Very-Short-Term Probabilistic Wind Power Forecasts by Sparse Vector Autoregression

Jethro Dowell, *Student Member, IEEE*, Pierre Pinson, *Senior Member, IEEE*

***Abstract*—A spatio-temporal method for producing very-short-term parametric probabilistic wind power forecasts at a large number of locations is presented. Smart grids containing tens, or hundreds, of wind generators require skilled very-short-term forecasts to operate effectively, and spatial information is highly desirable. In addition, probabilistic forecasts are widely regarded as necessary for optimal power system management as they quantify the uncertainty associated with point forecasts. Here we work within a parametric framework based on the logit-normal distribution and forecast its parameters. The location parameter for multiple wind farms is modelled as a vector-valued spatio-temporal process, and the scale parameter is tracked by modified exponential smoothing. A state-of-the-art technique for fitting sparse vector autoregressive models is employed to model the location parameter and demonstrates numerical advantages over conventional vector autoregressive models. The proposed method is tested on a dataset of 5 minute mean wind power generation at 22 wind farms in Australia. 5-minute-ahead forecasts are produced and evaluated in terms of point and probabilistic forecast skill scores and calibration. Conventional autoregressive and vector autoregressive models serve as benchmarks.**

***Index Terms*—Probabilistic forecasting, wind power, power system operations, renewable energy.**

I. Introduction

THE large-scale integration of wind power presents operational challenges for both power systems [1] and electricity markets [2] due to the stochastic nature of the wind itself. As a result, the reliable and economic operation of power systems with high wind penetration depends on wind power forecasts, not least in the smart grid paradigm of distributed and highly interconnected generation. Applications of very-short-term forecasts include balancing and the optimal operation of reserves [3], and wind farm control [4]. Furthermore, the stochastic nature of the wind and complexity of the problem calls for a spatio-temporal probabilistic treatment in order to make optimal decisions under inherent uncertainty.

A review of the state-of-the-art in short-term (<12 hours) wind power forecasting can be found in [5] and [6]; of particular relevance to this work is the conclusion that for forecast horizons of less than approximately 6 hours statistical methods using local information are superior to physical models (i.e. numerical weather predictions), which require hours of computation time and introduce imprecision as a result of spatial interpolation. Such statistical methods are typically non-spatial (for individual locations), examples include autoregressive modelling [7], Markov chains [8] and data mining [9], among others, plus various hybrid approaches such as [10], [11]. Examples of very-short-term (<1 hour) forecasting include Markov switching [12] and parametric probabilistic forecasting [13], both of which rely on autoregressive techniques.

Several spatial predictors have been proposed to capitalise on the spatio-temporal relationship between wind power generation at a few wind farms in a small region. Spatial correlation between wind speed and direction has been exploited in [14] by regressing on different spatial information depending on the wind direction, and in [15] by fitting vector autoregressive-type models. Using multiple wind farms as 'spatial sensors' was shown to improve wind power forecast skill at a target site in [16]. Recent contributions have sought to build efficient probabilistic spatial models with sparse Gaussian random fields, but are limited to modest spatial dimension [17], [18]. However, with the abundance of wind farms on many power systems today it is desirable to build a spatial predictor for tens, or hundreds, of wind farms, making computational cost and automated model fitting serious considerations.

In this paper we present a single predictor for very-short-term probabilistic forecasting on large, previously intractable, spatial scales. The model fitting procedure is completely data driven making it ideal for smart grid applications where many generators share a single, highly interconnected power system and capturing spatial dependence is desirable. We combine two state-of-the-art statistical techniques: a parametric probabilistic framework based on the logit-normal distribution, as in [13], [19], and model the location parameter of that distribution as a sparse vector autoregressive process [20]. Further, we propose a novel exponential smoothing scheme with dynamic forgetting factor to track the scale parameter and compare it to the boundary weighted scheme described in [13].

The framework for producing spatial probabilistic forecasts based on the logit-normal distribution and transformation is outlined in Section II. The spatio-temporal modelling of the location parameter and the procedure for fitting sparse vector autoregressive models are described Section III. The tracking of the scale parameter is addressed in Section IV. In Section V the proposed method is tested on to a case study of 22 wind farms in southeastern Australia and results are presented and discussed. Conclusions are drawn in Section VI.

II. Spatial Probabilistic Forecast Framework

The power generated by a wind farm at any given time is bounded between zero, when no turbines are operating, and nominal, when all turbines are generating their rated power output. As a result, wind power cannot be directly

J. Dowell is with the University of Strathclyde's Wind Energy Systems Centre for Doctoral Training, Glasgow, UK (email: jethro.dowell@strath.ac.uk).

P. Pinson is with the Centre for Electric Power and Energy, Technical University of Denmark, Kgs. Lyngby, Denmark (email: ppin@elektro.dtu.dk).

Manuscript received...

modelled using conventional unbounded Gaussian distributions. Truncated Gaussian, censored Gaussian and generalised logit-normal distributions have all been proposed to model the conditional density of wind power motivated by the desire to work in a linear Gaussian framework [13]. In what follows, data are normalised by their corresponding nominal power such that they occupy the range $[0, 1]$.

In the proceeding derivation we assume logit-normal distributed wind power observations and transform the measurement data along the lines of [13]. The complete distribution is a discrete-continuous mixture of the logit-normal distribution with the possibility of probability masses on the bounds of the interval $[0, 1]$.

The logit-normal transformation is given by

$$y = \gamma(x) = \ln\Big(\frac{x}{1-x}\Big) \ , \quad x \in (0,1) \ , \tag{1}$$

with inverse

$$x = \gamma^{-1}(y) = \Big(1 + e^{-y}\Big)^{-1} \ , \quad y \in \mathbb{R} \quad . \tag{2}$$

Assuming that the variable X is logit-normal distributed, the transformed variable $Y = \gamma(X)$ is normally distributed. The logit-normal distribution has density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x(1-x)} \exp\Big[-\frac{1}{2}\Big\{\frac{\gamma(x)-\mu}{\sigma}\Big\}^2\Big] \ , \tag{3}$$

where location and scale parameters μ and σ^2 are directly connected the mean and variance of $Y \sim N(\mu, \sigma^2)$. The location parameter can be interpreted as the expected value of wind power, and the scale parameter as a measure of spread.

