \leftarrow \boldsymbol{y} - \hat{\alpha} + \hat{\boldsymbol{\omega}}_1^{m-1} - \hat{\boldsymbol{\omega}}_2^{m-1}$.
    Set $\hat{\boldsymbol{\eta}}^m$ by fitting smoothing spline to $\boldsymbol{r}_\eta$ with respect to $\boldsymbol{V}$.
  Estimation of $\hat{\boldsymbol{\omega}}_1$
    Calculate partial residuals: $\boldsymbol{r}_{\omega_1} \leftarrow -(\boldsymbol{y} - \hat{\alpha} - \hat{\boldsymbol{\eta}}^m - \hat{\boldsymbol{\omega}}_2^{m-1})$.
    Set $\hat{\boldsymbol{\omega}}_1^m$ by fitting thin plate regression spline with non-negativity to $\boldsymbol{r}_{\omega_1}$ with respect to $\boldsymbol{V}$ and $\boldsymbol{D}$ for the data whose $D \in \mathcal{D}_1$.
  Estimation of $\hat{\boldsymbol{\omega}}_2$
    Calculate partial residuals: $\boldsymbol{r}_{\omega_2} \leftarrow \boldsymbol{y} - \hat{\alpha} - \hat{\boldsymbol{\eta}}^m + \hat{\boldsymbol{\omega}}_1^m$.
    Set $\hat{\boldsymbol{\omega}}_2^m$ by fitting thin plate regression spline with non-negativity to $\boldsymbol{r}_{\omega_2}$ with respect to $\boldsymbol{V}$ and $\boldsymbol{D}$ for the data whose $D \in \mathcal{D}_2$.
  Computation of convergence criterion
    $$\Delta \leftarrow \frac{||\hat{\boldsymbol{\eta}}^m - \hat{\boldsymbol{\eta}}^{m-1}|| + ||\hat{\boldsymbol{\omega}}_1^m - \hat{\boldsymbol{\omega}}_1^{m-1}|| + ||\hat{\boldsymbol{\omega}}_2^m - \hat{\boldsymbol{\omega}}_2^{m-1}||}{||\hat{\boldsymbol{\eta}}^{m-1}|| + ||\hat{\boldsymbol{\omega}}_1^{m-1}|| + ||\hat{\boldsymbol{\omega}}_2^{m-1}||}.$$
until $\Delta \leq \tau$ where τ is a pre-specified threshold.

to quantify the effect of a wind turbine upgrade action through the comparison of a pair of turbines that is preferably isolated from other turbines (Lee et al., 2015b).

For this model comparison, we use actual operational data collected from an onshore wind farm in the US. The wind farm houses more than 200 wind turbines and four meteorological mast towers spreading over a relatively large area. From this wind farm, we take six pairs of wind turbines (in total, 12 wind turbines) into consideration. The turbine pairs are chosen such that no other turbines except the pair are located within 10 times the rotor diameter d. Such arrangement is to find a pair of turbines that are free of other turbine's wake, so that the wake analysis result can be reasonably attributed to the wake of its pair turbine. Ammara et al. (2002) state, theoretical and experimental studies have generally suggested that wake velocity deficit is minimal beyond $10d$ downstream of a wind turbine. As such, we expect the wake power loss due to other turbines becomes virtually negligible by the $10d$ restriction. Meanwhile, Meyers and Meneveau (2012) state that turbine spacing of $7d$ is conventionally used in wind farm implementations. This suggests that $10d$ separation can be applied to other wind farms in some cases and that the isolation of a turbine pair can be applied more frequently with a shorter distance restriction, for example, $7d$ (see discussions in Section 5).

Figure 6 shows the relative locations of the six pairs of turbines on the wind farm. The circle around each turbine indicates the $10d$ radius from the turbine. All turbine pairs happen to have the northwestern-to-southeastern orientation. So we designate the turbine on the northwestern side as Turbine 1 and the one on the southeastern side as Turbine 2 for all turbine pairs. Table 1 provides the between-turbine distances, in terms of a multiple of the rotor diameter, and the relative positional angles between a pair of turbines. Based on the specific relative positions between a pair of turbines and the notations illustrated in Figure 5, we can divide wind direction into two distinct sectors of $\mathcal{D}_1$ and $\mathcal{D}_2$ for each turbine pair. For a wind direction $D \in \mathcal{D}_2$, Turbine 1 is wake free and Turbine 2 is in the wake, whereas for $D \in \mathcal{D}_1$, Turbine 2 is wake free and Turbine 1 is in the wake.

We have the operational data for the six pairs of turbines, taken during roughly a year-long period between 2010 and 2011. The datasets include wind power output, wind speed, wind direction, air pressure, and temperature; air pressure and temperature data are used to calculate air density. The data provided to us are reported every 10 minutes, and they are the averages of the raw measurements calculated over distinct 10-minute time intervals, following the IEC standard (IEC12.1, 2005). Overall, it gives us approximately 33,500–38,500 data records per turbine pair for one year after accounting for the missing values and eliminating data records outside the normal operating range (i.e., with a negative power output).

The wind power outputs are measured on the turbine. The wind related and other environmental variables are measured at a nearby mast tower. Wind speed is also measured on the turbines; all other variables are measured at the mast tower only. Considering that

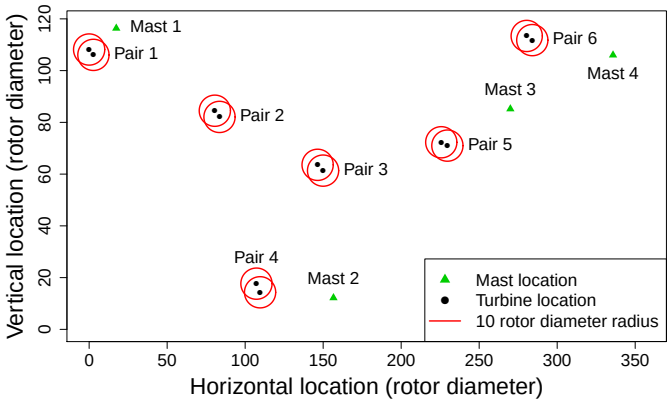

Figure 6: Locations of the six pairs of wind turbines and the four meteorological masts. The distances along both axes are expressed as a multiple of the rotor diameter of the turbines. All turbines have the same rotor diameter.

