Environmetrics 00, 1–25

DOI: 10.1002/env.XXXX

Benefits of spatio-temporal modelling for short term wind power forecasting at both individual and aggregated levels

Amanda Lenzi[a]* and Ingelin Steinsland[b] and Pierre Pinson[c]

Summary: The share of wind energy in total installed power capacity has grown rapidly in recent years around the world. Producing accurate and reliable forecasts of wind power production, together with a quantification of the uncertainty, is essential to optimally integrate wind energy into power systems. We build spatio-temporal models for wind power generation and obtain full probabilistic forecasts from 15 minutes to 5 hours ahead. Detailed analysis of the forecast performances on the individual wind farms and aggregated wind power are provided. We show that it is possible to improve the results of forecasting aggregated wind power by utilizing spatio-temporal correlations among individual wind farms. Furthermore, spatio-temporal models have the advantage of being able to produce spatially out-of-sample forecasts. We evaluate the predictions on a data set from wind farms in western Denmark and compare the spatio-temporal model with an autoregressive model containing a common autoregressive parameter for all wind farms, identifying the specific cases when it is important to have a spatio-temporal model instead of a temporal one. This case study demonstrates that it is possible to obtain fast and accurate forecasts of wind power generation at wind farms where data is available, but also at a larger portfolio including wind farms at new locations.
The results and the methodologies are relevant for wind power forecasts across the globe as well as for spatial-temporal modelling in general.

Keywords: wind power; aggregated forecast; probabilistic forecast; integrated nested Laplace approximation.

[a] *Applied Mathematics and Computer Science Department, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*
[b] *Department of mathematical sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*
[c] *Electrical Engineering Department, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*
* *Correspondence to: Applied Mathematics and Computer Science Department, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark. E-mail: amle@dtu.dk*

This paper has been submitted for consideration for publication in *Environmetrics*

1. INTRODUCTION

Wind power is a clean, renewable and widely available source of energy and electricity generated from wind power is increasing world wide. A challenge for utilizing wind power is that the generated amount of energy varies much and relatively fast over time due to variations in wind. An important tool for efficiently integrating wind power in a system with energy sources that can be controlled, e.g. thermal energy and hydro power, is high quality probabilistic forecasts for short term wind power production Ackermann [2005]. Moreover, accurate forecasting of wind power generation makes wind more competitive in the energy market, since it reduces the imbalance costs to producers Girard et al. [2013]. Recently, there has been an increasing amount of research in wind speed and wind power forecasts. Most of the developments are for point forecasts (e.g. Louka et al. [2008], Catalão et al. [2011]), i.e. the forecast consists of one value for each wind farm or location. To make better decisions one also needs to quantify the uncertainty of the forecast, and provide a probability density function (pdf) instead of a point forecast. This is called a probabilistic forecast. For a probabilistic forecast to be useful it needs to be calibrated and sharp Gneiting et al. [2007]. Calibrated refers to a forecast that is reliable: in the long term, 90% of the observed wind production should be within a 90% forecast interval, 80% of the observations within a 80% forecast interval and so forth. Sharpness refers to the spread of the predictive distribution, a sharper forecast is more concentrated and better when subject to calibration.

In recent years, more emphasis has been placed on probabilistic forecasts in order to quantify the inherent uncertainties in wind, see Pinson and Kariniotakis [2010] and Bremnes [2004]. From the point of view of a wind farm operator, probabilistic forecasts improve decision making regarding the management of the immediate regulating and spinning reserves, which is essential given the financial penalties that are incurred for deviating from the declared power levels. From the point of view of a system operator, the aggregated wind power generation over pre-defined areas is of particular importance. Some recent

contributions to the modelling and forecasting of aggregated wind power energy are Lau and McSharry [2010] and Focken et al. [2002], which do, however, not account for spatio-temporal dependencies.

To illustrate the challenge of forecasting individual and aggregated wind power simultaneously, we consider a toy example of two wind farms at one lead time and denote their forecasts X_1 and X_2 (these are random variables). The aggregated forecast for the system is $Y = X_1 + X_2$. We know from basic probability, see e.g Ross [2015], that the expected value for the system is $\mathrm{E}(Y) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$ and the variance is $\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)$. Hence, to obtain a forecast for the system Y we also need to model the dependency between the wind farms. This calls for a spatio-temporal model for wind power production. If the productions at the two farms in our toy example are dependent and have a positive covariance, but are assumed independent in the forecast, the variance of Y gets too small and the forecast for Y is not calibrated. Verification of multivariate probabilistic forecasts is an active field of research, for which new scores and diagnostic tools are being proposed and discussed, see, e.g., Pinson and Tastu [2013], Scheuerer and Hamill [2015], Thorarinsdottir et al. [2016] among others. A pragmatic approach is to evaluate relevant univariate probabilistic forecasts derived from the multivariate probabilistic forecast.

Wind speed, and hence wind power production, has temporal and spatial dependencies. In Section 2 we will see that this is also the case for western Denmark. Indeed, our approach of basing the forecast on recent observations relies on the temporal dependency. As demonstrated with our toy example, the spatial dependency also needs to be considered for the aggregated forecasts to be calibrated. Furthermore, borrowing information by utilizing the spatial correlation among individual wind farms has been shown to reduce the errors in point forecasts significantly Tastu et al. [2011], and has the advantage of producing models that are able to generate forecasts at locations that are not within the observation samples.

Several characteristics in a typical wind power series make it a challenging problem to generate accurate forecasts. First of all, wind power is bounded below by zero, when no turbines are operating, and above by the nominal capacity, when all turbines are generating their rated power output. In addition, wind power series are clearly non-Gaussian. In fact, the marginal distribution of wind power production data possesses tails that are heavier than the Gaussian distribution. Instead of using a classical Gaussian distribution, truncated Gaussian, censored Gaussian and generalized logit-normal distributions have been proposed to model the conditional density of wind power Gneiting et al. [2006] Pinson [2012]. Our approach is based on the logistic function, which has shown to be a suitable transformation to normalize wind power data Dowell and Pinson [2016].

We propose statistical models that yield calibrated probabilistic forecasts of wind power generation at multiple sites and lead times simultaneously. We define three different models that share the same data process, or likelihood, but differ in the process model. We start with a model consisting of a location specific intercept and an autoregressive component that captures the local variability without considering the dependency between the farms. This model is well suited for individual forecasts, but it is not calibrated for aggregated forecasts. To obtain reliable aggregated forecasts, we introduce two different models that capture the spatio-temporal features present in the data. The first has a common intercept and a spatio-temporal process, in which spatial and temporal dependency is modelled by a latent Gaussian field. The second is a combination of the previous two models, with a common intercept, an autoregressive process and a spatio-temporal term that varies in time with first order autoregressive dynamics. To meet the computational requirements the stochastic partial differential equations (SPDE) approach to spatial and temporal-spatial modelling is taken Lindgren et al. [2011] Blangiardo and Cameletti [2015], for which fast Bayesian inference can be performed using integrated nested Laplace approximations (INLA).

Moreover, we study the performance of the proposed models in forecasting wind power from

individual and aggregated farms under two different scenarios. In a first stage, we consider out-of-sample forecasts in terms of time, that is, they are obtained for wind farms inside the training set. However, there are situations where not enough data is available for all the wind farms, and even when it is available, the computational load to calculate forecasts for all of them can be very high. In those cases, it is important to have a method of forecasting that is as robust as possible, so that parameters estimated using only part of the portfolio can readily be used to forecast a larger data set, including wind farms at new locations. In such cases, temporal models that require local information for the parameter estimation cannot be used to obtain forecasts. Based on this, in a second stage, we consider spatially out-of-sample forecasts generated by the proposed spatio-temporal models. We develop and evaluate the forecasts for wind power production in western Denmark based on a data set for 349 wind farms with energy production observations every 15 minutes from 2006 to 2012.

In Section 2, we provide a short description of the wind power data that we use in our study and the data treatment. The hierarchical models used to generate probabilistic forecasts of wind power generation, as well as the framework for producing probabilistic forecasts with such models, are outlined in Section 3. In Section 4, we give details of the probabilistic forecasting scheme and outline the scores and the scenarios used for forecast evaluation. In Section 5, we show the results of a case study where we obtain spatio-temporal forecasts and spatially out-of-sample forecasts on the individual and aggregated level. Section 5 also contains the results of a simulation study, whereas conclusions of the work are drawn in Section 6.

2. DANISH WIND POWER PRODUCTION DATA

This project is based on a system of 349 wind farms in western Denmark. Observations of wind power production between January 2006 and March 2012 were provided by the

Transmission System Operator in Denmark and each measurement consists of temporal average over a 15-min time period.

The measurements at each site have been normalised by the nominal power of the corresponding wind farm, so that they are within the range [0,1]. Moreover, to avoid including long chains of zeros that come from temporary shutdown of the turbines for maintenance or missing data that are reflected as unreasonably long periods of zero wind power production, we choose to analyze only wind farms containing at most 10% of zero observations. The evaluation of the predictive performance of individual wind farms and aggregated wind power is done as % of nominal power, which is a common practice in the wind power field (e.g., Pinson [2012], Tastu et al. [2011], Dowell and Pinson [2016]).