Consider now the stochastic process $\{X_t\}$ and its transformation $\{Y_t\}$ with realisations $\{x_t\}$ and $\{y_t\}$, respectively. The full predictive distribution of X_t, including probability masses on the bounds, is given by the sum of the logit-normal distribution, $L(\mu_t, \sigma_t^2)$, and probability masses w_t^0 and w_t^1 corresponding to zero and nominal power, respectively. It is written as

$$X_t \sim \delta_0 w_t^0 + \delta_1 w_t^1 + (1 - w_t^0 - w_t^1) L(\mu_t, \sigma_t^2) \ , \tag{4}$$

with

$$\begin{aligned} w_t^0 &= \Phi\Big\{\frac{\gamma(\eta) - \mu_t}{\sigma_t}\Big\} \ , \\ w_t^1 &= 1 - \Phi\Big\{\frac{\gamma(1-\eta) - \mu_t}{\sigma_t}\Big\} \ , \end{aligned} \tag{5}$$

where δ_x is the Dirac delta function at x, Φ is the cumulative distribution function of a standard normal variable, and η is the order of the measurement precision. Wind power values less than η, or greater than $1 - \eta$, are considered to be 0 or 1, respectively. The key result is that the predictive density of $\{X_t\}$ is parametrized by the conditional mean and variance of $\{Y_t\} \sim N(\mu_t, \sigma_t^2)$ only.

In order to calculate density forecasts for the wind power at some future time, $\{X_{t+k}\}$, we need only forecast the location and scale parameters of the predictive distribution, which are the mean and variance of the transformed process $\{Y_{t+k}\}$. We therefore proceed by modelling $\{Y_{t+k}\}$ as an autoregressive process (AR), or a vector autoregressive process (VAR) in the spatial case. Indeed, the spatial case is the main focus of this paper.

The wind power measurements from multiple wind farms are logit-normal transformed and embedded in a vector-valued time series, and the expected future value for each vector element provides the forecast of the location parameter for the predictive distribution at the corresponding site. The scale parameter could be similarly modelled, but for simplicity it is assumed to be slowly varying and is tracked by an exponential smoothing scheme on a site-by-site basis.

For a vector-valued process, such as a series of measurements made at multiple locations, dependencies between vector elements may exist on a range of scales. Such spatio-temporal dependence can be captured by VAR models and produce more skilful forecasts than independent AR models. However, as the spatial dimension becomes large, VAR models quickly become difficult to estimate as the number of parameters increases with the square of the dimension, and useful spatial information is increasingly diluted. We therefore pursue a sparse parametrisation of VAR models whereby coefficients linking sites that exhibit spatial co-dependence are retained in the model, and those that do not are omitted. The resulting sparse-VAR (sVAR) is a refined parametrisation of the full VAR model and requires a fewer training data compared to the full VAR equivalent.

III. From VAR to sVAR

A. Definitions

First consider the problem of calculating the predictive density for the wind power generation at a single wind farm. The power measured at the wind farm at time t is contained in the time series $\{x_t\}$. The logit-normal transformation of $\{x_t\}$ is $\{y_t\}$ and we proceed by modelling this series as an autoregressive process of order p, denoted AR(p). The expression relating the future observation y_{t+k} to previous measurements is written

$$y_{t+k} = \sum_{\tau=1}^{p} a_\tau y_{t-\tau+1} + \epsilon_{t+k} \quad , \tag{6}$$

where a_τ is the autoregressive coefficient for the τ^{th} lag, and ϵ_t is additive Gaussian noise with finite variance σ^2. The expected value of y_{t+k} is

$$\hat{\mu}_{t+k} = \sum_{\tau=1}^{p} a_\tau y_{t-\tau+1} \tag{7}$$

which along with σ^2 parametrises the predictive distribution of $\{Y_{t+k}\} \sim N(\hat{\mu}_{t+k}, \sigma^2)$ conditional on the p previous measurements.

Next consider the problem of calculating the predictive density for the wind power generation at M spatially separate wind farms. The power measured at each wind farm at time t is contained in the vector valued time series $\{\mathbf{x}_t\}$ where each $\mathbf{x}_t \in [0, 1]^M$.

The logit-normal transformation and predictive distributions of $\{\mathbf{x}_t\}$ are all calculated by applying Equations (1)–(5) element-wise. We may then proceed to work with the transformed vector-valued time series $\{\mathbf{y}_t\}$, where $\mathbf{y}_t \in \mathbb{R}^M$.

The new time series is modelled as a vector autoregressive process of order p, VAR(p), expressed as

$$\mathbf{y}_{t+k} = \sum_{\tau=1}^{p} \mathbf{A}_\tau \mathbf{y}_{t-\tau+1} + \boldsymbol{\epsilon}_{t+k} \quad , \tag{8}$$

with matrices $\mathbf{A}_\tau \in \mathbb{R}^{M\times M}$ containing the VAR coefficients, and zero-mean Gaussian noise $\boldsymbol{\epsilon}_t \in \mathbb{R}^M$ with non-singular covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$. The expected value of $\mathbf{y}_{t+k}$ is given by

$$\hat{\boldsymbol{\mu}}_{t+k} = \sum_{\tau=1}^{p} \mathbf{A}_\tau \mathbf{y}_{t-\tau+1} \quad . \tag{9}$$

Typically the VAR coefficients and the noise covariance matrix are determined by maximum likelihood estimation, yielding the Yule-Walker equations for the case when the VAR(p) process is Gaussian and no constraints are placed on the parameters. However, estimating all pM^2 VAR coefficients quickly becomes impractical for models of large spatial dimension and can lead to noisy coefficient estimates and unstable predictions, particularly when insufficient training data are available. We therefore pursue a recently proposed method for the sparse estimation of the coefficient matrices to overcome these drawbacks.

