

Adaptive Generalized Logit-Normal Distributions for Wind Power Short-Term Forecasting

Amandine Pierrot
Technical University of Denmark
Kgs Lyngby, Denmark
amapi@dtu.dk

Pierre Pinson
Technical University of Denmark
Kgs Lyngby, Denmark
ppin@dtu.dk

Abstract—There is increasing interest in very short-term and higher-resolution wind power forecasting (from mins to hours ahead), especially offshore. Statistical methods are of utmost relevance, since weather forecasts cannot be informative for those lead times. Those approaches ought to account for the fact wind power generation as a stochastic process is non-stationary, double-bounded (by zero and the nominal power of the turbine) and non-linear. Accommodating those aspects may lead to improving both point and probabilistic forecasts. We propose to focus on generalized logit-normal distributions, which are naturally suitable and flexible for double-bounded and non-linear processes. Relevant parameters are estimated via maximum likelihood inference. Both batch and online versions of the estimation approach are described – the online version permitting to additionally handle non-stationarity through parameter variation. The approach is applied and analysed on the test case of the Anholt offshore wind farm in Denmark, with emphasis placed on 10-min-ahead forecasting.

Index Terms—Wind power, Probabilistic forecasting, Dynamic models, Bounded time-series

I. INTRODUCTION

Forecasting is of utmost importance to the integration of renewable energy into power systems and electricity markets. The attention of energy forecasting has increased tremendously over the years [1]. For instance, thinking of short-term operational problems, transmission system operators (TSOs) have to operate reserves optimally to keep the system in balance at reasonable costs. Indeed, in Denmark, the TSO has some time argued the 10-min lead time as the most important since wind power fluctuations at this horizon particularly affect the system balance, see [2] for instance. Emphasis here is on offshore wind power forecasting, since those short-term fluctuations in power generation are most significant offshore. Even though most efforts in wind power forecasting are placed on lead times from hours to days, many are investing in alternative approaches to improve the accuracy of very short-term forecasts, for instance leveraging detailed turbine-level

The research leading to this work is being carried out as a part of the Smart4RES project (European Union’s Horizon 2020, No. 864337). The sole responsibility of this publication lies with the authors. The European Union is not responsible for any use that may be made of the information contained therein. The authors additionally acknowledge Ørsted for providing the data for the Anholt offshore wind farm.

data [3]. Those very short-term lead times are not only crucial but also those it is the most difficult to improve the forecasts for, especially compared to the simple but very effective persistence benchmark. Forecasts characterize and reduce but do not eliminate uncertainty. Thus forecasts should be probabilistic in nature taking the form of probability distributions, following the argument of [4] among others.

Wind power generation is a stochastic process which is double-bounded by nature, both by zero when there is no production at all, and by the nominal power for high-enough wind speeds. For short-term forecasting, statistical methods have proved to be more skilled and accurate. However, those methods often rely on a Gaussian assumption – which cannot be appropriate for a double-bounded variable. In [5], it is proposed to move from the classical Gaussian assumption to a framework where the wind power variable follows a generalized logit-normal distribution. In this framework though, not all the parameters of the distribution are estimated and tracked, the shape parameter being selected upon cross-validation.

Consequently here, we propose to revisit this work and to estimate all the parameters of the generalized logit-normal distributions within a maximum likelihood framework. Such a framework is particularly suitable to obtain skilled probabilistic forecasts. In addition, emphasis is placed on describing both batch and recursive estimation approaches, in order to go towards an online learning approach as a basis for probabilistic forecasting. For a nice introduction to online learning, the reader is referred to [6]. Online learning (with exponential forgetting) makes it possible to accommodate the non-stationarity of wind power generation time-series. The models and estimation framework are first presented in Section II, and the resulting algorithms in Section III. They are then applied to 10-min-ahead point and probabilistic forecasting at the Anholt offshore wind farm in Section IV. Finally some concluding remarks and prospects are given in Section V.