Table 1: Between-turbine distances and relative positions of the six pairs of turbines. Bearing 1 to 2 indicates a relative direction of Turbine 1 to the location of Turbine 2, and Bearing 2 to 1 is similarly defined.

	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6
Between-turbine distance	$3.4d$	$3.8d$	$4.2d$	$4.1d$	$4.1d$	$3.9d$
Bearing 1 to 2 (°)	307.1	308.7	302.6	325.0	288.3	294.2
Bearing 2 to 1 (°)	127.1	128.7	122.6	145.0	108.3	114.2

there are only four mast towers on this wind farm and that some of the turbine pairs are relatively far away from any of the mast towers, we decide to use the wind speed measurements obtained at the turbines. For wind direction, air pressure, and temperature, the mast measurements are used as there is no other option. Of course, we take the data from the mast closest to a turbine pair. Specifically, we use the data from Mast 1 for Pairs 1 and 2, Mast 2 for Pairs 3 and 4, and Mast 3 for Pair 5 and 6.

Note that the wind speed measurements on a turbine is obtained after the rotor, namely in its own wake. IEC 61400-12-2 (IEC12.2, 2013) recommends dealing with this issue through a nacelle transfer function (NTF), which describes the relation between the free inflow wind speed and that measured at the turbine anemometer in the wake. The wind farm operator who provided us the data has informed us that the turbine wind speed measurements have been adjusted through a NTF and can be treated as if they were measured in front of the rotor.

We evaluate the performance of a model with respect to its out-of-sample prediction errors. For this, we split each turbine pair's annual data into training and testing subsets by a ratio of 80:20. In other words, we use randomly selected 80% of a given dataset to train the model and the remaining 20% to calculate the prediction error. To measure the prediction error, we use the root mean square error (RMSE) and mean absolute error (MAE) defined, respectively, by

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\tilde{y}_i - \hat{\tilde{y}}_i)^2}, \qquad \text{MAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |\tilde{y}_i - \hat{\tilde{y}}_i|,$$

where $\tilde{y}_i$ for $i = 1, \ldots, n_{\text{test}}$ is the power difference calculated for the testing dataset, $\hat{\tilde{y}}_i$ is the corresponding prediction of the power difference, and n_{test} is the number of data pairs in a test set, typically between 6,000 and 8,000 in amount.

We compare the proposed method with four other methods: a physics-based method

and three data-driven models. The physics based method is Jensen's model (Jensen, 1983). The three data-driven models are the data binning approach (Barthelmie et al., 2010), the IND model, and the GMRF wake model (You et al., 2017). We choose Jensen's model as a representative of the physics-based models because it is widely used in the commercial operation environments. The CFD-based models are limited to restrictive research settings that have access to supercomputing capability, and these models are computationally prohibitive even for the analysis of a pair of turbines when they are applied to commercial sized wind turbines in actual wind farm operations.

To implement these methods, certain tuning parameters need to be set. The Jensen's model has a wake decay constant. We set this value to 0.075, the value commonly used for onshore wind farms (WAsP, 2016). The implementation of the physics-based model additionally requires power curve estimation for which we followed the standard procedure described in IEC12.1 (2005). For the data binning approach, instead of using a subset of data with a restricted range of wind speed and wind direction, we simply chose to use the entire dataset. This extension is straightforward, and the result of using the whole dataset is better in terms of RMSE than using the restrictive subset. We generated wind direction bins using a bin width resolution of 5° following Barthelmie et al. (2010). For the IND model, which is an additive B-spline model taking wind speed and turbulence intensity as covariates, we have to determine the number and locations of knots for each covariate. We followed the suggestions in You et al. (2017) and used equidistant knots covering the range of each covariate. For the GMRF model, the authors of You et al. (2017) have generously implemented their method on our data (all six turbine sets) and reported the resulting RMSE and MAE values to us.

Table 2 and 3, respectively, present the RMSE and MAE values for the five methods and six turbine pairs, denoted as percentage values of the maximum power output. The proposed spline-based model is labeled as 'GAM with TPRS-N' in the table. The Jensen's model, due to its simplicity and additional errors induced by the use of a power curve based conversion, leaves a relatively large portion of variation in the original data unexplained and hence registers the highest RMSE values. Relative to the Jensen's model, all data-driven methods except the IND model significantly reduce the level of uncertainty by accounting for the variation observed in the data. The IND model, without explicitly accounting for the effect of wind direction, performs only slightly better than the Jensen's model, sometimes exhibiting even higher MAE values. This demonstrates the importance of modeling wind direction as an influential covariate.

Recall that the GMRF model was not specifically developed for the single wake situation. By construction, the GMRF model is designed to perform better with more turbines since it benefits from the spatial modeling of multiple turbines at different locations. Understandably, the method loses its benefits when applied to a pair of turbines. Still, the method shows significant improvement with an 18% reduction in RMSE and 14% in MAE, on average, as compared to the Jensen's model.