B. sVAR Fitting

A 2-stage procedure for fitting a sparse vector autoregressive model has been proposed by Davis *et al.* in [20]. The first stage selects symmetric pairs of coefficients to be included in the sparse model based on the corresponding pair of time series' conditional dependence. The second stage refines the initial selection based on ranking individual coefficients by their t-statistic. At each stage the set of coefficients selected is that which minimises the Bayesian information criterion (BIC). This approach is detailed in the remainder of this section, for further discussion see Davis *et al.* [20].

1) Stage 1: The goal of stage 1 is to determine the order of temporal regression, p, and choose N pairs of off-diagonal coefficients to be retained in the sparse model. This is achieved by eliminating pairs of series which are determined to be conditionally uncorrelated and setting the corresponding VAR coefficients (at all lags) to zero. All diagonal coefficients, i.e. those containing auto-covariate information, are retained in stage 1.

Let $\{\mathbf{y}_{t,i}\}$ denote the i^{th} marginal series of the process $\{\mathbf{y}_t\}$. If two distinct time series $\{\mathbf{y}_{t,i}\}$ and $\{\mathbf{y}_{t,j}\}$ $(i \neq j)$ are conditionally uncorrelated then their partial spectral coherence $PSC_{ij}(\omega) = 0$ for $\omega \in (-\pi, \pi]$. The PSC can be computed efficiently from the spectral density matrix $f^Y(\omega)$ of the process $\{\mathbf{y}_t\}$, where the (i,j)th element of $f^Y(\omega)$ is the usual (cross-)spectrum between $\{\mathbf{y}_{t,i}\}$ and $\{\mathbf{y}_{t,j}\}$. The PSC is the negative rescaled inverse of the spectral density matrix, as demonstrated in [21]. Let $g^Y(\omega) = f^Y(\omega)^{-1}$, then

$$PSC_{ij}(\omega) = -\frac{g^Y_{ij}(\omega)}{\sqrt{g^Y_{ii}(\omega) g^Y_{jj}(\omega)}} \ , \quad \omega \in (-\pi, \pi] \ , \tag{10}$$

where $g^Y_{ij}(\omega)$ denotes the (i,j)th entry of $g^Y(\omega)$.

In practice, however, the estimated PSC will not be exactly zero for a finite number of samples. We therefore rank each pair of time series by a summary statistic, $\hat{S}_{ij}$, calculated from the estimated PSC, which is denoted $P\hat{S}C_{ij}(\omega)$, taken to be the supremum of the squared PSC estimate, i.e.,

$$\hat{S}_{ij} = \sup_{\omega} |P\hat{S}C_{ij}(\omega)|^2 \quad . \tag{11}$$

Large values of $\hat{S}_{ij}$ indicate pairs of series which are likely to be conditionally correlated; we therefore consider the constrained VAR models containing the top N pairs of off-diagonal coefficients plus the M diagonal coefficients, all other coefficients are zero. This reduces the number of parameters to be estimated from pM^2 to $(M + 2N)p$.

Finally, we calculate the maximum likelihood estimate of the constrained VAR models for predetermined sets of values for p and N. When VAR parameters are constrained the parameter estimates and covariance matrix $\Sigma_{\boldsymbol{\epsilon}}$ are commingled and their estimates must be updated iteratively until convergence, see [22] for details. We choose the pair of parameters $(\tilde{p}, \tilde{N})$ that minimise the BIC to take forward to stage 2.

2) Stage 2: The first stage selects VAR coefficients based on conditional correlation according to the BIC, however, it is unable to discriminate between the $2\tilde{p}$ coefficients associated with each pair of series, nor between the $\tilde{p}$ diagonal coefficients associated with each individual series. The aim of the second stage is therefore to refine the selection of coefficients made by stage 1.

We proceed by ranking the non-zero VAR coefficient estimates from the stage 1 model $[\mathbf{A}_\tau]_{ij}$, $\tau = 1, ..., \tilde{p}$ by their t-statistic, which is

$$\Delta_{i,j,\tau} = \frac{[\mathbf{A}_\tau]_{ij}}{\text{s.e.}([\mathbf{A}_\tau]_{ij})} \quad . \tag{12}$$

The standard error, s.e.$(\cdot)$, of $[\mathbf{A}_\tau]_{ij}$ is computed from the asymptotic distribution of the constrained maximum likelihood estimator of the stage 1 model, see [22].

Large values of $\Delta_{i,j,\tau}$ imply significance in the model so we retain the n coefficients with the largest t-statistic values. Once again we calculate the BIC for a set of values of n and choose $n = \tilde{n}$ which gives the minimum BIC value. The resulting sVAR model has an autoregressive order of $\tilde{p}$ and contains $\tilde{n}$ non-zero coefficients; it is denoted sVAR($\tilde{p}$,$\tilde{n}$).

C. Implementation of sVAR

The spectral density matrix used in the calculation of partial spectral coherence must be estimated from available training data. The periodogram smoothed by a modified Daniell kernel is used here, as in [20], though alternative spectral density estimates could be employed.

The BIC is a smooth convex function of the number of parameters being estimated which allows for efficient implementation of the sVAR procedure: once the turning point of the function has been found, the minimum is known and the fitting algorithm can advance. Since the parameter estimation and BIC calculation are relatively expensive this represents a significant speed-up over a naive approach.

It is well documented that the properties of meteorological time series, including wind speed, change slowly over time with changes of season and climate; therefore, it is appropriate to allow the parameters of time series models to track this variation, if it is not modelled directly. The same applies to wind power as a weather-dependant process. Recursively updating AR parameters is frequently practised and can easily extend to VAR models; however, it is not possible to modify the sparsity structure of the proposed sVAR model in a simple way. Indeed, the idea of slowly varying parameters conflicts with abruptly choosing to include or remove a coefficient.

In order to capture these gradual changes the sVAR is trained on a window of the most recent measurements, and then re-trained in the same way periodically, i.e., at any time t, the model is trained on based on the past observations between $t-L$ and $t-1$, where L is the training window length. For comparison, the AR and VAR benchmarks are trained in the same fashion. Note that the parameters of an sVAR (with a fixed sparsity structure) could be updated in a recursive framework (such as a least squares update [23]) in the same way as a conventional AR or VAR model, but this would distract from our main investigation so is not done here.