II. MODEL AND ESTIMATION FRAMEWORK

A. Generalized Logit-Normal Distribution and its Parameters

For an original random variable $X \in (0, 1)$, the generalized logit transform Y is given by

$$Y = \gamma(X; \nu) = \ln \left(\frac{X^\nu}{1 - X^\nu} \right), \quad \nu > 0, \quad (1)$$

where ν is the shape parameter. When Y follows a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the original variable X follows a generalized logit-normal distribution $L_\nu(\mu, \sigma^2)$, see [5]. The probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{\nu}{x(1-x^\nu)} \exp\left[-\frac{1}{2}\left(\frac{\gamma(x; \nu) - \mu}{\sigma}\right)^2\right]. \quad (2)$$

Let X the wind power random variable. We want ν such as the transform variable Y is as close as possible to a Gaussian variable which then can be forecast in a Gaussian framework. As we have access to some realizations (x_t) of X and to the analytical expression of its density, we can then maximize the probability of observing the data (x_t) depending on ν, μ and σ^2 , that is estimate all the parameters of the distribution (2) using maximum likelihood inference.

In the case of wind power generation, the observations (x_t) are strongly correlated. We thus assume that $Y_t|Y_{t-1}, \dots, Y_{t-p} \sim \mathcal{N}(\mu_t, \sigma^2)$ where $\mu_t = \sum_{k=1}^p \phi_k Y_{t-k}$, that is the distribution of $X_t|X_{t-1}, \dots, X_{t-p}$ is a generalized logit-normal distribution of density

$$\frac{1}{\sqrt{2\pi\sigma^2}} \frac{\nu}{x_t(1-x_t^\nu)} \exp\left[-\frac{1}{2}\left(\frac{y_t - \sum_{k=1}^p \phi_k y_{t-k}}{\sigma}\right)^2\right], \quad (3)$$

where $y_t = \gamma(x_t; \nu)$. While the density in (3) is defined only for $x \in (0, 1)$, the wind power generation can take values 0 and 1. We thus choose to look at the observations $x_t \in [0, 1]$ as a coarsened version of X , see [7]. This coarsened data framework has been formalized by [8] and [9].

B. Maximum Likelihood Inference

Let $\Phi = (\phi_1, \dots, \phi_p)^\top \in \mathbb{R}^p$. The maximum likelihood inference is based on the likelihood function, given by

$$L(\nu, \Phi, \sigma^2 | \mathbf{x}) = \prod_{t=1}^N f(x_t | x_{t-1}, \dots, x_{t-p}, \nu, \Phi, \sigma^2), \quad (4)$$

which is the probability of the observed data under the model f , assuming the realizations of $X_t|X_{t-1}, \dots, X_{t-p}$ are independent and identically distributed. We think of $L(\nu, \Phi, \sigma^2 | \mathbf{x})$ as a function of ν, Φ and σ^2 , the data (x_t) being fixed. The method of maximum likelihood chooses the values $(\nu, \Phi, \sigma^2) = (\hat{\nu}, \hat{\Phi}, \hat{\sigma}^2)$ to maximize $L(\nu, \Phi, \sigma^2 | \mathbf{x})$. The logarithm of L being easier to maximize, especially when exponential distributions are involved, it is used instead of the likelihood. For model f the negative log-likelihood function is

$$\begin{aligned} \tilde{l}(\nu, \Phi, \sigma^2 | \mathbf{x}) &= \frac{N-p}{2} \ln(\sigma^2) - (N-p) \ln(\nu) \\ &+ \sum_{t=p+1}^N \ln(1-x_t^\nu) \\ &+ \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{Y}\Phi)^\top (\mathbf{y} - \mathbf{Y}\Phi) + C, \end{aligned} \quad (5)$$

where $\mathbf{y} = (y_{p+1}, \dots, y_N)^\top \in \mathbb{R}^{N-p}$, \mathbf{Y} is a matrix with columns $B\mathbf{y}, B^2\mathbf{y}, \dots, B^p\mathbf{y} \in \mathbb{R}^{(N-p) \times p}$, B being the backshift