The data binning approach, while fitting the trend of data without any restriction, in fact attains competitive prediction errors. This should not come as a surprise, as the binning approach is an extreme version of a localized fitting method and can adapt to local data features, as long as one uses a small enough binning resolution and there are dense enough data points to fit. The data binning approach is often competitive or the best method in terms of out-of-sample prediction when comparing with other fitting methods. The fact that its RMSE and MAE are in fact larger than those of our proposed spline-based model suggests that the data binning approach overfits the (training) data. We also reiterate that the binning approach is less insightful at providing wake characteristic, as we argued in an early section. In addition, the data binning approach will run into a dimensionality issue rather quickly, if there are more variables to be binned other than the wind direction, an issue that has been discussed at length in Lee et al. (2015b).

Our proposed model demonstrates its superiority over other alternatives in terms of the prediction error of the power difference. It yields the smallest RMSE values across all six turbine pairs and the smallest MAE values for five among the six pairs. Its RMSE (MAE) is, on average, 30% (24%) smaller than that of the Jensen's model, 26% (21%) smaller than the IND model, 15% (12%) smaller than the GMRF model, and 6% (7%) smaller than the data binning approach.

Table 2: Comparison of prediction error in terms of RMSE. The value in the table is the percentage of power difference relative to the maximum power of the turbine. The boldface values are the smallest in each column.

	RMSE					
	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6
Jensen's model	11.03%	8.87%	11.09%	9.71%	9.56%	10.20%
IND model	10.77%	8.52%	10.05%	9.31%	8.54%	9.85%
GMRF model	8.46%	7.52%	8.88%	7.97%	7.98%	8.77%
Data binning approach	7.78%	6.67%	8.18%	8.00%	7.06%	7.51%
GAM with TPRS-N	**6.68**%	**6.27**%	**8.02**%	**7.58**%	**6.83**%	**6.99**%

Table 3: Comparison of prediction error in terms of MAE. The value in the table is the percentage of power difference relative to the maximum power of the turbine. The boldface values are the smallest in each column.

	MAE					
	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6
Jensen's model	5.44%	5.30%	6.73%	5.70%	6.31%	5.65%
IND model	5.52%	5.19%	6.23%	5.42%	5.41%	5.81%
GMRF model	4.84%	4.69%	5.68%	**4.70**%	4.97%	5.44%
Data binning approach	4.34%	4.35%	5.32%	5.04%	4.89%	4.57%
GAM with TPRS-N	**3.75**%	**4.08**%	**5.23**%	4.77%	**4.47**%	**4.34**%

5. Analysis of wind turbine wakes in actual operations of wind turbines

In this section, we quantify annual wake power loss in actual wind turbine operations. Quantification of the wake power loss based on an annual period supports economic assessment of wake effect in terms of annual energy production (AEP), a key performance metric that is contractual binding and recommended by IEC (IEC12.1, 2005). Doing so also provides practical insights into the economic impact of decisions and actions attempting to alleviate the wake power loss.

To quantify the annual wake power loss and derive the wake characteristics revealed during an annual period, we apply the proposed method to the entire year-long dataset. Figure 7 illustrates the fitted wake effect. By our estimation, the wake loss is strictly positive, but what we show in the plot is actually $-\hat{\omega}_1(V, D) \cdot \mathbb{1}_{\mathcal{D}_1}(D) + \hat{\omega}_2(V, D) \cdot \mathbb{1}_{\mathcal{D}_2}(D)$, so that one sees both positive and negative portions. The power difference of some pairs of turbines, when plotted against wind direction, exhibits large variation with several peaks and troughs. Even under such a noisy circumstance, our model captures the wake power loss signals well, by focusing on where the wake power loss is expected. In the figure, the vertical dashed lines indicate the bearings, i.e., $\theta_1 = 0$ and $\theta_2 = 0$. Comparing Figure 7(e) to Figure 4 (generated from Pair 5), it is obvious that our wake loss estimation method captures the signals much better than the data binning approach could, making the subsequent derivation of the wake characteristics more convincing. We also observe from Figure 7 that the wind direction associated with the highest power loss is not exactly aligned with the bearings of the turbine pairs. This implies that there are measurement errors in wind direction. When applying the data binning approach, practitioners typically generate angle bins starting from a bearing by making it as the midpoint of an angle bin (and propagate with a resolution of 5 degrees, for example) and then regard the wake loss estimate of this