Figure 1 (a) shows the spatial correlation of wind power production between one wind farm located in the southern part of Denmark and the remaining wind farms of the portfolio. The higher correlations come from farms that are closer, while the correlations of wind farms far from it are almost zero. Next, we check the dependency of the temporal correlation at fixed locations. Figure 1 (b) shows the mean autocorrelation function of wind power production among wind farms located in western Denmark. The autocorrelation function of the normalized wind power production at a single farm has a slow decay and on average, it drops down to zero after about 40 hours.

[Figure 1 about here.]

Wind power generated by a farm over a period of time is non-Gaussian and bounded between zero and one after the normalization. In fact, wind power distribution has a sharper peak than the Gaussian distribution and is also significantly right-skewed. In all the approaches to be described next, we apply the logit-normal transformation to the normalized wind power data following the procedure in Pinson [2012].

Let $X(\mathbf{s}, t)$ denote the normalized wind power production at location $\mathbf{s} \in \mathcal{D}_s$ and time $t \in \mathcal{D}_t$, with respective observations or measurements indicated by $x(\mathbf{s}, t)$. The logit-normal

transformation is given by

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

with inverse

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

To represent the logit-normal transformation in the cases where measurements are equal to zero and one, we follow the approach by Lesaffre et al. [2007] for modelling outcome scores in $[0,1]$.

Moreover, to evaluate the performance of aggregated wind power forecasts, we obtain the normalized aggregated wind power at lead time h by

$$x_A(t+h) = \frac{\sum_{j=1}^{N} c_j x(\mathbf{s}_j, t+h)}{\sum_{j=1}^{N} c_j}, \tag{3}$$

where c_j is the capacity of wind farm at location $\mathbf{s}_j$ and N is the total number of wind farms in the portfolio.

3. MODELS AND FITTING SCHEME

In this section, we introduce three different statistical models for wind power production. We start with a simpler autoregressive model, where each wind farm is considered as an independent replicate of the same process. Next, we describe two versions of a spatio-temporal model, in which spatial correlation is captured by a latent Gaussian field with a Matérn covariance function. The simplest version has only a spatio-temporal component, while the other has both, an autoregressive process and a spatio-temporal model. The section ends with the estimation procedure and how we obtain probabilistic forecasts.

3.1. Likelihood

We denote by $Y(\mathbf{s},t)$ the normalized logit-normal transformed wind power generation at location $\mathbf{s}$ and time t, which is calculated using (1). We assume the following distribution for $Y(\mathbf{s},t)$ at the first level of the hierarchical models considered in this section

$$Y(\mathbf{s},t) \sim \text{Normal}\left(\mu(\mathbf{s},t), \sigma_e^2\right), \tag{4}$$

with σ_e^2 being the variance of the measurement error, defined by a Gaussian white noise process both serially and spatially uncorrelated. The term $\mu(\mathbf{s},t)$ is the mean of the random process and can be defined by other process levels giving rise to different hierarchical models that are described in the following sections.

3.2. Latent Gaussian structure

3.2.1. Temporal model (Model T)

We start with a time series model where each wind farm is considered as an independent replicate of the same random process. The independence assumption is of course a simplification, since the wind power production in one location is probably dependent on the production in other locations. We assume that $\mu(\mathbf{s},t)$, in (4), is constant in time and can be modelled as

$$\mu(\mathbf{s},t) = b(\mathbf{s}) + w_{\mathbf{s}}(t), \tag{5}$$

where $b(\mathbf{s})$ is an intercept specific for each location and $w_{\mathbf{s}}(t)$ is an autoregressive process that can be written as

$$w_{\mathbf{s}}(t) = \rho_1 w_{\mathbf{s}}(t-1) + \nu_{\mathbf{s}}(t), \tag{6}$$

with $t = 2, \ldots, T$ and $|\rho_1| < 1$. The term $\nu_{\mathbf{s}}$ is uncorrelated with $w_{\mathbf{s}}(t)$ and independent identically distributed as $\nu_{\mathbf{s}} \sim N(0, \sigma^2_\nu)$.

3.2.2. Spatio-temporal model (Model S-T)

This model is a spatio-temporal process with temporal dynamics as in Cameletti et al. [2013]. This type of model is commonly used for modelling air quality because of its flexibility in including time and space dependency, as well as the effect of covariates (see e.g. Fassò and Finazzi [2011] and Cocchi et al. [2007]). The mean function $\mu(\mathbf{s}, t)$ in (4) is given by

$$\mu(\mathbf{s}, t) = b_0 + z(\mathbf{s}, t), \tag{7}$$

where b_0 is an intercept that is common to all wind farms and constant in time and space. The term $z(\mathbf{s}, t)$ refers to a spatio-temporal process that varies in time with first order autoregressive dynamics

$$z(\mathbf{s}, t) = \rho_2 z(\mathbf{s}, t-1) + w(\mathbf{s}, t), \tag{8}$$

with $t = 2, \ldots, T$ and $|\rho_2| < 1$. Moreover, $w(s, t)$ is a zero-mean Gaussian field, assumed to be temporally independent with covariance function

$$\mathrm{Cov}(w(\mathbf{s}, t), w(\mathbf{s}', t')) = \begin{cases} \sigma^2_w C(h), & \text{if } t = t' \\ 0, & t \neq t' \end{cases}$$

for $\mathbf{s} \neq \mathbf{s}'$. The correlation function C depends on the locations $\mathbf{s}$ and $\mathbf{s}'$ through the distance $h = ||\mathbf{s} - \mathbf{s}'||$. This means that the process is assumed to be second-order stationary and isotropic (see Cressie [1992]). The marginal variance is $\mathrm{Var}(\mathbf{s}, t) = \sigma^2_w$ and $C(h)$ is the correlation function defined by the Matérn, given by

$$C(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa h)^{\nu} K_{\nu}(\kappa h), \tag{9}$$

where K_1 is the modified Bessel function of second kind, order ν. The parameter κ can be used to select the range, while ν is a smoothness parameter determining the mean-square differentiability of the underlying process. More precisely, the range is defined to be $r = \sqrt{8\nu}/\kappa$. Although the parameter ν is fixed to 1 for computational reasons, it remains flexible enough to handle a broad class of spatial variation Rue et al. [2009]. Applications with fixed parameter ν include Ingebrigtsen et al. [2014], Cameletti et al. [2013] and Munoz et al. [2013].

3.2.3. Temporal + Spatio-temporal model (Model ST+T)

This is a model defined by an autoregressive process at each location to capture the individual variability and a spatio-temporal process with temporal dynamics to take into account the spatial dependence among wind farms. Specifically, $\mu(\mathbf{s}, t)$ from (4) is defined as

$$\mu(\mathbf{s}, t) = b_0 + w_{\mathbf{s}}(t) + z(\mathbf{s}, t), \tag{10}$$

where b_0 is a fixed unknown intercept that is shared by all wind farms. The process $w_{\mathbf{s}}(t)$ is assumed to have autoregressive dynamics as defined in (6). Finally, $z(\mathbf{s}, t)$ is a spatio-temporal component that has the structure of (8) and its spatio-temporal covariance function is the same as in (9).

For all the models described above, a log-Gamma prior is assumed for the parameters in the Matérn covariance as well as for the precision parameters σ_e^2 and σ_ν^2. For the fixed effect b's we assume Gaussian priors. The correlations ρ's are specified over the parametrization $\log(\frac{1+\rho}{1-\rho})$ with prior Gaussian distributions.

3.3. Inference and prediction

The key feature of the models described above is that they can be handled within the theoretical and computational framework developed by Rue et al. [2009] and Lindgren et al. [2011]. The approach by Rue et al. [2009] allows us to directly compute accurate and fast approximations of the posterior marginals. In addition, the method by Lindgren et al. [2011] is computationally efficient for inferential purposes: instead of using a Gaussian random fields (GRF) with dense covariance matrix, the computations are carried out with a Gaussian Markov random field (GMRF) with sparse precision matrix. The original idea comes from the work of Whittle [1954] and Whittle [1963], where it is shown that the solution to the SPDE

$$(\kappa^2 - \Delta)^{\alpha/2} x(\mathbf{u}) = \mathcal{W}(\mathbf{u}), \quad \mathbf{u} \in \mathbb{R}^d, \alpha = \nu + D/2, \kappa > 0, \nu > 0, \tag{11}$$

is a GRF with Matérn covariance function. The innovation process $\mathcal{W}$ on the right hand side of (11) is Gaussian white noise and Δ is the Laplacian.

An approximation to the solution of the SPDE in (11) can be obtained using the finite element method (FEM), a numerical technique for solving partial differential equations Lindgren et al. [2011]. This is done by representing the infinite dimensional GRF by a linear combination of finite basis function

$$x(\mathbf{u}) = \sum_k \psi_k(\mathbf{u}) w_k \tag{12}$$

where the w_k's are random weights chosen so that the representation in (12) approximates the distribution of the solution to the SPDE in (11). The ψ_k's are basis functions defined on a triangulation of the domain, i.e. a subdivision into non-intersecting triangles. Figure 2 shows the triangulation of western Denmark data set described in Section 2.