operator, C is a constant which does not depend on ν, Φ or σ^2 . Computing the first derivatives of (5) w.r.t. the parameters of the distribution we can retrieve stationary points. It is worth noting that those points are minimizers only if the negative log-likelihood is convex. Taking the derivative of (5) w.r.t. Φ , resp. σ^2 , and setting it equal to zero, leads to the usual maximum likelihood estimators

$$\hat{\Phi} = (\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{y}, \quad \hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{Y}\hat{\Phi})^\top (\mathbf{y} - \mathbf{Y}\hat{\Phi})}{N-p}. \quad (6)$$

Taking the derivative of (5) w.r.t. ν , we thus need to solve

$$-\frac{N-p}{\nu} - \sum_{t=p+1}^N \frac{\ln(x_t) x_t^\nu}{1-x_t^\nu} + \frac{(\mathbf{u} - \mathbf{U}\hat{\Phi})^\top (\mathbf{y} - \mathbf{Y}\hat{\Phi})}{\sigma^2} = 0, \quad (7)$$

where $\mathbf{u} = \frac{\partial \mathbf{y}}{\partial \nu}$ with $u_t = \ln(x_t)(1 + \frac{x_t^\nu}{1-x_t^\nu})$, $\mathbf{U} = \frac{\partial \mathbf{Y}}{\partial \nu}$ with columns $B\mathbf{u}, B^2\mathbf{u}, \dots, B^p\mathbf{u}$. Unlike $\hat{\Phi}$ and $\hat{\sigma}^2$, $\hat{\nu}$ does not have a closed-form solution and a descent algorithm is then to be used to solve (7).

III. BATCH AND RECURSIVE ALGORITHMS

A. Batch Algorithm

We use both the closed-form solutions in (6) for $\hat{\Phi}$ and $\hat{\sigma}^2$, and a Newton-Raphson algorithm to solve (7) in order to estimate the shape parameter ν . The computation of the Newton-Raphson step requires the second derivative of (5) w.r.t. ν , i.e.

$$\begin{aligned} \frac{\partial^2 \tilde{l}}{\partial \nu^2} &= \frac{N-p}{\nu^2} - \sum_{t=p+1}^N \ln(x_t)^2 \frac{x_t^\nu}{(1-x_t^\nu)^2} \\ &+ \frac{(\mathbf{v} - \mathbf{V}\hat{\Phi})^\top (\mathbf{y} - \mathbf{Y}\hat{\Phi})}{\sigma^2} + \frac{\|\mathbf{u} - \mathbf{U}\hat{\Phi}\|_2^2}{\sigma^2}, \end{aligned} \quad (8)$$

where $\mathbf{v} = \frac{\partial \mathbf{u}}{\partial \nu}$ with $v_t = u_t \ln(x_t) \frac{x_t^\nu}{1-x_t^\nu}$, $\mathbf{V} = \frac{\partial \mathbf{U}}{\partial \nu}$ with columns $B\mathbf{v}, B^2\mathbf{v}, \dots, B^p\mathbf{v}$.

The full algorithm is described in Algorithm 1 and has showed very fast convergence on numerous simulations of samples distributed according to the generalized logit-normal distribution with different values of Φ, σ^2 and ν .

Algorithm 1 Batch MLE with diagonalization

Set $i \leftarrow 1$ and let $\nu_1 = 1, \epsilon = 0.001$.

repeat

1. *Update.* $\Phi_i = (\mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{y}; \sigma_i^2 = \frac{(\mathbf{y} - \mathbf{Y}\Phi_i)^\top (\mathbf{y} - \mathbf{Y}\Phi_i)}{N-p}$.
2. *Compute the Newton step and decrement for ν .*
 $\Delta \nu_{nt} = -\frac{\nabla_\nu \tilde{l}}{\nabla_\nu^2 \tilde{l}}; \lambda^2 = \frac{(\nabla_\nu \tilde{l})^2}{\nabla_\nu^2 \tilde{l}}$.
3. *Stopping criterion.* **quit** if $\lambda^2/2 \leq \epsilon$.
4. *Line search.* Choose step size t by backtracking line search.
5. *Update.* $\nu_{i+1} = \nu_i + t\Delta \nu_{nt}$.

until termination test satisfied.