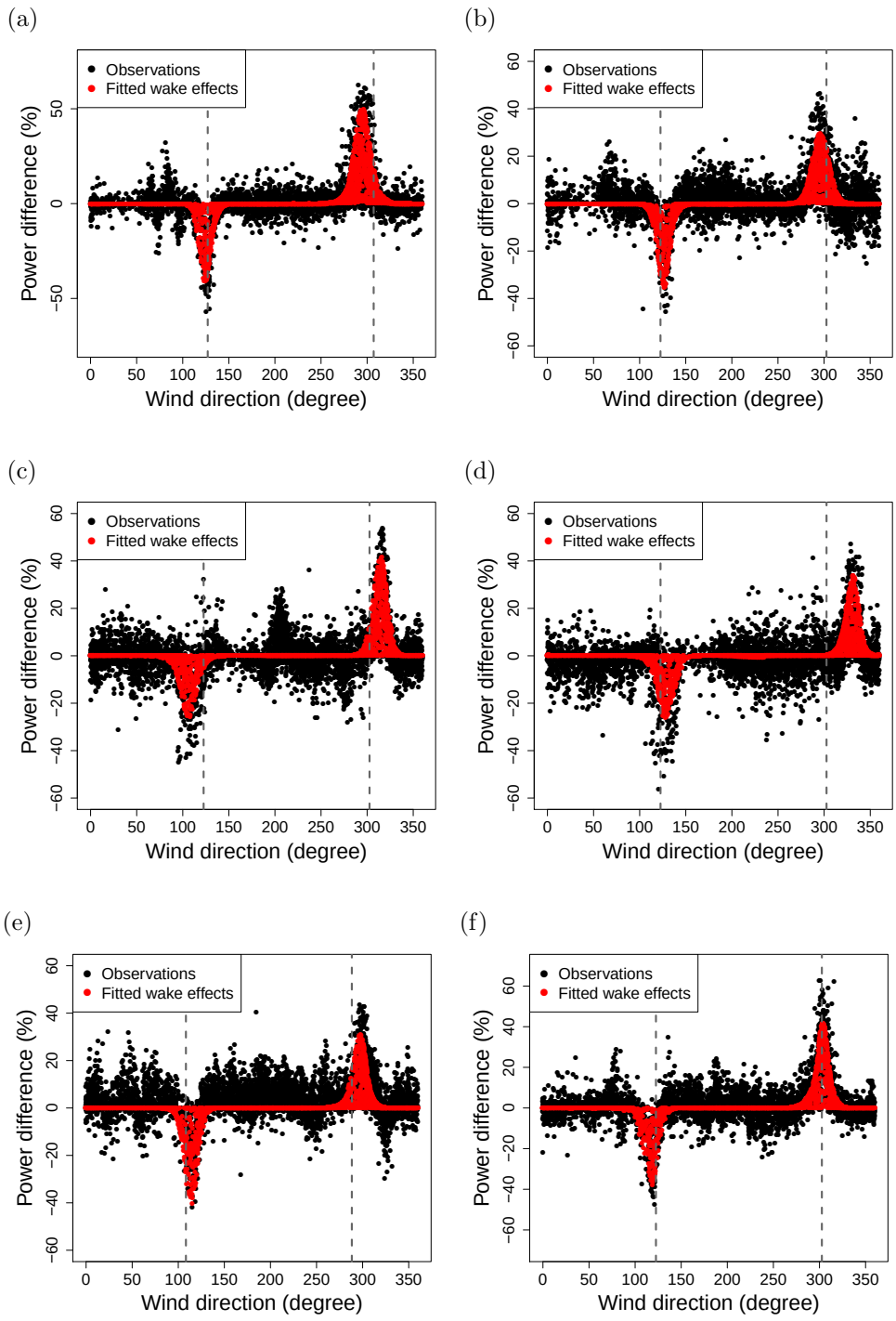

Figure 7: Estimated wake effects using GAM with TPRS-N: (a) for Pair 1; (b) for Pair 2; (c) for Pair 3; (d) for Pair 4; (e) for Pair 5; (f) for Pair 6. The red shaded areas represent the fitted wake loss in terms of $-\hat{\omega}_1(V,D)\cdot\mathbb{1}_{\mathcal{D}_1}(D)+\hat{\omega}_2(V,D)\cdot\mathbb{1}_{\mathcal{D}_2}(D)$. Two dashed vertical lines indicate wind direction that is parallel to the line connecting the pair of turbines.

Table 4: Wake depth and width for the six pairs of turbines.

	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6
Depth: Turbine 1	41.7% (59.1%)	35.4% (44.0%)	26.3% (32.9%)	26.8% (43.8%)	40.2% (47.0%)	39.5% (43.1%)
Depth: Turbine 2	50.7% (58.2%)	29.8% (34.5%)	42.4% (48.7%)	33.8% (42.2%)	31.0% (44.5%)	42.0% (49.8%)
Width: Turbine 1	40.1°	42.7°	53.1°	52.2°	41.1°	53.4°
Width: Turbine 2	61.8°	47.9°	44.6°	43.6°	46.8°	49.6°

specific bin as the wake depth. It turns out that, in the presence of measurements errors in wind direction, such a practice has an obvious disadvantage and will surely underestimate the wake depth due to the discrepancy between a bearing and the actual wind direction with the highest wake loss (see Figure 7(d), for an extreme example).

Table 4 shows the wake characteristics for the six turbine pairs. The first two rows are the wake depths, namely the magnitude of the wake power losses. The last two rows are the wake widths. The wake depth is identified as the peak of the wake loss estimate representing the maximum power loss. The wake width is supposed to be determined by the angles around the bearings at which the power loss eventually becomes zero. However, given noisy signals spreading over a large range of wind directions, the fitted wake power loss is not completely zero even in the regions where it is unquestionably wake free. To estimate the wake width, we therefore find the range of wind direction for which loss is greater than 1% of the rated power of the turbine.

For the wake depth, Table 4 presents two percentage values for each turbine. The one outside the parenthesis is the wake power loss relative to the rated power of that turbine whereas the one inside the parenthesis is the loss relative to the free stream equivalent power output. Recall that throughout the paper, we have not shown the actual power values due to the confidentiality agreement in place but have shown the normalized power values, in a percentage, relative to the rated power. These turbines belong to the general 2 MW turbine class. It does not mean that the rated power of the turbines is exactly 2 MW but it is in that vicinity. Using this information, one can estimate the wake power loss magnitude in the unit of mega-watts.

In the literature, however, the wake power loss is often expressed as the ratio of the loss over the free stream equivalent power output (Barthelmie and Jensen, 2010; Hansen et al., 2012; Adaramola and Krogstad, 2011), so we have

$$\frac{\hat{\omega}_t(V_i, D_i)}{\hat{y}_t(V_i, D_i) + \hat{\omega}_t(V_i, D_i)}, \quad t = 1, 2, \quad i = 1, \ldots, n,$$

where $\hat{y}_t(V_i, D_i)$ denotes the expected power generation given (V_i, D_i). Depending on

(V_i, D_i), $\hat{y}_t(V_i, D_i)$ could be the expected power in the wake of another turbine, so that we mean to recover the free stream equivalent power output by adding $\hat{y}_t(V_i, D_i)$ and $\hat{\omega}_t(V_i, D_i)$. To calculate $\hat{y}_t(V_i, D_i)$, we first define a neighborhood of (V_i, D_i), i.e., $\mathcal{N}_i = \{(V, D) : V \in (V_i - \epsilon_V, V_i + \epsilon_V], D \in (D_i - \epsilon_D, D_i + \epsilon_D]\}$ where ϵ_V and ϵ_D are predetermined constants. We set $\epsilon_V = 0.25$ m/s and $\epsilon_D = 2.5°$, following the common practice (IEC12.1, 2005; Barthelmie et al., 2010). Then, we calculate $\hat{y}_t(V_i, D_i)$ by taking the average of the power outputs whose corresponding wind speed and direction is a member of $\mathcal{N}_i$. This is a two-dimensional binning with $2\epsilon_V$ and $2\epsilon_D$ as the respective bin width. The second percentage values in Table 4, namely the ones inside the parenthesis, are the wake power loss expressed in this conventional fashion.