[Figure 2 about here.]

Next, the posterior estimates of parameters and hyperparameters are computed using INLA Rue et al. [2009]. This method approximates the integral involved in the calculation of the marginal posterior distributions of the hyperparameters by Laplace approximation, making use of the Markov structure of the latent variables in the computation. We use the R-INLA package to perform inference and prediction. For more information on the package see http://www.r-inla.org.

4. FORECAST EVALUATION

4.1. Probabilistic forecasting scheme

We evaluate the predictive performance of the models described in Section 3, using a time moving window approach with data from western Denmark in 2009, so that each training set consists of $L = 2 \times 96 = 192$ observations, i.e., two days. In total, the model is fit to $364/2 = 182$ different data sets. We obtain forecasts for lead times $h = 1, \ldots, 20$, that is, from 15 minutes up to 5 hours following the training data. Notice that we have compared different lengths of data window L with respect to the root mean squared error (RMSE) and continuous ranked probability score (CRPS). The temporal model presented in Section 3.2.1 is very sensitive to the window length, such that less than two days of observations in the training set resulted in poor estimation at all lead times. On the other hand, the spatio-temporal models in Section 3.2.2 and 3.2.3 showed to be robust for different values of L, with small changes in the forecast performance for different training sets.

Moreover, because of the high-time resolution of the Danish wind power time series (15-minutes) and the dependency structure in space and time of Model S-T and Model ST+T, the fitting can be very computationally expensive. One way to deal with high-time resolution data is to define the model on a set of knots instead of all time points. Knot-based linear combinations are widely used to tackle computational problems in large data sets (e. g.

Paciorek [2007] and Wikle and Cressie [1999]). To fit the spatio-temporal component $z(\mathbf{s}, t)$ in (8), we define a set of equally spaced knots at every 12 data points (3 hours), such that the points in time are reduced to only 17 knots, instead of the original 192 observations. Note that the component $w_{\mathbf{s}}(t)$ in models Model T and Model ST+T is fitted to the complete training data, since it does not involve spatio-temporal interactions.

We evaluate probabilistic forecasts of wind power production from individual wind farms and aggregated.

Let $\hat{X}(\mathbf{s}_j, t+h)$ denote the random variable of the wind power forecast at wind farm $\mathbf{s}_j$ and lead time h. The aggregated forecast of wind power generation is taken as

$$\hat{X}_{A(t+h)} = \frac{\sum_{j=1}^{N} c_j \hat{X}(\mathbf{s}_j, t+h)}{\sum_{j=1}^{N} c_j}, \tag{13}$$

where c_j is the capacity of wind farm $\mathbf{s}_j$ and N is the number of wind farms. To find the pdf of the aggregated forecasts, $\hat{f}_{X_{A(t+h)}}$, the joint distribution for all wind farms $\{\hat{X}(\mathbf{s}_1, t+h), \hat{X}(\mathbf{s}_2, t+h), \dots \hat{X}(\mathbf{s}_N, t+h)\}$ needs to be assessed. Finally, point forecast of aggregated wind power production is obtained as the mean (or median) of $\hat{f}_{X_{A(t+h)}}$.

4.2. Point and probabilistic forecast scores

We assess the quality of predictive performance of the models proposed in Section 3 using both point and probabilistic forecast scores. We obtain point forecast at a specific location as the mean of the forecast density. For each lead time, point forecast of individual power is assessed using the root mean squared error (RMSE), where the mean is taken over all wind farms and data sets,

$$\text{RMSE}(t+h) = \sqrt{\frac{1}{DN} \sum_{i=1}^{D} \sum_{j=1}^{N} (x(\mathbf{s}_{ij}, t+h) - \hat{x}(\mathbf{s}_{ij}, t+h))^2} \tag{14}$$

where D is the number of data sets, N is the number of wind farms and $\hat{x}(\mathbf{s}_{ij}, t+h) = \gamma^{-1}(\hat{y}(\mathbf{s}_{ij}, t+h))$ is the predicted value of $x(\mathbf{s}_{ij}, t+h)$.

To evaluate the performance of forecast densities, we use the continuous ranked probability score (CRPS). Gneiting and Raftery [2007] showed that CRPS is a strictly proper scoring rule for the evaluation of probabilistic forecasts of a univariate quantity that assesses calibration and sharpness simultaneously Gneiting and Raftery [2007]. A lower score indicates a better density forecast. It is defined as

$$\text{CRPS}(F, x) = \int_{-\infty}^{\infty} (F(y) - \delta_{\{y \geq x\}})^2 dy \tag{15}$$

where F is the cumulative distribution function of the density forecast and y is the observation. With the available samples, we can approximate the mean CRPS at each lead time by

$$\begin{aligned} \text{CRPS}_{F,x}(t+h) = &\frac{1}{DN} \sum_{i=1}^{D} \sum_{j=1}^{N} \Big(\frac{1}{n} \sum_{k=1}^{n} |\hat{x}^{(k)}(\mathbf{s}_{ij}, t+h) - x(\mathbf{s}_{ij}, t+h)| \\ &- \frac{1}{2n^2} \sum_{k,l=1}^{n} |\hat{x}^{(k)}(\mathbf{s}_{ij}, t+h) - \hat{x}^{(l)}(\mathbf{s}_{ij}, t+h)| \Big), \end{aligned} \tag{16}$$

where n is the number of samples. Again, the mean CRPS is taken over all the wind farms and data sets in the training set.

Reliability, also referred to as calibration, of probabilistic forecasts is assessed with reliability diagrams. In a calibrated forecast, the observed levels should match the nominal levels for specific quantile forecasts, which results in points aligning with the diagonal in the reliability diagram. To construct reliability diagrams, we start by introducing an indicator variable $\mathcal{I}^{(\alpha)}(\mathbf{s}_{ij}, h)$, which is defined for a quantile forecast $\hat{q}^{(\alpha)}(\mathbf{s}_{ij}, t+h)$ issued at lead time

h and wind farm $\mathbf{s}_i$ of the training data j, with observed value $x(\mathbf{s}_{ij}, t+h)$ as follows

$$\mathcal{I}^{(\alpha)}(\mathbf{s}_{ij}, h) = \begin{cases} 1 & \text{if} \quad x(\mathbf{s}_{ij}, t+h) \leq \hat{q}^{(\alpha)}(\mathbf{s}_{ij}, t+h) \\ 0, & \text{otherwise} \end{cases}$$

The indicator variable $\mathcal{I}^{(\alpha)}(\mathbf{s}_{ij}, h)$ shows whether the actual outcome lies below the α quantile forecast (hits) or not (miss). Next, $n_{h,1}^{(\alpha)}$ denotes the sum of hits and $n_{h,0}^{(\alpha)}$ the sum of misses over all the realizations

$$n_{h,1}^{(\alpha)} = \sum_{i=1}^{D} \sum_{j=1}^{N} \mathcal{I}^{(\alpha)}(\mathbf{s}_{ij}, h) \quad \text{and} \quad n_{h,0}^{(\alpha)} = DN - n_{h,1}^{(\alpha)}.$$

An estimation $\hat{a}_h^{(\alpha)}$ of the actual coverage $a_h^{(\alpha)}$ is then obtained by calculating the mean of $\mathcal{I}^{(\alpha)}(\mathbf{s}_{ij}, h)$ over the N wind farms in the D validation sets

$$\hat{a}_h^{(\alpha)} = \frac{1}{DN} \sum_{i=1}^{D} \sum_{j=1}^{N} \mathcal{I}^{(\alpha)}(\mathbf{s}_{ij}, h) = \frac{n_{h,1}^{(\alpha)}}{n_{h,1}^{(\alpha)} + n_{h,0}^{(\alpha)}}. \tag{17}$$

Here, we use nominal levels from 5% to 95% in steps of 5%. Since the number of observations used to calculate the reliability diagrams is of limited size and the observed proportions are equal to the nominal ones only asymptotically Toth et al. [2003] Bröcker and Smith [2007], we follow the idea of Bröcker and Smith [2007] of generating consistency bars for reliability diagrams.

4.3. Evaluation scheme

We evaluate probabilistic forecasts of Danish wind power production from two different scenarios. First, we consider time forward forecast performances at the locations of the training set. The spatio-temporal models, i.e, Model S-T and Model ST+T, have the advantage of being able to provide forecasts where recent observations are not available.

Based on this, in a second evaluation scheme, we study the performances of spatially out-of-sample forecasts, which are based on k-fold cross-validation with $k = 5$. Notice that overall, 5 to 10-fold cross-validation is recommended as a good compromise between bias and variance (Breiman and Spector [1992]; Kohavi et al. [1995]). The forecast performance measures from the second scenario are obtained by combining the estimates from the 182 data sets in the training set.

Finally, we validate our results with a simulation study consisting of 200 simulated spatio-temporal data sets. In each data set, logit transformed wind power production measurements, $y(\mathbf{s}_i, t)$, are "observed" at 200 wind farms belonging to the wind power data set (see left plot of Figure 1). To mimic the case study based on the Danish wind data set, we simulate data every 15 minutes for 2 days and 5 hours. In total, there are $2 \times 96 + 20 = 212$ measurements taken at each location. All data sets are generated according to Model ST+T directly using the SPDE model construction. We use the set of parameters found for one specific data set of the training set from fitting Model ST+T to the logit transformed Danish wind power data.