B. Recursive Algorithm

The batch algorithm is well suited if the data are known to be stationary to second order, that is assuming the parameters of the distribution (2) do not change over the course of time. But if, as we suspect in the case of wind power data, the time series is not stationary and the parameters are not constant, then the batch algorithm is not appropriate and alternative solutions are required. Recursive estimation allows for such a parametric time-variability and provides information not only on the existence of non-stationarity but also on the possible nature of the parametric variations (see e.g. [10]).

As the inference relies on the likelihood function, it is straightforward to derive a recursive algorithm which only requires the first derivatives of (5) w.r.t. to the parameters. Let introduce $\hat{\Theta}_t = (\hat{\Phi}_t, \hat{\sigma}_t^2, \hat{\nu}_t)$ the estimate of the parameters at time t . The recursive estimation procedure relies on a Newton-Raphson step for obtaining the estimate $\hat{\Theta}_t$ as a function of the previous estimate $\hat{\Theta}_{t-1}$, see e.g. [11] and [12]. Let introduce the time-dependent negative log-likelihood objective function to be minimized at time t

$$S_t(\Theta) = -\frac{1}{n_\alpha} \sum_{j=p+1}^t \alpha^{t-j} \ln(f_j(\Theta)), \quad (9)$$

where $f_j(\Theta) = f(x_j | x_{j-1}, \dots, x_{j-p}; \Theta)$, α is a forgetting factor, $\alpha \in (0, 1)$, allowing for exponential forgetting of past observations, $n_\alpha = \frac{1}{1-\alpha}$ is the effective number of observations used for normalizing the weighted negative log-likelihood function. Applying one Newton-Raphson step we have

$$\hat{\Theta}_t = \hat{\Theta}_{t-1} - \frac{\nabla_{\Theta} S_t(\hat{\Theta}_{t-1})}{\nabla_{\Theta}^2 S_t(\hat{\Theta}_{t-1})}. \quad (10)$$

As

$$\begin{aligned} \nabla_{\Theta} S_t(\hat{\Theta}_{t-1}) &= \alpha \nabla_{\Theta} S_{t-1}(\hat{\Theta}_{t-1}) \\ &\quad - (1-\alpha) \nabla_{\Theta} \ln(f_t(\hat{\Theta}_{t-1})), \end{aligned} \quad (11)$$

assuming that $\hat{\Theta}_{t-1}$ minimizes $S_{t-1}(\Theta)$, we get

$$\nabla_{\Theta} S_t(\hat{\Theta}_{t-1}) = -(1-\alpha) \nabla_{\Theta} \ln(f_t(\hat{\Theta}_{t-1})). \quad (12)$$

From (11) we also get

$$\begin{aligned} \nabla_{\Theta}^2 S_t(\hat{\Theta}_{t-1}) &= \alpha \nabla_{\Theta}^2 S_{t-1}(\hat{\Theta}_{t-1}) \\ &\quad - (1-\alpha) \nabla_{\Theta}^2 \ln(f_t(\hat{\Theta}_{t-1})). \end{aligned} \quad (13)$$