The scale parameter should also be allowed to track changes in dynamics resulting meteorological variation, and that is the subject of the next section.

IV. Dynamic Tracking of Scale Parameter

The scale parameter $\sigma^2_{t+k,i}$ of $\{\mathbf{Y}_{t+k,i}\}$ is estimated recursively by exponential smoothing for each site $i \in \{1,...,M\}$ independently, i.e. assuming no spatial dependence. To avoid notational clutter the second index is dropped in this section.

We apply two variations on exponential smoothing and compare their performance. First, the boundary weighted forgetting factor down-weights of observations when the location parameter is close to the bounds akin to [13]. The logit-normal transformation is particularly sensitive in these regions and this approach is designed to robustify the smoothing scheme. A second scheme is also proposed with a dynamic forgetting factor motivated by regime-switching type behaviour often exhibited by weather-dependent processes.

1) Boundary Weighted Forgetting Factor: In a modification to standard exponential smoothing, observations are down weighted by a factor ω_t when the expected power $\gamma^{-1}(\hat{\mu}_{t+k})$ is close to the bounds due to the sensitivity of the logit-normal transformation in these regions [13]. The factor ω_t is given by

$$\omega_t = 4\gamma^{-1}(\hat{\mu}_{t+k})\big(1-\gamma^{-1}(\hat{\mu}_{t+k})\big). \tag{13}$$

and the smoothing scheme is written

$$\hat{\sigma}^2_{t+k} = \lambda^*_t\hat{\sigma}^2_t + (1-\lambda^*_t)(y_t-\hat{\mu}_t)^2 \tag{14}$$

where $\lambda^*_t = 1-(1-\lambda)\omega_t$.

2) Dynamic Forgetting Factor: The behaviour of wind power generation can switch quickly between periods of smooth generation and periods of volatile generation. In the event of such a switch it is necessary to briefly but dramatically reduce the forgetting factor in order to *forget* out of date, mismatched information. Therefore, when the difference between the squared residual, ϵ^2_t, and estimated scale parameter $\hat{\sigma}^2_t$ is large, the forgetting factor is reduced. The dynamic forgetting factor is given by the logit function as follows,

$$\lambda^*_t = \lambda - \frac{b}{1+\exp[c(a-\mathcal{E}_t)]} \quad , \tag{15}$$

where $\mathcal{E}_t = |\hat{\sigma}^2_t - \epsilon^2_t|$. The parameters a and b control the threshold location and the minimum value that λ^*_t can take, respectively, and c controls the gradient of the transition.

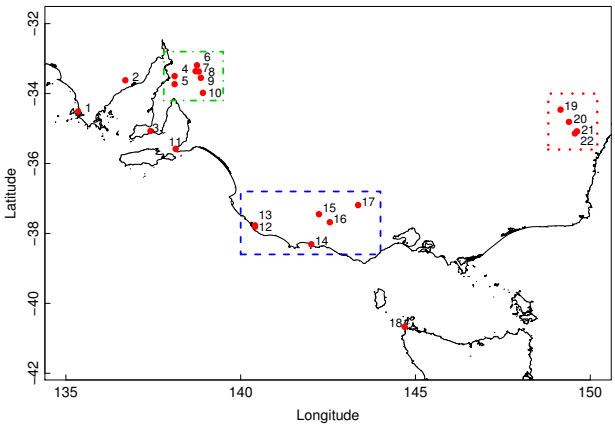


Fig. 1: Location of 22 sites located in S.E. Australia used in the data model. Boxed regions correspond to those in Figure 3.

V. Application and Case Study

A. Dataset

The proposed approach is tested on 5 minute mean wind power data provided by the Australian Energy Market Operator [24], which comprises recordings of wind farm power generation at 22 wind farms in southeastern Australia. Data from 2012 and 2013 are available comprising 210 528 measurements at each site; all have been normalised by the nominal power of the corresponding wind farm so that they occupy the range [0,1]. Wind farm locations are plotted in Figure 1. The 2012 data are used as a training set on which the implementation of the fitting procedure is optimised by cross-validation, and the parameters of the exponential smoothing scheme are chosen. The 2013 data are then used to evaluate the performance of the predictor, the results of which are presented and discussed in Section V-C. The results comprise the analysis of more than 2.3 million individual forecasts. The complete dataset as used in this paper is available to download from [25]. In this study, we only predict for $t+1$ (one step ahead), though cases with forecast for $t+k$ could be similarly be considered.

B. Implementation

The size of data window, L, used to train the AR, VAR and sVAR is determined heuristically, by cross-validation using the training dataset. The chosen window length is that which minimises the point prediction root-mean-squared error (RMSE) since this is the cost function minimised in the predictors' estimation. A new model is fit for each calendar month to

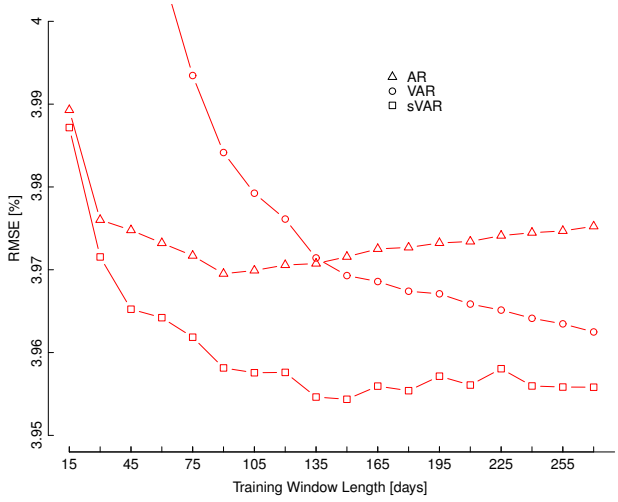


Fig. 2: Variation of root mean squared error (RMSE) of AR, VAR and sVAR models with training window length.