The peak power loss relative to the free stream equivalent (the value inside the parenthesis) ranges from 33% to 59%. The wake width for the 12 turbines ranges from 40° to 62° with concentration around 40°–53°. The wake depth commonly stated in the literature is in the range of 30%–40% (Barthelmie et al., 2010, 2009; Sanderse et al., 2011), which appears to be at the lower bound of our estimates. In addition, our wake width estimates are also noticeably larger than the 25° to 40° range stated previously (Barthelmie et al., 2010; McKay et al., 2013; Troldborg et al., 2011). We believe that the difference can be attributed to two major factors. The first one is that our estimation can identify the wake region more accurately, producing better estimates of the two main characteristics, whereas the methods in the literature rely on *ad hoc* data segmentation and partition and often use a partial set of data based on a pre-selected range of wind direction, and consequently, their wake power loss estimates do not capture the characteristics as well as our estimator does. The second factor is that the historical estimates are usually the averages over multiple turbines, understandably leading to a narrower range.

Table 5 shows how each term in the power difference model (4) affects the power generation of a turbine pair in an annual period, namely the AEP power difference or AEP loss. The first row is the between-turbine power production difference independent of wake effect, expressed relative to the rated power. The second and third rows present the wake loss. Similarly to Table 4, the values outside the parentheses is the loss relative to the rated power, whereas the values inside the parentheses is the loss relative to the free stream equivalent. Both percentages represent the AEP wake loss but use different baselines.

The wake loss relative to the rated power is in fact related to the capacity factor of a wind turbine (Wikipedia, 2017). Recall that the capacity factor is the ratio of the actual power production of a turbine for a selected period of time, say, one year, over the supposed power

Table 5: Annual power loss for the six turbine pairs.

	Percentage measure (%)					
	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6
Turbine difference ($\tilde{\eta}_{2-1}$)	0.22	1.10	0.18	-1.39	3.41	0.92
Wake loss: Turbine 1 (ω_1)	0.83 (1.64)	0.78 (1.62)	0.48 (1.08)	0.62 (1.36)	0.60 (1.30)	0.65 (1.31)
Wake loss: Turbine 2 (ω_2)	2.05 (4.13)	1.24 (2.68)	1.39 (3.15)	1.11 (2.39)	1.00 (2.34)	1.76 (3.69)
Average loss for the pair	1.44 (2.87)	1.01 (2.14)	0.94 (2.10)	0.86 (1.88)	0.80 (1.80)	1.20 (2.48)

production the turbine could have produced, had it operated at its maximum capacity (i.e., at the rated power) all the time; the typical range of the capacity factor is 25%–35%. The wake loss relative to the rated power, therefore, can be seen as the direct reduction to a turbine's capacity factor. We hereby refer to the corresponding AEP loss as the capacity factor AEP loss, and refer to the AEP loss relative to the free stream equivalent as the traditional AEP loss which is computed, if using Turbine 1 group as an example, by

$$\frac{\sum_{i=1}^{n} \hat{\omega}_1(V_i, D_i)}{\sum_{i=1}^{n} \{\hat{y}_1(V_i, D_i) + \hat{\omega}_1(V_i, D_i)\}}.$$

The fourth row is the average AEP loss for a pair of turbines. The average is weighted by the number of data points in the respective wake regions to account for the annual distribution of the AEP loss for the turbine pairs. For this reason, the values in the fourth row may be slightly different from the simple average of the two individual losses. The traditional AEP loss for a pair is computed by

$$\frac{\sum_{i=1}^{n} \{\hat{\omega}_1(V_i, D_i) + \hat{\omega}_2(V_i, D_i)\}}{\sum_{i=1}^{n} \{\hat{y}_1(V_i, D_i) + \hat{\omega}_1(V_i, D_i) + \hat{y}_2(V_i, D_i) + \hat{\omega}_2(V_i, D_i)\}}.$$

The average capacity factor AEP loss is computed by setting the denominator in the above equation to be $\sum_{i=1}^{n} \{(\text{rated power}) + (\text{rated power})\} = 2n \cdot (\text{rated power})$.

From Table 5, one may notice that the magnitude of the between-turbine difference is sizeable, sometimes even larger than that of the wake effect. This result suggests that modeling of the between-turbine difference as a separate term in the power difference model is important to our mission to estimate the wake effect; otherwise, the estimate of the wake effect can be biased considerably.

One can immediately observe that the AEP losses are much smaller than the peak power loss (wake depth). This is expected because the annual loss is the average over all kinds of wind speed and direction conditions in an entire year. Under many circumstances, the wake loss is much smaller than the peak loss. The capacity factor AEP loss is between 0.5–2.0%,

meaning that if the turbine's actual capacity factor is 25%, then its ideal capacity factor, if the turbine always operated wake free, could have been 25.5% to 27%. This difference, while appearing as a small percentage, should not be taken lightly. Consider a wind farm housing 200 turbines all in the 2 MW turbine class. A 1% capacity factor AEP loss for the whole farm translates to $1.3 million loss in revenue at the wholesale price of $37 per MWh (Statista, 2016).