As

$$\begin{aligned} \nabla_{\Theta}^2 \ln(f_t(\hat{\Theta}_{t-1})) &= \frac{\nabla_{\Theta}^2 f_t(\hat{\Theta}_{t-1})}{f_t(\hat{\Theta}_{t-1})} \\ &\quad - \frac{\nabla_{\Theta} f_t(\hat{\Theta}_{t-1}) (\nabla_{\Theta} f_t(\hat{\Theta}_{t-1}))^T}{f_t(\hat{\Theta}_{t-1})^2}, \end{aligned} \quad (14)$$

assuming f_t is (almost) linear in Θ in the neighborhood of $\hat{\Theta}_{t-1}$, the first term in (14) vanishes and we obtain the following approximation

$$\nabla_{\Theta}^2 \ln(f_t(\hat{\Theta}_{t-1})) = -\mathbf{h}_t \mathbf{h}_t^T, \quad (15)$$

where $\mathbf{h}_t = \frac{\nabla_{\Theta} f_t(\hat{\Theta}_{t-1})}{f_t(\hat{\Theta}_{t-1})} = \nabla_{\Theta} \ln(f_t(\hat{\Theta}_{t-1}))$.

Let $\hat{\mathbf{R}}_t = \nabla_{\Theta}^2 S_t(\hat{\Theta}_t)$ and assume that the objective criterion S is smooth in the vicinity of $\hat{\Theta}_t$, and the adaptation step small enough so that

$$\hat{\mathbf{R}}_t = \nabla_{\Theta}^2 S_t(\hat{\Theta}_t) \simeq \nabla_{\Theta}^2 S_t(\hat{\Theta}_{t-1}). \quad (16)$$

This is a classic assumption for deriving recursive estimation methods for stochastic systems (see [13]). The two-step recursive scheme at time t is then

$$\hat{\mathbf{R}}_t = \alpha \hat{\mathbf{R}}_{t-1} + (1-\alpha) \mathbf{h}_t \mathbf{h}_t^T, \quad (17)$$

$$\hat{\Theta}_t = \hat{\Theta}_{t-1} + (1-\alpha) \hat{\mathbf{R}}_t^{-1} \mathbf{h}_t. \quad (18)$$

Equation (17) derives from (13) and (15). Equation (18) derives from (10), (12) and (16). The final algorithm is available in Algorithm 2.

Algorithm 2 Recursive MLE

Let $\Phi_0 = \mathbf{0}$, $\sigma_0^2 = 1$, $\nu_0 = 1$, $\mathbf{h}_0 = \mathbf{0}$, $R_0 = 0_{(p+2, p+2)}$.

repeat

1. *Update.* $\hat{\mathbf{R}}_i = \alpha \hat{\mathbf{R}}_{i-1} + (1-\alpha) \mathbf{h}_i \mathbf{h}_i^T$.

2. *Update.* $\hat{\Theta}_i = \hat{\Theta}_{i-1} + (1-\alpha) \hat{\mathbf{R}}_i^{-1} \mathbf{h}_i$ if $i > 100 + p$.

until t the forecasting time.

IV. VERY-SHORT-TERM WIND POWER FORECASTING APPLICATION

We apply the proposed models to a real dataset consisting of wind power generation from a large wind farm, Anholt in Denmark, from July 1, 2013 to August 31, 2014. Emphasis is placed on the maximum likelihood framework and its online learning derivation. For a comparison of the generalized logit-normal distribution to other distributions (e.g., Beta) for the purpose of wind power forecasting, see [5].

A. Data Description

Active power is available for 110 wind turbines at a temporal resolution of every 10 minute. The time series are scaled individually according to the nominal power of the wind turbines. The average generation over the wind farm is then computed depending on the number of wind turbines being available at each time step, in order to handle missing values. The resulting random variable is then $X_t \in [0, 1]$, the average active power generated in the wind farm at time t .

We are interested in forecasting X_{t+1} (point forecasting) and its distribution (probabilistic forecasting) knowing the realization of X_t ; the lead time is therefore 10-minute-ahead. We split our data into two datasets:

- a training/cross-validation dataset from July 1, 2013 to March 31, 2014, resulting in 39,450 observations,
- a test dataset from April 1 to August 31, 2014, resulting in 22,029 observations.

The training set is used to fit all models, the cross-validation set to select hyper-parameters if needed and the test set to compare the proposed methodology to the benchmarks. It is

worth noting the training set is long enough for the Algorithm 2 to be recursive yet on the training period, after a short warm-up of 100 iterations.