5. RESULTS

In this section we show the results from a case study, where we use the models described in Section 3 to forecast individual and aggregated wind power in Denmark. As described in Section 4.3, we evaluate and discuss the performances of our models when we consider time forward forecasts at the locations of the training set. We call these spatio-temporal forecasts, and we also show the case of spatially out-of-sample forecasts, i.e, for wind farms that are not in the training set. Furthermore, we illustrate the results from a simulation study based on our case study. Details of the probabilistic forecasting scheme can be found in Section 4.1, while the methodology used to rank point and probabilistic forecasts is in Section 4.2.

5.1. Spatio-temporal forecast performance

Figure 3 summarizes the spatio-temporal forecast performances of the three models introduced in Section 3 in terms of RMSE and CRPS. As we can see from Figure 3 (a), Model T and Model ST+T outperformed Model S-T with respect to RMSE and CRPS when forecasting individual wind farms at lead times 1-6 (i.e, from 15 minutes up to 2 hours ahead). For higher lead times, the three models have similar performance. In terms of aggregated wind power production, Model T performed similar to Model ST+T in terms of point forecast (RMSE), but it has poor performance according to CRPS values, as shown in Figure 3 (b).

Reliability diagrams for each model at lead times $h = 1, 7, 13$ and 19 are presented in Figure 4. These diagrams compare the theoretical and the observed proportions of a set of quantiles from forecasts made at all wind farms and data sets in the training set. The forecasts at individual wind farms produced by the three models presented in Section 3 perform similarly well in terms of reliability, with points close to the diagonal for most quantiles, see Figure 4 (a). Since the number of observations used to calculate reliability diagrams is relatively small (182 data sets in the training set), consistency bars for the evaluation of forecasts from aggregated farms are also plotted, as shown in Figure 4 (b). The aggregated forecasts provided by Model ST+T are the best calibrated among the three models for most of the quantiles at all lead times, followed by Model S-T. Even though the performance of Model T is comparable with the performance of the other models in terms of aggregated forecast density mean (RMSE), we can see that this model does not produce reliable probabilistic forecasts for the aggregated data. This fact is more obvious for the lower quantiles; more than 50% of the observed aggregated forecasts are below the nominal 5% quantile at lead times $h = 7, 13$ and 19.

[Figure 3 about here.]

[Figure 4 about here.]

We further explore aggregated probabilistic forecasts from models in Section 3 with plots containing the 5%, 50% and 95% quantiles of the aggregated forecast densities together with the actual observed aggregated power produced at four different data sets in the training set, as shown in Figure 5. We noticed that Model T results in forecast densities that are consistently too narrow. On the other hand, Model ST+T provides the widest aggregated forecast densities among the three models in most of the data sets, which produces calibrated forecasts at all lead times. This is confirmed in Figure 4 (b) and will be further explored in the simulation studies in Section 5.3.

[Figure 5 about here.]

5.2. Spatially out-of-sample forecast performance

Figure 6 shows the out-of-sample forecast performances in terms of RMSE and mean CRPS for individual wind farms (a) and aggregated wind power (b). They are computed as the mean of the RMSE and CRPS from the 5-fold cross validations as described in Section 4.3. It can be seen that Model ST+T outperforms Model S-T at all lead times when predicting wind power at individual wind farms under RMSE and CPRS. When looking at aggregated out-of-sample forecasts, while for shorter lead times than 2 hours, Model S-T is better than Model ST+T in terms of RMSE, for longer horizons, Model ST+T out-performs Model S-T under the same score. In terms of CRPS, Model ST+T produces better aggregated forecasts at lead times 1-20 (i.e., from 15 minutes to 5 hours ahead).

Reliability diagrams at lead times $h = 1, 7, 13$ and 19 are presented in Figure 7. We observe from Figure 7 (a) that Model S-T and Model ST+T provide relatively well calibrated forecast densities for individual farms. In terms of aggregated forecasts, we can see from Figure 7 (b) that Model ST+T is calibrated, since the line is always within the consistency bars. On the other hand, aggregated forecast densities obtained with Model S-T are poorly calibrated for

quantiles lower than 0.75. Indeed, 20% of the observations are below the 5% forecast quantile at lead times 1, 7, 13 and 19.

[Figure 6 about here.]

[Figure 7 about here.]

5.3. Simulation study

From the data analysis in Section 5.1 and 5.2, we see that Model ST+T is the only model among the three that produces individual and aggregated calibrated forecasts. In this section, we simulate 200 data sets according to this model. We set the parameters equal to the estimates given by the fit of this model to one of the training data sets from our case study. More details on the evaluation scheme can be found in Section 4.3.

RMSE and mean CRPS of the three different models for forecasting simulated data at lead times 1-20 are shown in Figure 8. According to RMSE and CRPS for individual wind farms, the three models perform similarly, while, according to CRPS for aggregated forecasts, Model ST+T out-performs Model T and Model S-T.

Results from individual and aggregated forecasts calibration are shown in Figure 9. The forecasts from all three models are calibrated for individual wind farms, as shown in Figure 9 (a). We observe from Figure 9 (b) that the aggregated forecasts produced by the Model ST+T are better in terms of calibration than the forecasts from the other models, which is in agreement with the results from the analysis of the aggregated Danish wind power data, as shown in Figure 4 (b). In fact, the aggregated forecasts produced by Model ST+T are well calibrated at lead times $h = 1, 13$, and 19, since the line is always inside the consistency bars.

The simulations show that when we fit simulated data from Model ST+T using Model S-T, the spatial range r (see (9)) is underestimated. In fact, when data is generated with $r = 62.1$, the first and third quartiles of the 200 estimates of this parameter from Model

ST+T are 27.7 and 164.6, while with Model S-T the estimated quartiles are 25.2 and 28.0, respectively. Thus, a larger estimated spatial dependency results in a larger variance to the aggregated forecasts and makes it possible to borrow more information from close wind farms when doing out-of-sample predictions, causing the variance of a sum to increase. Hence, this explains both why the aggregated forecasts are not calibrated for Model S-T as well as why Model ST+T gives better spatially out-of-sample predictions than model S-T.

[Figure 8 about here.]

[Figure 9 about here.]

6. CONCLUSIONS

In this article we have presented hierarchical spatio-temporal models for obtaining probabilistic forecasts of wind power generation at multiple locations and lead times. We started with a time series model consisting of an autoregressive process with a location specific intercept. The results for individual probabilistic forecasts were satisfactory in terms of skill scores and reliability, however, the aggregated probabilistic forecasts were not calibrated. After finding the unsatisfactory results for the reliability of aggregated forecasts, we introduced two different spatio-temporal models. The first has a common intercept for all farms and a spatio-temporal model that varies in time with first order autoregressive dynamics and has spatially correlated innovations given by a zero mean Gaussian process with Matérn covariance. The second model has a common intercept, an autoregressive process to capture the local variability and the spatio-temporal term. To deal with the non-Gaussianity of wind power series, a parametric framework for distributional forecasts based on the logit-normal transformation was used.

In a case study, the proposed models have been used to produce probabilistic forecasts of wind power at wind farms in western Denmark from 15 minutes up to 5 hours ahead for a test

period of one year. Using the SPDE approach that is implemented in the R-INLA library, we obtained fast and accurate forecasts of wind power generation at wind farms where data is available, but also at a larger portfolio including wind farms at locations that are not included in the training set. We provided detailed analysis on the forecast performances based on appropriate metrics tailored for probabilistic forecasts. To better understand the properties of our methods, we analysed artificial data sets from a simulation study.

Our results showed that all the proposed approaches produce calibrated short-term forecasts for individual wind farms. However, we found that modeling spatial dependency is required to achieve calibrated aggregated probabilistic forecasts. Indeed, our case study showed that spatial dependency is important for aggregated properties, and individual forecasts do not reveal this. Moreover, when we simulated from the spatio-temporal model containing an autoregressive term (Model ST+T), we obtained results that are in accordance with our case study, where the proposed models performed equally well for individual forecasts, while aggregated probabilistic forecasts benefit from having a spatio-temporal model with the autoregressive term. Model ST+T was introduced due to unsatisfactory reliability for the aggregated forecasts. Hence, evaluating aggregated forecasts can be a tool for investigating and improving models, even when spatially out-of-sample forecasts are the purpose of the modelling. Indeed, results from spatially out-of-sample forecast performances showed that when predicting wind power at new locations that are not included in the training set, having the autoregressive term in the spatio-temporal model improved the forecast performance.

This work was motivated by the need to produce accurate short term probabilistic forecasts at multiple wind farms and lead times, which will ultimately be applied on a national scale. A possible extension of the models described in this work is to include weather forecast information in the linear predictor. This approach usually requires ensemble forecasts to be generated from sophisticated numerical weather prediction (NWP) models and has shown

to produce reliable wind power forecasts up to 10 days ahead Taylor et al. [2009].