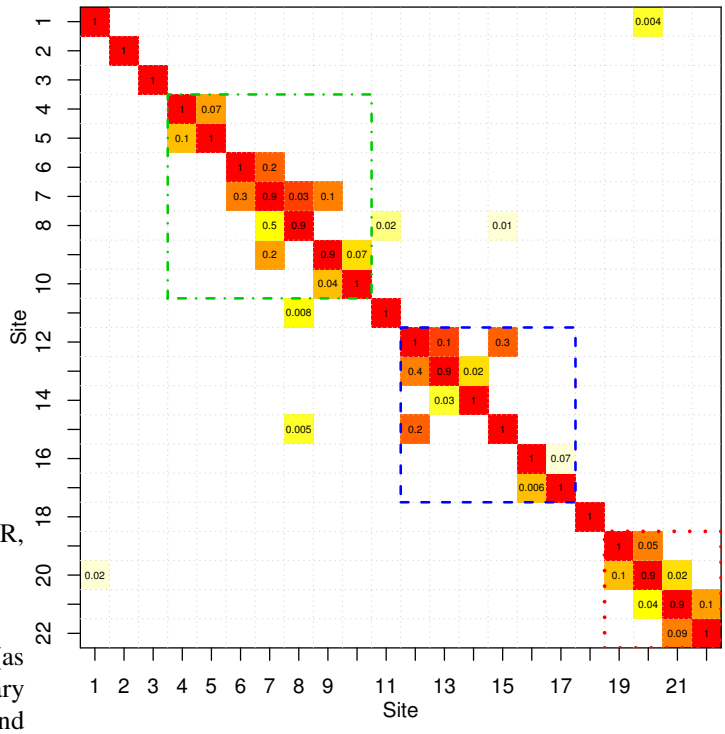


Fig. 3: Superposition of January 2013 sVAR coefficient matrices taking absolute values and displaying 1 s.f. Blank entries correspond to coefficients not included in the sparse model and are therefore equal to zero at all lags. Boxed regions correspond to those in Figure 1.

be forecasts to track changes in the time series dynamics (as discussed in Section III-C); this choice is somewhat arbitrary but provides a satisfactory trade-off between accuracy and computational expense. Results of the window length selection procedure are illustrated in Figure 2. The optimal window length is $L = 60$ days for the AR model and $L = 150$ days for the sVAR. As already mentioned, the conventional VAR model is extremely data-hungry and computationally expensive to fit and as a result a VAR model cannot be fit with more than $L = 270$ days of training data on the computer being used (64-bit operating system, 8GB of RAM, Intel Core i7-2600 3.4GHz processor). Each VAR model is therefore trained on the maximum $L = 270$ days of data.

The optimal window length is directly related to the number of parameters being estimated in each of the three models. The AR has pM parameters so only requires a modest amount of training data, whereas the VAR has pM^2 parameters and as a result requires much more training data to produce reliable parameter estimates. The sVAR offers a compromise: increase the number of parameters to take advantage of spatial information, but only include those parameters deemed significant.

The basic forgetting factor for both exponential smoothing schemes is chosen such that the effective memory is 2000 samples ($\lambda = 0.9995$). The parameters of the dynamic forgetting factor exponential smoothing scheme are chosen by expert judgement such that the forgetting factor does not drop bellow 0.5 ($b = 0.4995$), such that the forgetting factor is reduced when the squared residuals exceed 0.1 ($a = 0.1$), and such that the gradient of the logit function is sharp ($c = 50$).

C. Results

The proposed technique is implemented on the test dataset in the manor determined by the cross-validation exercise described above.

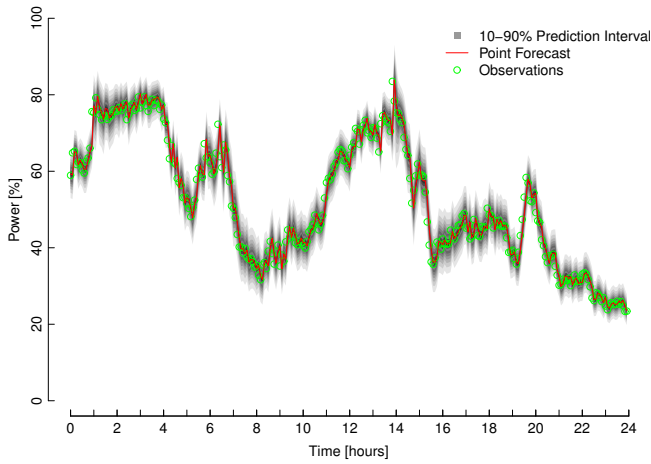
The 2-stage method for fitting an sVAR model results in the inclusion of 5%–10% of the possible pM^2 parameters. The number of lags is typically $\tilde{p} = 3$. A superposition of the VAR coefficient matrices, taking the absolute value of each element, from one sVAR model is illustrated in Figure 3. There is a strong diagonal structure with off-diagonal coefficients appearing in blocks corresponding to groups sites that are close to one another geographically, precisely the sites one would expect to display spatio-temporal dependence.

The 10 minute-ahead sVAR forecasts made over a 24 hour period, and the behaviour of the variable forgetting factor are presented in Figure 4. Prediction intervals from 10%–90% are illustrated by shading. The variable forgetting factor behaves as intended, decreasing to allow fast learning when the behaviour switches, and then returning to normal. The width of the prediction intervals behave accordingly and widen quickly during volatile periods, and narrowing during periods of relative calm.

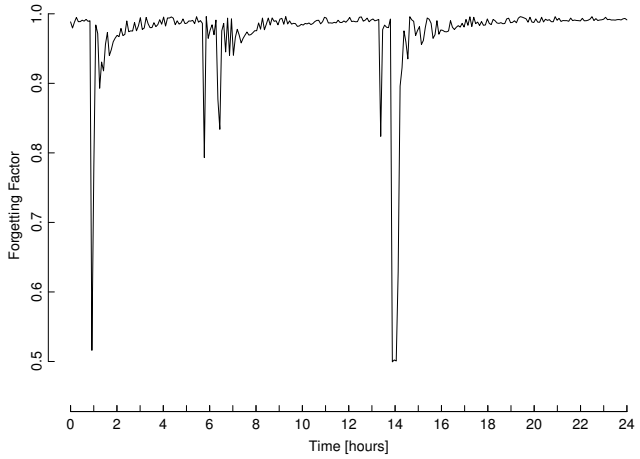
Both point and probabilistic forecast scores are used to quantify the skill of the proposed and benchmark methods. Point forecasts are assessed using the familiar root mean squared error, RMSE $= \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x}_t)^2}$, and mean absolute error, MAE $= \frac{1}{T}\sum_{t=1}^{T}|x_t - \hat{x}_t|$, where $\hat{x}_t = \gamma^{-1}(\hat{\mu}_t)$ is the predicted value of x_t.