B. Point Forecasting

In order to evaluate and compare the performance of the proposed methods for point forecasting we use the Root Mean Square Error (RMSE). When a model requires hyper-parameters to be selected before estimating the parameters, we use the following procedure:

- The candidate models are fitted over a grid of hyper-parameters' values from July 1 to October 31, 2013;
- they are then retrained in a time-series cross-validation scheme, from November 1, 2013 to March 31, 2014, for which the size of the training window increases as we evolve through the validation set (consistent with a leave-one-out setup);
- the hyper-parameters leading to the smallest RMSE on the cross-validation set are selected;
- finally the final model is fitted over the whole training/cross-validation set and used for forecasting on the test set.

a) *Benchmarks:* We compare our methods to three benchmarks: the persistence, a normal auto-regressive (NAR) model and its recursive version. The persistence consists in taking $\hat{x}_{t+1} = x_t$. The normal AR model assumes $X_t|X_{t-1}, \dots, X_{t-p} \sim \mathcal{N}(\mu_t, \sigma^2)$ where $\mu_t = \sum_{k=1}^p \phi_k X_{t-k}$. In this Gaussian setup the forecasts are unbounded and happen to be greater than 1 or lower than 0. Thus we need to truncate *a posteriori* the out-of-range predictions so they lie in the interval $[0, 1]$. We test AR models up to lag $p = 5$ and observe that no significant improvement is provided beyond lag 2 for both batch and recursive approaches. We thus select $p = 2$. For the recursive AR model we also need to select the forgetting factor α , which exponentially weights data in the past. In a similar way, it is selected such as $\alpha = 0.995$.

b) *Forecasting using generalized logit-normal distributions:* Let $\delta > 0$ such as each value being lower than δ (resp. greater than $1 - \delta$) is set to δ (resp. $1 - \delta$) and consider those "corrected" observations as the realizations of $X \in (0, 1)$. In a symmetric way, forecasts being lower than δ (resp. greater than $1 - \delta$) will be set to 0 (resp. 1). δ is selected over cross-validation along with p . Algorithm 1 converges in 11 iterations towards the estimated values $\hat{\nu} = 1.39$, $\hat{\Phi} = (1.363, -0.370)^\top$ and $\hat{\sigma}^2 = 0.11$ for the selected combination of hyper-parameters $\delta = 0.005$ and $p = 2$. For Algorithm 2, we choose $\delta = 0.005$, $p = 2$ and $\alpha = 0.9994$ upon cross-validation. See in Fig. 1 the estimated parameters of the generalized logit-normal distributions over the test period.

c) *Results:* The point forecasting performance over the test set of the benchmarks and the (GLNAR) proposed algorithms are available in Table I. It is worth noting that the test set consists in 22,023 observations, which is a volume of data large enough to claim for significant results. The best point forecasts are obtained by the model using adaptive generalized

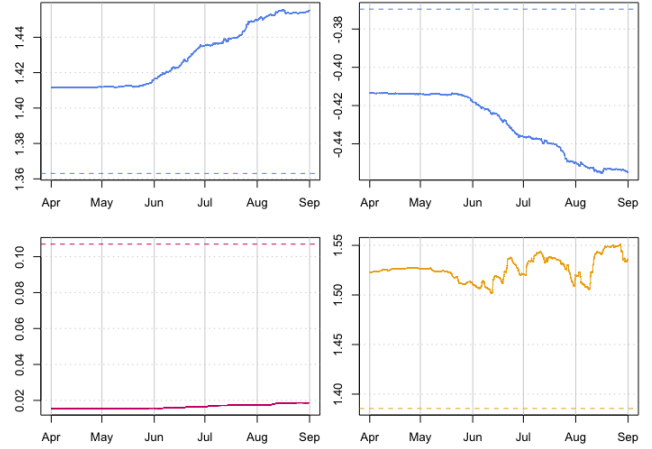


Fig. 1. Parameters of the generalized logit-normal distribution for $p = 2$ and $\alpha = 0.9994$: $\hat{\Phi}$ (top), $\hat{\sigma}^2$ (bottom left) and $\hat{\nu}$ (bottom right).