ACKNOWLEDGEMENTS

The authors are grateful to Energinet.dk (system operator in Denmark) for providing the data and to Robin Girard at MinesParistech, France for checking the quality of the data. The authors also thank the Danish Strategic Council for Strategic Research through the project 5s-Future Electricity Markets (No. 12-132636/DSF), Research Concil of Norway, project 250362 and CAPES for support.

REFERENCES

Thomas Ackermann. *Wind power in power systems*. John Wiley & Sons, 2005.

Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal bayesian models with R-INLA*. John Wiley & Sons, 2015.

Leo Breiman and Philip Spector. Submodel selection and evaluation in regression. the x-random case. *International Statistical Review/Revue Internationale de Statistique*, pages 291–319, 1992.

John Bjørnar Bremnes. Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, 7 (1):47–54, 2004.

Jochen Bröcker and Leonard A Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

Michela Cameletti, Finn Lindgren, Daniel Simpson, and Håvard Rue. Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, 97(2):109–131, 2013.

João Paulo da Silva Catalão, Hugo Miguel Inácio Pousinho, and Víctor Manuel Fernandes Mendes. Short-term wind power forecasting in portugal by neural networks and wavelet transform. *Renewable Energy*, 36(4):1245–1251, 2011.

Daniela Cocchi, Fedele Greco, and Carlo Trivisano. Hierarchical space-time modelling of pm 10 pollution. *Atmospheric Environment*, 41(3):532–542, 2007.

Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

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

Alessandro Fassò and Francesco Finazzi. Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics*, 22(6):735–748, 2011.

Ulrich Focken, Matthias Lange, Kai Mönnich, Hans-Peter Waldl, Hans Georg Beyer, and Armin Luig. Short-term prediction of the aggregated power output of wind farmsa statistical analysis of the reduction of the prediction error by spatial smoothing effects. *Journal of Wind Engineering and Industrial Aerodynamics*, 90(3):231–246, 2002.

Robin Girard, K Laquaine, and Georges Kariniotakis. Assessment of wind power predictability as a decision factor in the investment phase of wind farms. *Applied Energy*, 101:609–617, 2013.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting, Kristin Larson, Kenneth Westrick, Marc G Genton, and Eric Aldrich. Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method. *Journal of the American Statistical Association*, 101(475):968–979, 2006.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Rikke Ingebrigtsen, Finn Lindgren, and Ingelin Steinsland. Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, 8:20–38, 2014.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.

Ada Lau and Patrick McSharry. Approaches for multi-step density forecasts with application to aggregated wind power. *The Annals of Applied Statistics*, pages 1311–1341, 2010.

Emmanuel Lesaffre, Dimitris Rizopoulos, and Roula Tsonaka. The logistic transform for bounded outcome scores. *Biostatistics*, 8(1):72–85, 2007.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Petroula Louka, Georges Galanis, Nils Siebert, Georges Kariniotakis, Petros Katsafados, I Pytharoulis, and G Kallos. Improvements in wind speed forecasts for wind power prediction purposes using kalman filtering. *Journal of Wind Engineering and Industrial Aerodynamics*, 96(12):2348–2362, 2008.

Facundo Munoz, M Grazia Pennino, David Conesa, Antonio López-Quílez, and José M Bellido. Estimation and prediction of the spatial occurrence of fish species using bayesian latent gaussian models. *Stochastic Environmental Research and Risk Assessment*, 27(5):1171–1180, 2013.

Christopher J Paciorek. Bayesian smoothing with gaussian processes using fourier basis functions in the spectralgp package. *Journal of Statistical Software*, 19(2):nihpa22751, 2007.

Pierre Pinson. Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):555–576, 2012.

Pierre Pinson and George Kariniotakis. Conditional prediction intervals of wind power generation. *IEEE Transactions on Power Systems*, 25(4):1845–1856, 2010.

Pierre Pinson and Julija Tastu. Discrimination ability of the energy score. Technical report, Technical University of Denmark, 2013.

Sheldon Ross. *A first course in probability*. Pearson, 2015.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Michael Scheuerer and Thomas M Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

Julija Tastu, Pierre Pinson, Ewelina Kotwa, Henrik Madsen, and Henrik Aa Nielsen. Spatio-temporal analysis and modeling of short-term wind power forecast errors. *Wind Energy*, 14(1):43–60, 2011.

James W Taylor, Patrick E McSharry, and Roberto Buizza. Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*, 24(3):775–782, 2009.

Thordis L Thorarinsdottir, Michael Scheuerer, and Christopher Heinz. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, 25(1):105–122, 2016.

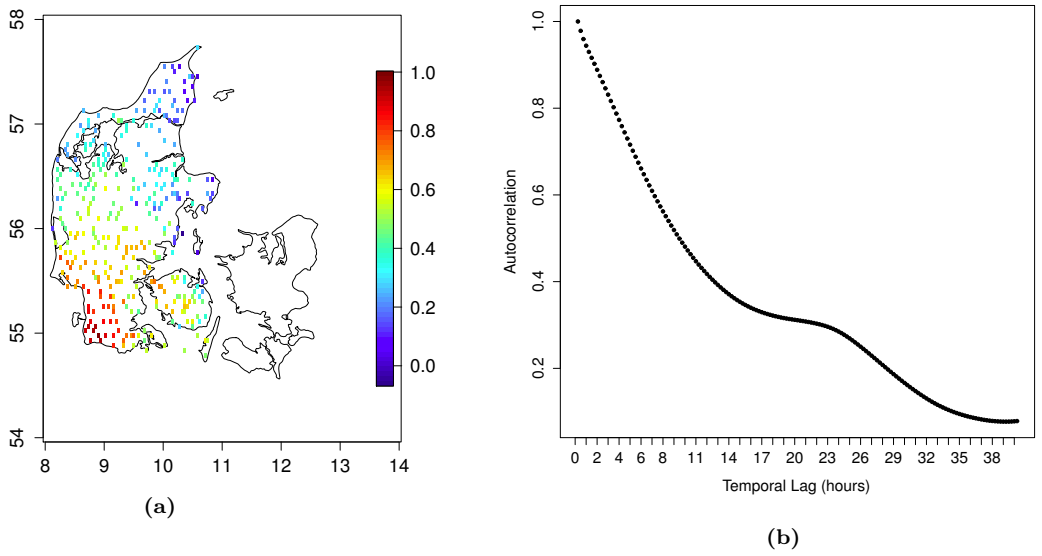
Zoltan Toth, Oliver Talagrand, Guillem Candille, and Yuejian Zhu. Forecast verification: A practitioners guide in atmospheric science, 2003.

Peter Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

Peter Whittle. Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):974–994, 1963.

Christopher K Wikle and Noel Cressie. A dimension-reduced approach to space-time kalman filtering. *Biometrika*, 86(4):815–829, 1999.

FIGURES



(a) **(b)**

Figure 1. (a) Map of spatial correlation of wind power production between one wind farm located in the southern part of western Denmark and the remaining wind farms. The correlations between wind farms in a closer proximity are clearly higher than between wind farms that are farther apart. (b) Mean autocorrelation function of wind power production at wind farms located in western Denmark. The autocorrelations decay slowly.

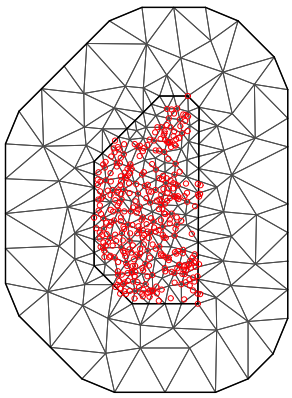
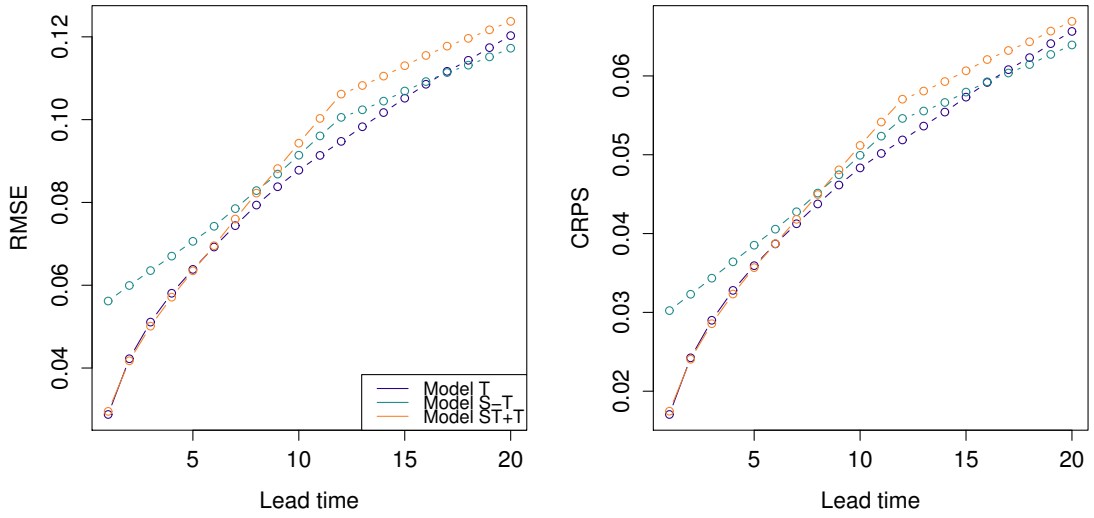
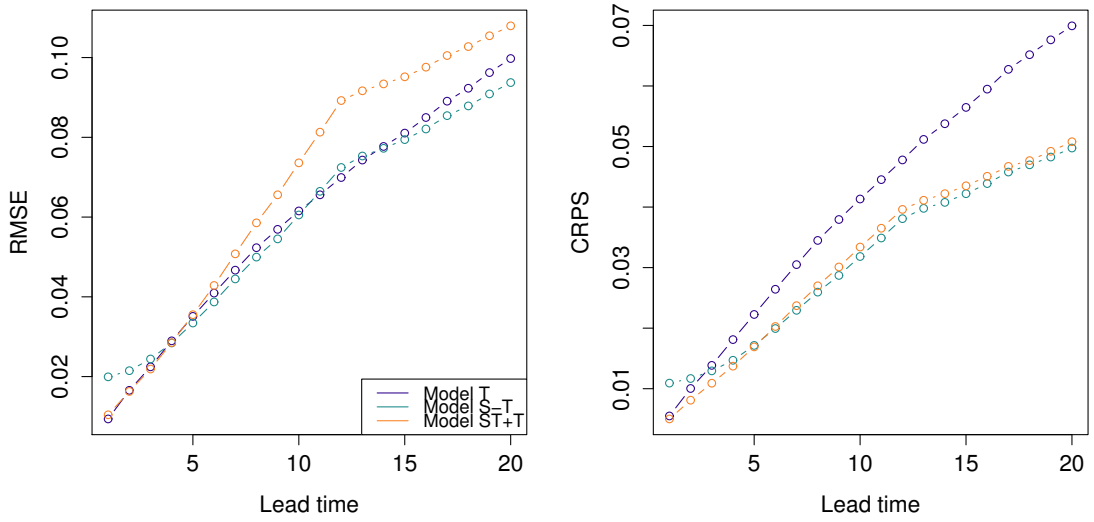