The skill of the distributional forecasts is quantified by the continuous rank probability score (CRPS) and log score [26]. The CRPS is given by

$$\text{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int_0^1 \{F(x|\hat{\mu}_t, \hat{\sigma}_t) - \mathbf{1}(x \geq x_t)\}^2 dx \qquad (16)$$



(a) Point and probabilistic forecasts made 5 minutes (1 step) ahead.



(b) Value of the dynamic forgetting factor, λ_t^*.

Fig. 4: Probabilistic forecasts and value of the dynamic forgetting factor at site 9 for July 11th 2013.

TABLE I: Mean skill scores (RMSE, MAE and CRPS as % of nominal power) across all sites with % improvement (Δ%) for dynamic vs boundary weighted (BW) forgetting factor.

		Persistence	AR	VAR	sVAR
	RMSE	3.956	3.970	3.962	3.954
	MAE	2.308	2.347	2.358	2.343
BW λ	CRPS	n/a	1.843	1.837	1.801
	Log Score	n/a	5.080	5.067	5.909
Dynam. λ	CRPS	n/a	1.751	1.751	1.745
	Log Score	n/a	4.634	4.629	4.622
Δ% vs	CRPS Δ%	n/a	*5.0%*	*4.7%*	*3.0%*
BW λ	Log Score Δ%	n/a	*8.8%*	*8.6%*	*21.8%*

TABLE II: Mean probabilistic forecast skill scores with dynamic forgetting factor broken down by calendar month (CRPS as % of nominal power). The best scores are highlighted in bold.

Month		AR	VAR	sVAR
January	CRPS	1.910	**1.896**	1.897
	Log Score	4.788	4.788	**4.781**
February	CRPS	1.826	1.819	**1.812**
	Log Score	4.752	4.755	**4.749**
March	CRPS	1.796	1.780	**1.779**
	Log Score	4.685	4.691	**4.681**
April	CRPS	**1.375**	1.383	1.380
	Log Score	4.351	4.355	**4.337**
May	CRPS	**1.617**	1.637	1.634
	Log Score	4.570	**4.565**	**4.565**
June	CRPS	1.486	1.500	**1.483**
	Log Score	4.434	4.435	**4.425**
July	CRPS	**1.544**	1.567	1.548
	Log Score	4.460	4.449	**4.436**
August	CRPS	1.831	1.840	**1.829**
	Log Score	4.712	4.697	**4.686**
September	CRPS	1.717	1.710	**1.700**
	Log Score	4.606	4.595	**4.594**
October	CRPS	2.001	1.999	**1.990**
	Log Score	4.759	**4.739**	**4.739**
November	CRPS	2.020	**2.007**	2.009
	Log Score	4.790	4.778	**4.777**
December	CRPS	1.883	**1.871**	1.875
	Log Score	4.703	4.697	**4.692**
All	CRPS	1.751	1.751	**1.745**
	Log Score	4.634	4.629	**4.622**

where F is the cumulative form of the predictive distribution and $\mathbf{1}(\cdot)$ is the indicator function. CRPS rewards sharpness and reduces to MAE when the forecast is deterministic.

The log score is the mean negative log of the predictive distribution evaluated at the corresponding observation, Log Score $= \frac{1}{T}\sum_{t=1}^{T} -\ln\left(f(x_t|\hat{\mu}_t,\hat{\sigma}_t)\right)$. Due to its logarithmic nature, the log score is not as robust as the CRPS: measurements in the tails of the predictive distribution heavily penalised and the score returns ∞ if a single measurement falls where the predictive distribution is numerically zero.

Point and probabilistic forecast skill scores are listed in Table I and probabilistic scores are broken-down by calendar month in Table II. The persistence point forecast, which is simply $\hat{x}_{t+k} = x_t$, is also included in Table I. Point forecast scores show that the sVAR improves on all the benchmarks in terms RMSE, and all but persistence in terms of MAE. Persistence does not offer probabilistic information, which is required for optimal decision making under uncertainty, hence the move to more sophisticated approaches.

With the boundary weighted tracking of the scale parameter, the sVAR performs very well in terms of CRPS but has a poor log score, when compared to the other models. The high log score is an effect of the very sharp predictive distribution close to the upper and lower bounds where measurements are more likely to be found in the tails of the distribution. The AR and VAR models, with their higher variance and broader predictive distributions, are not exposed to this affect as frequently and this is reflected in their comparatively low log scores.

When the scale parameter is tracked by the proposed dynamic forgetting factor scheme, all three models see significant improvement in both CRPS and log score compared to the boundary weighted scheme. Notably, the improved behaviour of the predictive distributions close to the bounds has brought the log score of the sVAR in line with the AR and VAR models. In this case, the sVAR is performs marginally better than the two benchmarks in terms of both CRPS and log score.