logit-normal distributions. One can note that the model which uses a constant generalized logit-normal distribution gets poorer performance than the recursive AR model. Therefore the assumption that seems to matter the most here is the time-varying parameters assumption. Moreover, the estimated value of the scale parameter is significantly larger in the batch setup than in the recursive one, while the shape parameter is significantly lower. It may confirm that the recursive setup is more appropriate to the characteristics of the time series and thus allows for a better discrimination between the scale and the shape parameters of the distribution.

TABLE I
10-MINUTE-AHEAD RMSE OVER THE TEST PERIOD, AND RESPECTIVE IMPROVEMENTS OVER PERSISTENCE

Model	RMSE	Imp. over persist.
persistence	3.27%	-
batch NAR	2.79%	14.68%
recursive NAR	2.72%	16.82%
batch GLNAR	2.74%	16.21%
recursive GLNAR	2.70%	17.43%

*Best forecast bolded.

C. Probabilistic Forecasting

Let F_t a predictive cumulative distribution function at time t . The Continuous Ranking Probabilistic Score (CRPS) is defined by

$$\text{CRPS} = \frac{1}{T} \sum_{t=1}^T \text{crps}(F_t, x_t) = \int_{-\infty}^{\infty} \text{BS}(y) dy, \quad (19)$$

where

$$\text{crps}(F_t, x_t) = \int_{-\infty}^{\infty} \{F_t(y) - \mathbf{1}(y \geq x_t)\}^2 dy, \quad (20)$$

and BS is the Brier score

$$\text{BS}(y) = \frac{1}{T} \sum_{t=1}^T \{F_t(y) - \mathbf{1}(x_t \leq y)\}^2. \quad (21)$$

See for example [14] and [15]. To evaluate the performance of the proposed models for probabilistic forecasting we use the

CRPS instead of the RMSE, following the scheme described at the beginning of section IV-B.

a) *Benchmarks:* We compare our method to four benchmarks: climatology, probabilistic persistence, and probabilistic versions of the batch and recursive AR models. Climatology consists in computing empirical quantiles on the training set. We test different grids and choose upon cross-validation to estimate the predictive cumulative distribution from the quantiles $\{0, 0.01, \dots, 0.99, 1\}$. On the test set the quantiles are updated whenever a new observation is recorded. Probabilistic persistence consists in dressing the point persistence prediction with the most recent observed values of the persistence error. We choose the number of observed values upon cross-validation to be 20. For probabilistic AR forecasts, the least squares estimator of the variance of the residuals is used in both batch and recursive modes, and we assume those residuals to follow a Gaussian distribution $\mathcal{N}(0, \hat{\sigma}^2)$. The forecast distribution of x_t is then a Gaussian distribution $\mathcal{N}(\hat{x}_t, \hat{\sigma}^2)$ where \hat{x}_t is the point forecast from the AR model. The hyper-parameters p and α for the recursive model are selected upon cross-validation with CRPS, which leads to $p = 2$ as for point forecasting, but to a different α which is now equal to 0.983 instead of 0.995.

b) *Forecasting using generalized logit-normal distributions:* The lag p selected upon cross-validation with CRPS remains equal to 2 in both batch and recursive algorithms, while δ and α change. For Algorithm 1, now $\delta = 0.006$ which leads to slightly different estimated parameters of the distribution: $\hat{\nu} = 1.37$ and $\hat{\Phi} = (1.358, -0.365)^\top$, while the variance $\hat{\sigma}^2 = 0.11$ remains the same. For Algorithm 2, now $\delta = 0.004$ and α decreases from 0.9994 for point forecasting to 0.9986 for probabilistic forecasting. See in Fig. 2 the estimated parameters of the generalized logit-normal distributions over the test period, which show higher time-variability because of the lower value of the forgetting factor.