Figure 2. The western Denmark triangulation. The red dots denote the observation locations of the wind power production data.

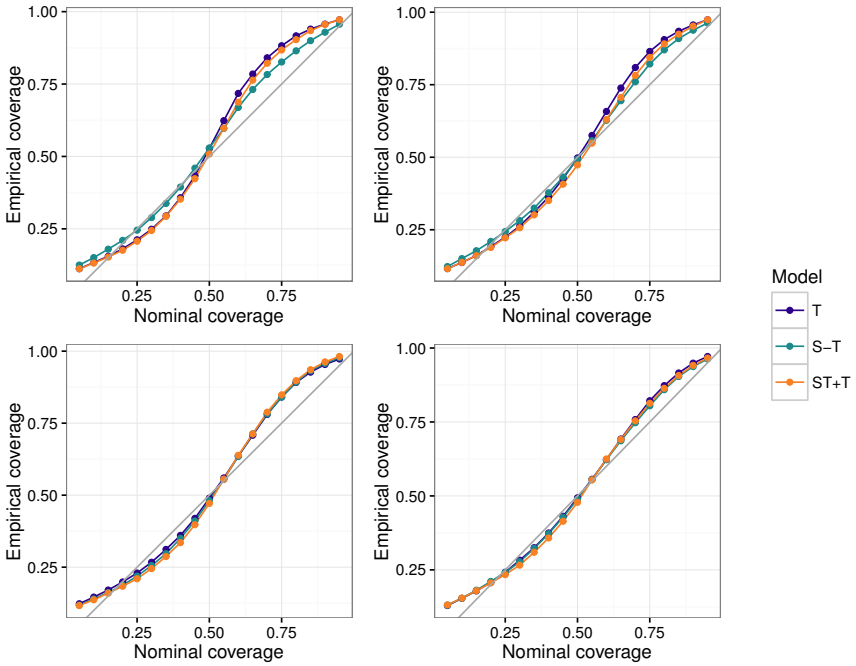


(a)

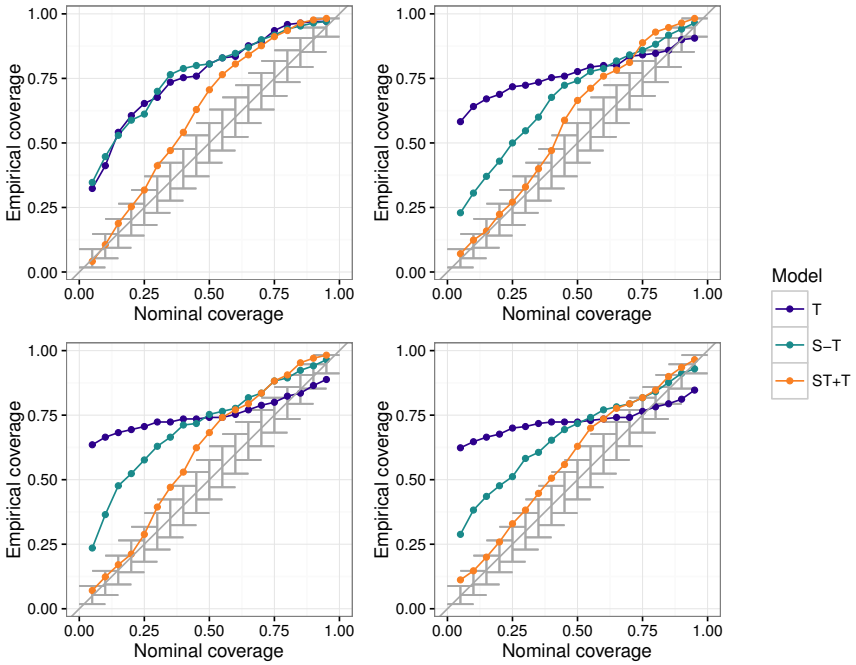


(b)

Figure 3. RMSE and CRPS (as % of nominal power) of spatio-temporal wind power forecasts at lead times $1, \ldots, 20$ (i.e., from 15 minutes up to 5 hours) for Model T (blue), Model S-T (green) and Model ST+T (orange). (a) Forecasts for individual wind farms. (b) Forecasts for aggregated wind farms.



(a)



(b)

Figure 4. Reliability diagram of spatio-temporal wind power forecasts at lead time 1 (*Top left*), 7 (*Top right*) , 13 (*Bottom left*) and 19 (*Bottom right*). The diagrams were calculated using Model T (blue), Model S-T (green) and Model ST+T (orange). (a) Forecasts for individual wind farms. (b) Forecasts for aggregated wind farms.