One may also notice that the wake loss endured by Turbine 2 in a pair is always greater than that of Turbine 1. This can be explained by the relative positions of the turbines and the prevailing wind direction over this farm during that particular year. Figure 8 presents the wind rose plots for three pairs of the turbines. The plots show that the north-western wind, for which Turbine 2 of each pair endures power loss, is more frequent and stronger than the south-eastern wind for which Turbine 1 experiences power loss. Unsurprisingly, we observe the AEP loss of Turbine 1 group is usually less than 0.83% (1.64%), whereas the AEP loss for Turbine 2 group is greater than 1.00% (2.34%) and can be as high as 2.05% (4.13%).

In the literature, it is well-known that turbine spacing is a decisive factor affecting the magnitude of wake power loss (Laan et al., 2015; Sanderse et al., 2011; Barthelmie and Jensen, 2010). We therefore suspect that the variation of the annual power loss between the individual turbine pairs can be explained by the between-turbine distance of each pair. Using the average AEP loss for the six turbine pairs (the fourth row in Table 5) and the corresponding between-turbine distances, we fit a simple linear regression model as has

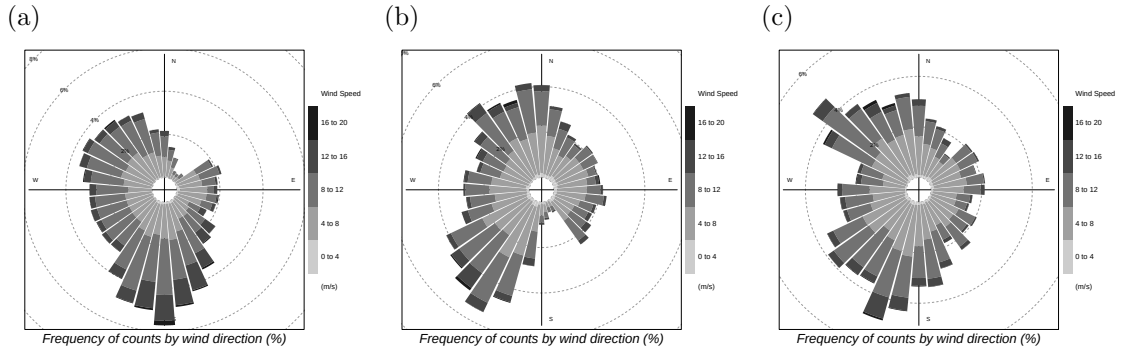

Figure 8: Wind rose plots illustrating the relative frequency of incoming wind for different direction sectors and for different speed ranges: (a) for Pair 1; (b) for Pair 3; (c) for Pair 5. These wind rose plots of Pair 1, 3, and 5 are representative of those of Pair 2, 4, and 6, respectively, because such group of two pairs shares the same mast that takes the wind direction measurements.

been done by Barthelmie and Jensen (2010). Figure 9 shows the scatter plots and the regression line fitting a respective AEP loss. For the capacity factor AEP loss, the p-values of the intercept and slope estimate are 0.005 and 0.013, respectively. For the traditional AEP loss, the corresponding p-values are 0.006 and 0.022. These results confirm that the turbine spacing indeed by and large explains the pair-wise difference in the AEP losses. Additionally, an extrapolation based on the fitted regression lines suggests that the wake loss would diminish after the turbine spacing reaches either $5.3d$ or $5.6d$, depending on which AEP loss is used in the analysis. Nevertheless, in either circumstance, the $10d$ separation used in this study to isolate a particular turbine pair from the rest of turbines appears safe enough to render the turbine pairs free of wake of any other turbine on the wind farm.

Regressing the turbines' inherent production difference (the first row in Table 5) on the between-turbine distance, on the other hand, suggests that there is no significant correlation between them. The p-values of the intercept and slope estimate in this case are 0.81 and 0.77, respectively, with R^2 of 0.02. As such, unlike the wake effect, the between-turbine production difference does not seem to be affected by the between-turbine distance. The obvious distinction of the regression results between the two cases supports that our model is capable of separating the power difference ($\tilde{y}$) into the between-turbine production difference part (i.e., $\tilde{\eta}$) and the wake effect part (i.e., ω_1 and ω_2), so that our wake effect estimate is well derived.

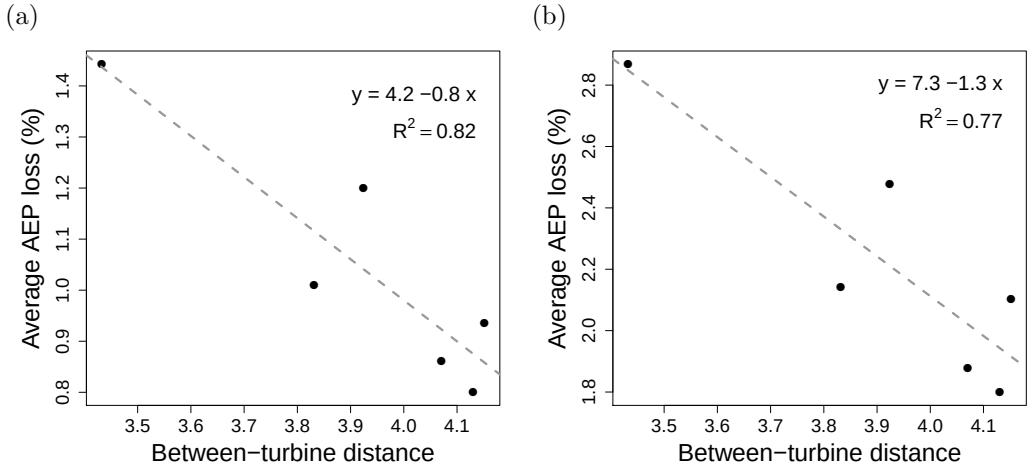

Figure 9: Relation between AEP losses and turbine spacing: (a) for the capacity factor AEP loss; (b) for the traditional AEP loss. The between turbine distance is expressed as a multiple of the rotor diameter.