Reliability (or calibration) of probabilistic forecasts is critical and can be assessed with quantile-quantile reliability diagrams, such as in Figure 5. A calibrated forecast with nominal proportion α should cover the observation α% of the

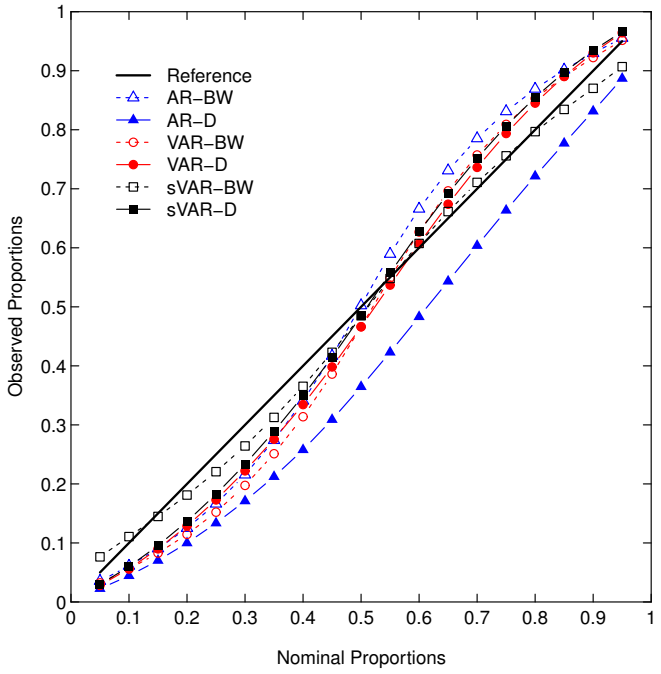


Fig. 5: Reliability diagram for the AR, VAR and sVAR models with boundary weighted (BW) and dynamic (D) forgetting factors.

time. In Figure 5 nominal quantiles from 5% to 95% in steps of 5% are evaluated.

The forecasts produced by the sVAR with the boundary weighted scale factor smoothing is reliable and the best calibrated of the six forecasts, followed by the sVAR with dynamic smoothing. The boundary weighted smoothing scheme results in better calibration than the dynamic smoothing scheme for the sVAR and AR models, but the opposite is true for the conventional VAR. The calibration of forecasts produced by the AR model with dynamic smoothing is particularly poor.

D. Discussions

It has been demonstrated that the proposed approach produces forecasts that are a non-negligible improvement on two competitive benchmarks in terms of several skill scores and reliability, while also offering attractive numerical properties through sparse parametrisation. The sVAR makes it possible to model data of high spatial dimension that would be impractical, or impossible, with a conventional VAR approach. In addition, the data-driven detection of dependence structures means that the benefits of a spatial treatment can be realised without knowledge of precise locations, or in situations where many generators are located in a small area, as is commonplace in the smart grid paradigm. This technique is equally applicable to other forecasting problems where VARs have been used, such as wind speed [15] and solar power forecasting [27], including short-term forecasting at other temporal resolutions, e.g. hourly.

However, the sVAR comes with some limitations: Regression parameters are commonly updated by a process of recursive estimation [23], or replaced with coefficient functions of some covariate such as wind direction [14], [28]. While in principle these techniques could be applied to an sVAR model, they would not be able to capture possible changes in the sparsity structure.

Computational cost is of interest: while the MLE of a single constrained VAR model takes around 2 minutes, compared to 4 for the full VAR, the calculation is repeated making the total time to fit an sVAR an order of magnitude larger than the conventional VAR. However, the stopping criterion described in Section III-C may be refined, and other speed-ups are possible such as parallelising the fitting procedure. There exist alternative methods for fitting sparse regression models, such as quasi-MLE, [29], and penalised linear regression (e.g. *lasso* [30]) which can be implemented by very efficient algorithms which are available in common software packages. However, reformulating the problem as one of linear regression comes at a cost as both the temporal ordering of samples and any error cross-covariance between spatial locations is negated.

Furthermore, retaining full covariance information may offer opportunities for future development. While the deterministic part of the forecast methodology described in this paper utilises spatial information, the scale parameter, and by extension the predictive distribution, for each location are calculated independently. A more general probabilistic forecast could consider the full joint predictive distribution taking into account the full covariance structure of observations.

The framework facilitated by the logit-normal transformation allows us to work in the familiar Gaussian domain, however, a generalisation of this transformation has been proposed in [13] for wind power forecasting. By including a shape parameter to control the skewness of the transformation, the properties of the transformed data may be improved. The optimal shape parameter to fit the marginal distribution of the data can be calculated by standard techniques, however, the same is not true of the conditional distributions, which are of concern here. In [13] the optimal shape parameter for the conditional distributions of a univariate time series is determined by an iterative process, which would be extremely time consuming in the spatial case, particularly if individual shape parameters were assigned to each location. Furthermore, the effects of using different shaped transformations on the spatio-temporal dependencies of the transformed data are unknown. For these reasons we leave the generalised logit-normal transformation for future investigation.

VI. Conclusions

This paper develops a large-scale spatial technique for producing very-short-term probabilistic forecasts of wind power generation at multiple locations. A parametric framework for distributional forecasts based on the logit-normal transformation and distribution is combined with a spatio-temporal model for the distribution's location parameter, and two competing smoothing schemes for it's scale parameter are presented. The location parameter is first modelled as a vector autoregressive process, and then as a sparse vector autoregressive process (sVAR), dramatically reducing the number of coefficients requiring estimation, and by extension the computational expense of model fitting and the volume training data required.

In a case study, the proposed sVAR technique has been used to produce 5 minute ahead probabilistic forecasts of wind power at 22 wind farms in southeastern Australia for a test period of 1 year. The performance of the sVAR is compared to conventional VAR and AR models yielding improvement in terms of both deterministic and probabilistic skill scores, as well as in the reliability of the distributional forecasts.

This work was motivated by the desire to produce accurate very-short-term forecasts at multiple wind farms, ultimately on a national scale, i.e., at 100s of wind farms. Future work should extend to spatial dimensions of this order, other forecast horizons, and consider building an adaptive sVAR, possibly with a dynamic sparsity structure. The parametric framework could also be extended by moving to the generalised logit-normal distribution and transformation which would require the development of an efficient method for determining the optimal shape parameter(s) with respect to conditional distributions.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Australian Energy Market Operator for their supply of wind power data and Stefanos Delikaraoglou for its pre-processing in addition to the support of the UK's Engineering and Physical Sciences Research Council via the University of Strathclyde's Wind Energy Systems Centre for Doctoral Training, grant number EP/G037728/1, and COST Action ES 1002. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which have improved the quality of this paper.