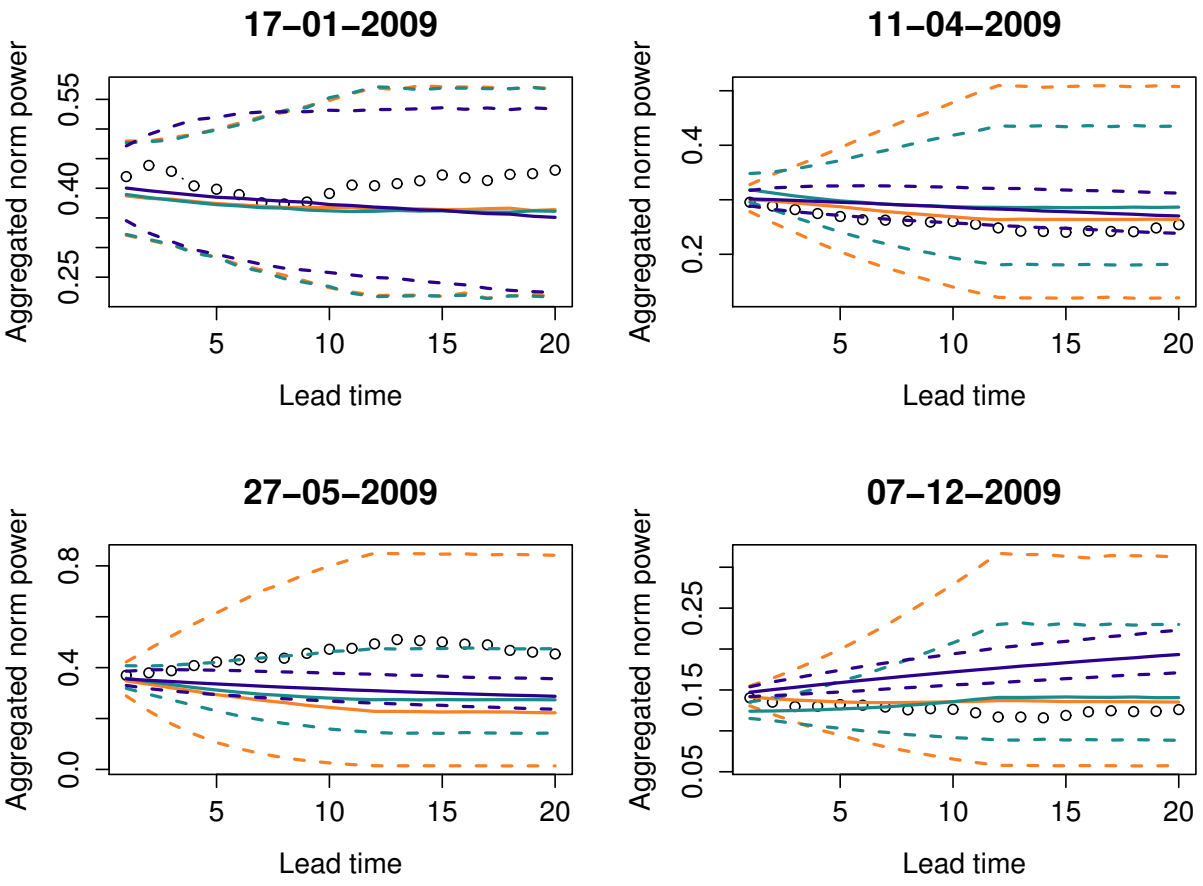
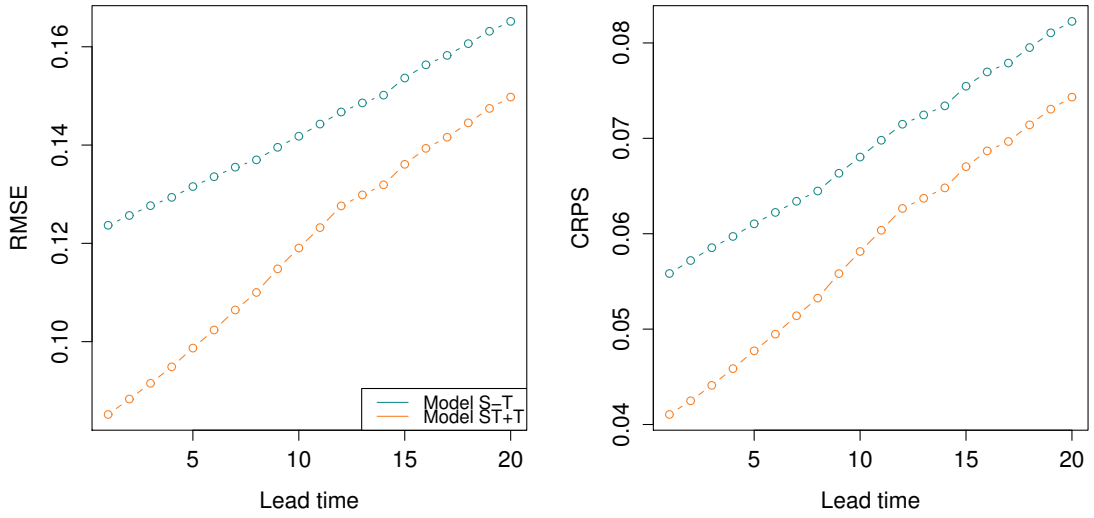
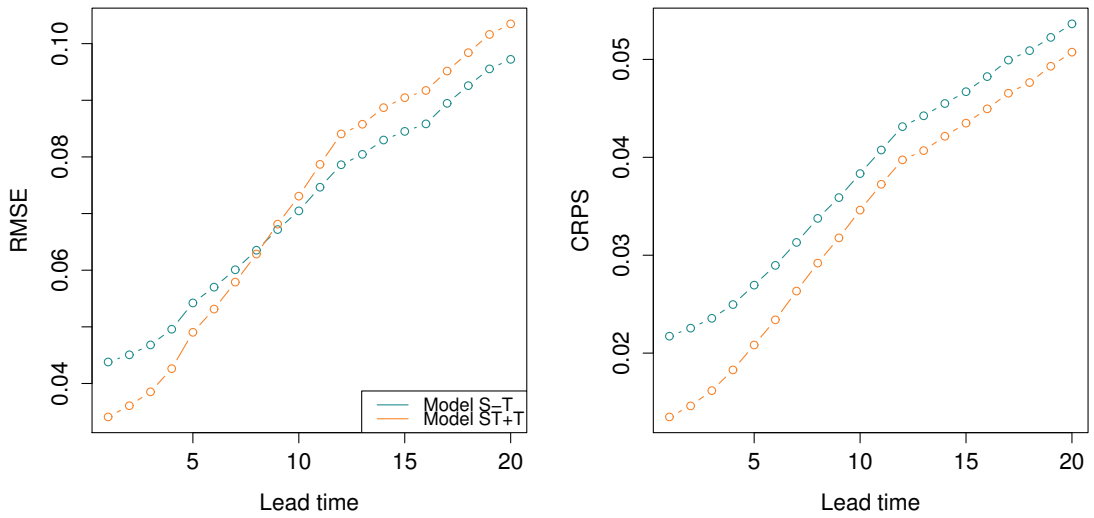


Figure 5. 5% and 95% quantiles (dashed lines), as well as the median (solid lines) of the aggregated forecast densities from four different data sets in the training set, together with the actual observed aggregated power produced (circles) at lead times 1-20 (i.e., from 15 minutes up to 5 hours). The forecast densities correspond to Model T (blue), Model S-T (green) and Model ST+T (orange). An example of a data set where all the models have forecast densities that cover the actual aggregated production is shown in the *Top left* plot. In the *Top right* plot, the observations lie close to the median of the forecast densities from Model S-T and Model ST+T, but close to the 5% quantile of the forecast density from Model T. *Bottom left* and *Bottom right* plots illustrate cases where Model T has forecast densities that are too narrow and fail to predict the aggregated wind power, while the forecasts from Model ST+T provide densities that are wide enough to cover the true value at all lead times.

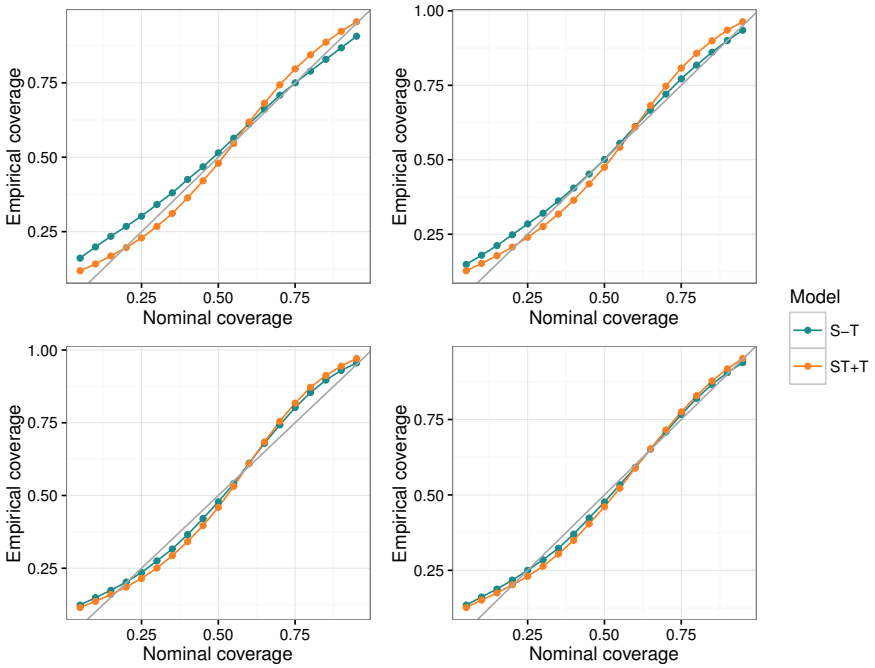


(a)

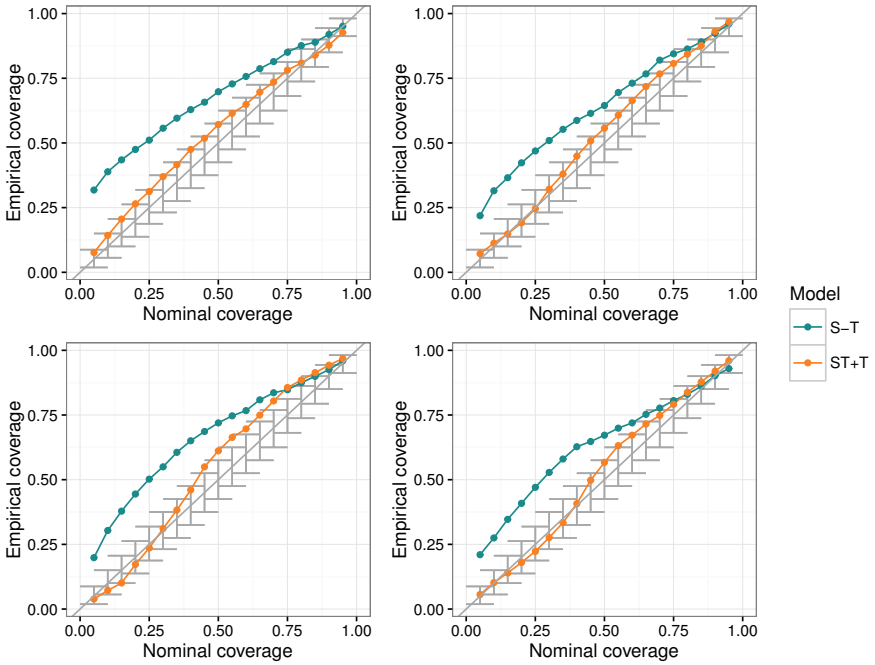


(b)

Figure 6. RMSE and CRPS (as % of nominal power) of spatially out-of-sample wind power forecasts at lead times $1, \ldots, 20$ (i.e., from 15 minutes up to 5 hours) for Model T (blue), Model S-T (green) and Model ST+T (orange). (a) Forecasts for individual wind farms. (b) Forecasts for aggregated wind farms.