6. Concluding remarks

Our study presents a data-driven wake effect model based on a spline model structure. A non-negativity constraint is incorporated in the model estimation to make sure that the estimate of wake power loss is consistent with the physical understanding. Our spline-based wake model produces the smallest prediction error when compared with one physics-based wake model and three data-driven wake models.

Nowadays, a systematic and effective way of combining physics-based models and data-driven models is to incorporate important constraints suggested by physical understanding into a data-driven model. Such constraints may include non-negativity, monotonicity, or convexity/concavity for various orders of derivatives of the response. Our proposed wake model is in fact in line with such a trend. We understand that there is no limit as to how physics-based models may be integrated with a data-driven model. We do want to take this opportunity to stress the importance and benefit of combining the physics-based models and data-driven models; much more needs to be done in future work.

Application of the proposed wake model is limited to the analysis of single wake behaviors due to its model structure established upon the pair-wise difference between two turbines' power output. The general power production function introduced prior to the pair-wise analysis does allow a potential extension to multiple turbines case at the wind farm level. Another possibility is to combine the strength of this spline wake model with that of You et al. (2017) and develop a wake model that accounts for spatial relationship among multiple turbines while imposing non-negativity on wake loss at the same time. Such data-driven models could be promising alternatives to complement complicated physics-based wake models for evaluating the wake power loss of a wind farm.

Acknowledgements

We thank You, Byon, Jin and Lee for implementing their approach and providing the prediction error assessment result. Hwangbo and Ding's research was partially supported by the National Science Foundation under grant no. CMMI-1300560.

References

Adaramola, M. S. and Krogstad, P.-Å. (2011), "Experimental investigation of wake effects on wind turbine performance," *Renewable Energy*, 36, 2078–2086.

Aigner, D., Lovell, C. K., and Schmidt, P. (1977), "Formulation and estimation of stochastic frontier production function models," *Journal of Econometrics*, 6, 21–37.

American Wind Energy Association (2016), "U.S. number one in the world in wind energy production," http://www.awea.org/MediaCenter/pressrelease.aspx?ItemNumber=8463, accessed: 2016-11-03.

Ammara, I., Leclerc, C., and Masson, C. (2002), "A viscous three-dimensional differential/actuator-disk method for the aerodynamic analysis of wind farms," *Journal of Solar Energy Engineering*, 124, 345–356.

Barthelmie, R. J., Hansen, K., Frandsen, S. T., Rathmann, O., Schepers, J. G., Schlez, W., Phillips, J., Rados, K., Zervos, A., Politis, E. S., et al. (2009), "Modelling and measuring flow and wind turbine wakes in large wind farms offshore," *Wind Energy*, 12, 431–444.

Barthelmie, R. J. and Jensen, L. E. (2010), "Evaluation of wind farm efficiency and wind turbine wakes at the Nysted offshore wind farm," *Wind Energy*, 13, 573–586.

Barthelmie, R. J., Pryor, S. C., Frandsen, S. T., Hansen, K. S., Schepers, J. G., Rados, K., Schlez, W., Neubert, A., Jensen, L. E., and Neckelmann, S. (2010), "Quantifying the impact of wind turbine wakes on power output at offshore wind farms," *Journal of Atmospheric and Oceanic Technology*, 27, 1302–1317.

DOE (2015), "Wind Vision: A New Era for Wind Power in the United States," Tech. rep., U.S. Deparment of Energy, Washington DC.

Duchon, J. (1977), "Splines minimizing rotation-invariant semi-norms in Sobolev spaces," in *Constructive theory of functions of several variables*, Springer, pp. 85–100.

Duckworth, A. and Barthelmie, R. J. (2008), "Investigation and validation of wind turbine wake models," *Wind Engineering*, 32, 459–475.

Emami, A. and Noghreh, P. (2010), "New approach on optimization in placement of wind turbines within wind farm by genetic algorithms," *Renewable Energy*, 35, 1559–1564.

Energy Information Administration (2016), "Word Pro - S7 - US Energy Information Administration," https://www.eia.gov/totalenergy/data/monthly/pdf/sec7_5.pdf, accessed: 2016-11-03.

Gebraad, P. M. O., Teeuwisse, F. W., Wingerden, J. W., Fleming, P. A., Ruben, S. D., Marden, J. R., and Pao, L. Y. (2016), "Wind plant power optimization through yaw control using a parametric model for wake effects—a CFD simulation study," *Wind Energy*, 19, 95–114.

Hansen, K. S., Barthelmie, R. J., Jensen, L. E., and Sommer, A. (2012), "The impact of turbulence intensity and atmospheric stability on power deficits due to wind turbine wakes at Horns Rev wind farm," *Wind Energy*, 15, 183–196.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, vol. 43, CRC Press.

Hwangbo, H., Johnson, A. L., and Ding, Y. (2015), "Power curve estimation: functional estimation imposing the regular ultra passum law," *Working Paper*, available at SSRN: http://ssrn.com/abstract=2621033.

IEC12.1 (2005), *IEC 61400-12-1 Ed 1, Wind Turbines - Part 12-1: Power Performance Measurements of Electricity Producing Wind Turbines*, Geneva, Switzerland: International Electrotechnical Commission.

IEC12.2 (2013), *IEC 61400-12-2, Wind Turbines - Part 12-2: Power Performance of Electricity Producing Wind Turbines Based on Nacelle Anemometry*, Geneva, Switzerland: International Electrotechnical Commission.