REFERENCES

[1] T. Ackermann, Ed., *Wind power in power systems*, 2nd ed. John Wiley & Sons: New York, 2012.
[2] J. Morales, A. Conejo, H. Madsen, P. Pinson, and M. Zugno, *Integrating Renewable in Electricity Markets*. Springer, 2014.
[3] P. Sørensen, N. A. Cutululis, A. Vigueras-Rodriguez, L. E. Jensen, J. Hjerrild, M. H. Donovan, and H. Madsen, "Power fluctuations from large wind farms," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 958–965, 2007.
[4] J. Kristoffersen and P. Christiansen, "Horns rev offshore wind farm: its main controller and remote control system," *Wind Engineering*, vol. 27, pp. 351–359, 2003.
[5] G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, and C. Draxl, *The State-of-the-Art in Short-Term Prediction of Wind Power*. ANEMOS.plus, 2011.
[6] X. Zhu and M. G. Genton, "Short-term wind speed forecasting for power system operations," *Int. Stat. Rev.*, vol. 80, no. 1, pp. 2–23, 2012.
[7] P. Pinson and H. Madsen, "Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models," *J. Forecasting*, vol. 31, no. 4, pp. 281–313, 2012.
[8] M. Yoder, A. S. Hering, W. C. Navidi, and K. Larson, "Short-term forecasting of categorical changes in wind power with markov chain models," *Wind Energy*, pp. n/a–n/a, 2013.
[9] A. Kusiak, H. Zheng, and Z. Song, "Short-term prediction of wind farm power: A data mining approach," *IEEE Trans. Energy Conversion*, vol. 24, no. 1, pp. 125–136, 2009.
[10] J. Catalão, H. M. I. Pousinho, and V. Mendes, "Hybrid wavelet-PSO-ANFIS approach for short-term wind power forecasting in portugal," *IEEE Trans. Sustainable Energy*, vol. 2, no. 1, pp. 50–59, 2011.
[11] Y. Liu, J. Shi, Y. Yang, and W.-J. Lee, "Short-term wind-power prediction based on wavelet transform, support vector machine and statistic-characteristics analysis," *IEEE Trans. Ind. Appl.*, vol. 48, no. 4, pp. 1136–1141, 2012.
[12] P. Pinson, L. E. A. Christensen, H. Madsen, P. E. Sørensen, M. H. Donovan, and L. E. Jensen, "Regime-switching modelling of the fluctuations of offshore wind generation," *J. Wind Eng. Ind. Aero.*, vol. 96, no. 12, pp. 2327–2347, 2008.
[13] P. Pinson, "Very-short-term probabilistic forecasting of wind power with generalized logit-normal distributions," *J. Roy Stat. Soc.: Series C*, pp. 555–576, 2012.
[14] A. S. Hering and M. G. Genton, "Powering up with space-time wind forecasting," *J. Am. Stat. Assoc.*, vol. 105, pp. 92–104, 2010.
[15] J. Dowell, S. Weiss, D. Hill, and D. Infield, "Short-term spatio-temporal prediction of wind speed and direction," *Wind Energy*, vol. 17, no. 12, pp. 1945–1955, 2014.
[16] J. Tastu, P. Pinson, P.-J. Trombe, and H. Madsen, "Probabilistic forecasts of wind power generation accounting for geographically dispersed information," *IEEE Tran. Smart Grid*, vol. 5, no. 1, pp. 480–489, 2014.
[17] J. Tastu, P. Pinson, and H. Madsen, *Space-time trajectories of wind power generation: Parameterized precision matrices under a Gaussian copula approach*, 2014, to appear.
[18] M. Wytock and J. Zico Kolter, "Large-scale probabilistic forecasting in energy systems using sparse gaussian conditional random fields," in *IEEE Conference on Decision and Control*, 2013.
[19] A. Lau and P. McSharry, "Approaches for multi-step density forecasts with application to aggregated wind power," *The Annals of Applied Statistics*, vol. 4, no. 3, pp. 1311–1341, 2010.
[20] R. A. Davis, P. Zang, and T. Zheng, "Sparse vector autoregressive modelling," *arXiv:1207.0520*, 2012.
[21] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, pp. 157–172, 2000.
[22] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Heidelberg: Springer-Verlag, 2005.
[23] L. Ljung and T. Söderström, *Theory of Recursive Identification*. MIT Press, 1983.
[24] Australian Energy Market Operator, "AEMO 5 minute wind power data, 2011–2012." [Online]. Available: http://www.aemo.com.au/
[25] http://www.jethrodowell.com.
[26] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *J. Roy. Stat. Soc.: Series B*, vol. 69, pp. 243–268, 2007.
[27] R. Bessa, A. Trindade, and V. Miranda, "Spatial-temporal solar power forecasting for smart grids," *IEEE Trans. Ind. Informatics*, vol. PP, no. 99, pp. 1–1, 2014.
[28] P. Pinson, H. A. Nielsen, H. Madsen, and T. S. Nielsen, "Local linear regression with adaptive orthogonal fitting for the wind power application," *Statist. Comput.*, vol. 18, pp. 59–71, 2008.
[29] T. McElroy and D. Findley, "Fitting constrained vector autoregression models," in *Empirical Economic and Financial Research*, pp. 451–470, Springer, 2015.
[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Society: Series B*, vol. 58, pp. 267–288, 1996.

Jethro Dowell (S'15) was born in the UK in 1989. He received the M.Phys. degree in Mathematics and Theoretical Physics from the University of St Andrews, UK, in 2011. Since 2011 he has been with the University of Strathclyde's Centre for Doctoral Training in Wind Energy Systems, Glasgow, UK, as a Ph.D. student where his main research interests are short-term wind and wind power forecasting.

Jethro is a member of the IET and the IEEE Power & Energy Society.

Pierre Pinson (M'11–SM'13) received the M.Sc. degree in Applied Mathematics from the National Institute for Applied Sciences (INSA Toulouse, France) and the Ph.D. degree in Energetics from Ecole des Mines de Paris (France).

He is a Professor at the Technical University of Denmark, Centre for Electric Power and Energy, Department of Electrical Engineering, also heading a group focusing on Energy Analytics & Markets. His research interests include among others forecasting, uncertainty estimation, optimization under uncertainty, decision sciences, and renewable energies.

Prof. Pinson acts as an Editor for the IEEE TRANSACTIONS ON POWER SYSTEMS, for the *International Journal of Forecasting* and for *Wind Energy*.