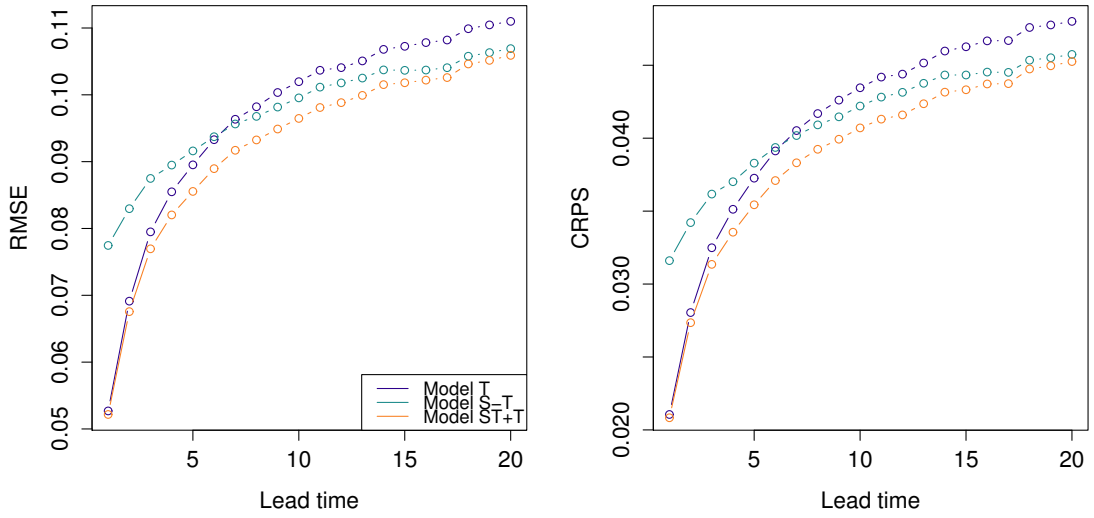


(a)

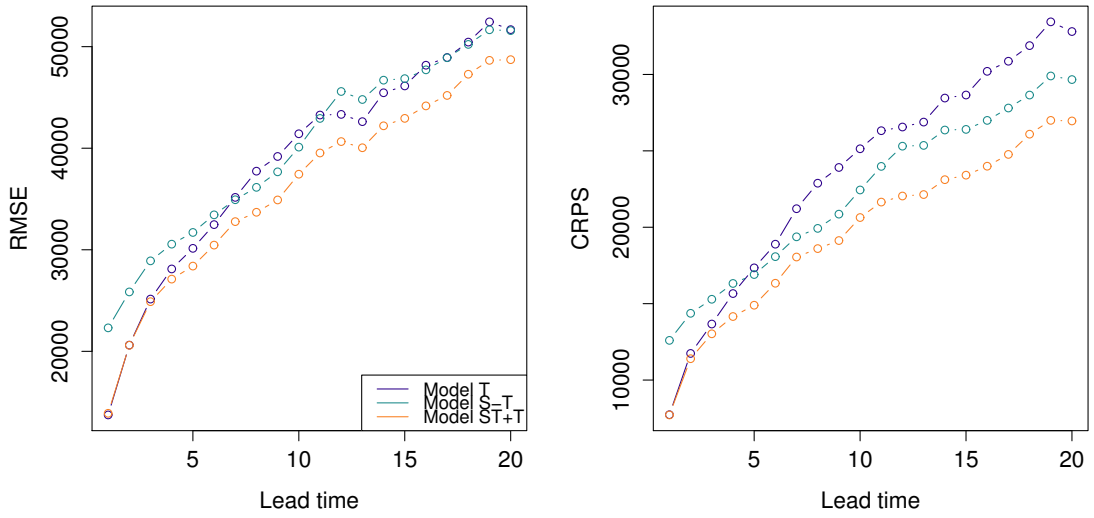


(b)

Figure 7. Reliability diagram of spatially out-of-sample wind power forecasts at lead time 1 (*Top left*), 7 (*Top right*) , 13 (*Bottom left*) and 19 (*Bottom right*). The diagrams were calculated using Model T (blue), Model S-T (green) and Model ST+T (orange). (a) Forecasts for individual wind farms. (b) Forecasts for aggregated wind farms.

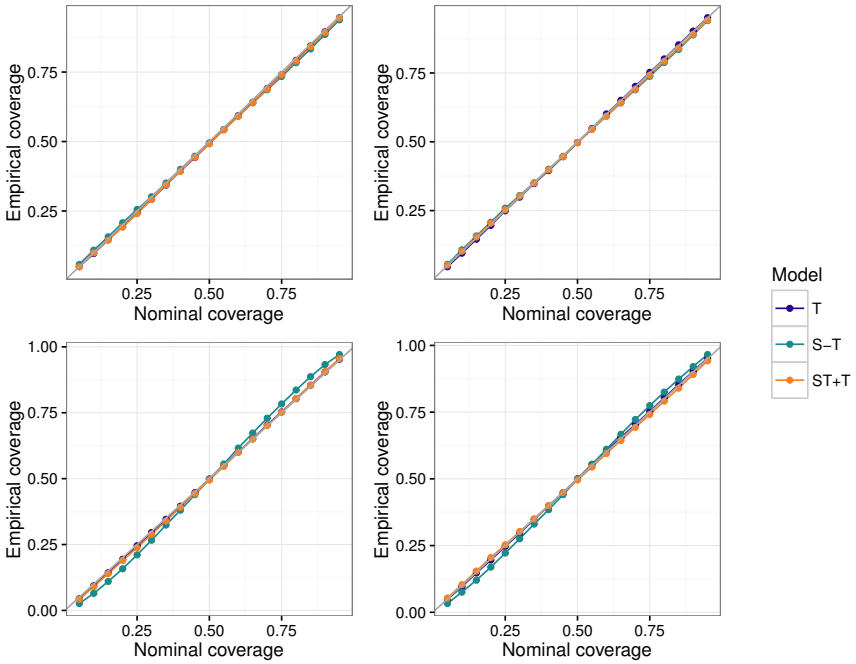


(a)

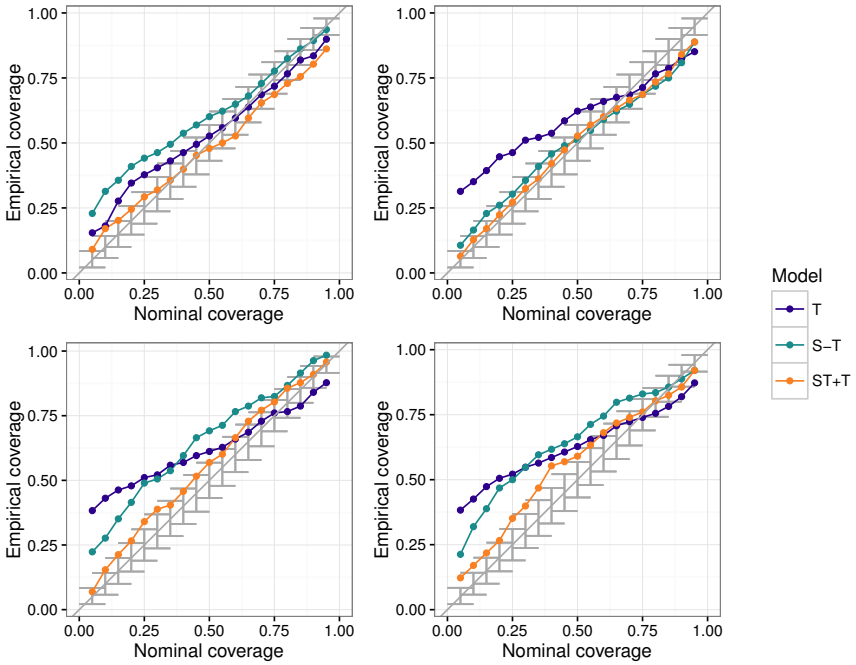


(b)

Figure 8. RMSE and CRPS (as % of nominal power) of forecasts from simulated data at lead times $1, \ldots, 20$ (i.e., from 15 minutes up to 5 hours) for Model T (blue), Model S-T (green) and Model ST+T (orange). (a) Forecasts for individual wind farms. (b) Forecasts for aggregated wind farms.



(a)



(b)

Figure 9. Reliability diagram for forecasts from simulated data at lead time 1 (*Top left*), 7 (*Top right*), 13 (*Bottom left*) and 19 (*Bottom right*). The diagrams were calculated using Model T (blue), Model S-T (green) and Model ST+T (orange). (a) Forecasts for individual wind farms. (b) Forecasts for aggregated wind farms.