Jensen, N. O. (1983), "A note on wind generator interaction," Risø-m; no. 2411, Risø National Laboratory, http://orbit.dtu.dk/files/55857682/ris_m_2411.pdf.

Kusiak, A. and Song, Z. (2010), "Design of wind farm layout for maximum wind energy capture," *Renewable Energy*, 35, 685–694.

Laan, M. P., Sørensen, N. N., Réthoré, P.-E., Mann, J., Kelly, M. C., Troldborg, N., Schepers, J. G., and Machefaux, E. (2015), "An improved k-ϵ model applied to a wind turbine wake in atmospheric turbulence," *Wind Energy*, 18, 889–907.

Lee, G., Ding, Y., Genton, M. G., and Xie, L. (2015a), "Power curve estimation with multivariate environmental factors for inland and offshore wind farms," *Journal of the American Statistical Association*, 110, 56–67.

Lee, G., Ding, Y., Xie, L., and Genton, M. G. (2015b), "A kernel plus method for quantifying wind turbine performance upgrades," *Wind Energy*, 18, 1207–1219.

McKay, P., Carriveau, R., and Ting, D. S.-K. (2013), "Wake impacts on downstream wind turbine performance and yaw alignment," *Wind Energy*, 16, 221–234.

Meyers, J. and Meneveau, C. (2012), "Optimal turbine spacing in fully developed wind farm boundary layers," *Wind Energy*, 15, 305–317.

Nadaraya, E. A. (1964), "On estimating regression," *Theory of Probability & Its Applications*, 9, 141–142.

Prospathopoulos, J. M., Politis, E. S., Rados, K. G., and Chaviaropoulos, P. K. (2011), "Evaluation of the effects of turbulence model enhancements on wind turbine wake predictions," *Wind Energy*, 14, 285–300.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer-Verlag New York, 2nd ed.

Rasmussen, C. E. and Williams, C. K. (2006), *Gaussian Processes for Machine Learning*, The MIT Press.

Sanderse, B., Pijl, S. P., and Koren, B. (2011), "Review of computational fluid dynamics for wind turbine wake aerodynamics," *Wind Energy*, 14, 799–819.

Statista (2016), "Wind power wholesale power prices United States 2015 — Statistic," `https://www.statista.com/statistics/217841/us-cumulative-capacity-weighted-average-wind-power-price/`, accessed: 2016-12-28.

Troldborg, N., Larsen, G. C., Madsen, H. A., Hansen, K. S., Sørensen, J. N., and Mikkelsen, R. (2011), "Numerical simulations of wake interaction between two wind turbines at various inflow conditions," *Wind Energy*, 14, 859–876.

WAsP (2016), "Wind resources for energy production of wind turbines - WAsP," `http://www.wasp.dk/wasp`, accessed: 2016-12-11.

Watson, G. S. (1964), "Smooth regression analysis," *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Wikipedia (2017), "Capacity factor — Wikipedia, The Free Encyclopedia," `"https://en.wikipedia.org/wiki/Capacity_factor"`, accessed: 2017-01-19.

Wood, S. N. (2003), "Thin plate regression splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 95–114.

You, M., Byon, E., Jin, J., and Lee, G. (2017), "When wind travels through turbines: a new statistical approach for characterizing heterogeneous wake effects in multi-turbine wind farms," *IIE Transactions*, 49, 84–95.

Appendix

A. The effect of the reduced rank k on the prediction performance

In this section, we illustrate the effect of the reduced rank k on the prediction performance of the proposed estimator. We use the same datasets that were used in Section 4 and the same criteria of RMSE and MAE for quantifying the prediction accuracy. We trace the prediction performance of the spline method while changing the value of the reduced rank k.

Figure A.1 shows how the RMSE and MAE vary with different k values, where we use five different values of $k = 10, 20, 30, 40, 50$. The prediction tends to be more accurate as k increases, but the additional benefit of using a large k diminishes once k reaches a certain level. When $k = 20$, the values of both RMSE and MAE become quite comparable to those for a higher k value. As such, we believe that $k = 30$ is a large enough choice for the rank reduction purpose.

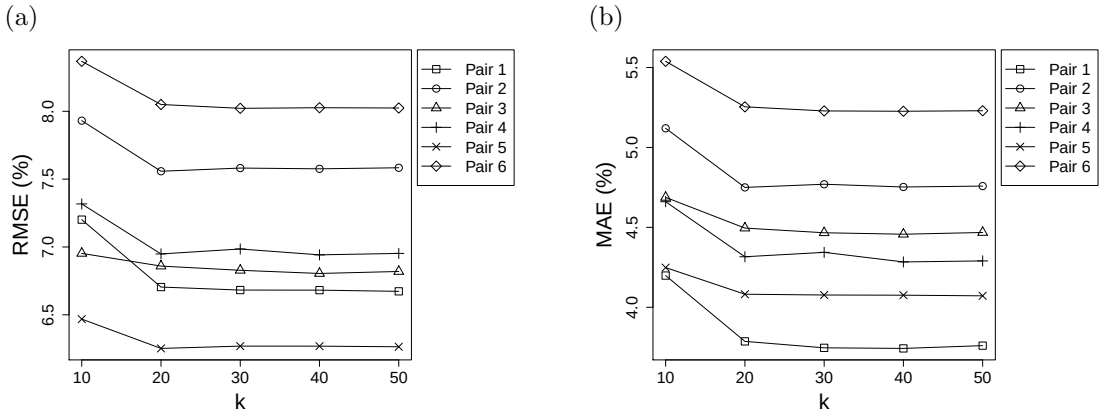

Figure A.1: Prediction performance varying with different k's: (a) in terms of RMSE; (b) in terms of MAE.