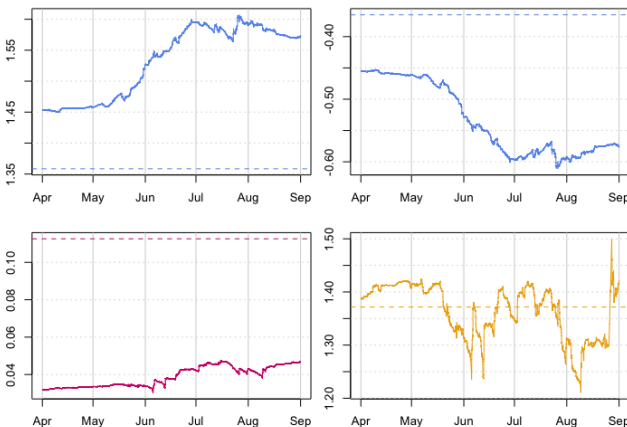


Fig. 2. Temporal evolution of the parameters of the generalized logit-normal distributions for $p = 2$ and $\alpha = 0.9994$: $\hat{\Phi}$ (top), $\hat{\sigma}^2$ (bottom left) and $\hat{\nu}$ (bottom right).

c) *Results:* The CRPS computed over the test set for all the benchmarks and the proposed models are available in Table

II. The climatology’s predictive cumulative distribution function F_{t+1} remains unchanged whatever the value of x_t , which explains the very poor global performance of this method. The performance of the predictive cumulative distributions assuming a Gaussian setup and that of the approach using a constant generalized logit-normal distribution are close as for point forecasting. However, for probabilistic forecasting, the approach using adaptive generalized logit-normal distributions outperforms the other methods. The Brier scores are plotted in Fig. 3. As expected the methods using the generalized logit transformation perform better close to the bounds of the interval $[0, 1]$.

TABLE II
10-MINUTE-AHEAD CRPS OVER THE TEST PERIOD, AND RESPECTIVE IMPROVEMENTS OVER CLIMATOLOGY AND PERSISTENCE

Model	CRPS	Imp. over clim.	Imp. over persist.
climatology	22.04%	-	-
prob. persistence	1.36%	93.85%	-
batch NAR	1.28%	94.17%	5.28%
recursive NAR	1.23%	94.34%	9.40%
batch GLNAR	1.21%	94.52%	10.90%
recursive GLNAR	1.06%	95.17%	21.57%

*Best forecast bolded.

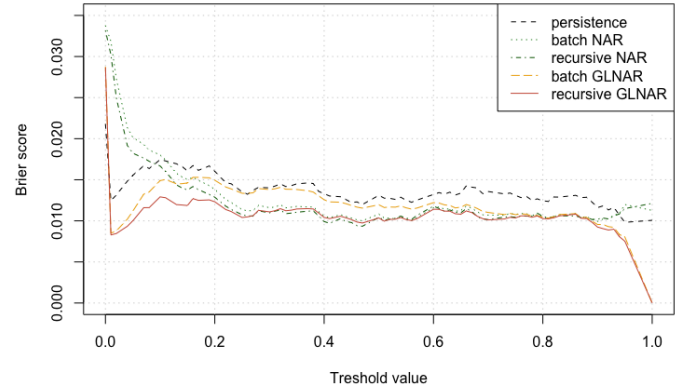


Fig. 3. Brier score computed over the test set for all methods but climatology, as a function of the chosen threshold.

The CRPS and the Brier score give indications about the sharpness of the distributions. In order to check the calibration we show the results of two tools: the reliability diagram in Fig. 4 and a marginal calibration plot which is the difference between the average predictive \bar{F} on the test set and the empirical cumulative distribution function in Fig. 5. For the reliability diagram, the closer to the diagonal, the better the calibration, the empirical probabilities getting closer to the nominal ones. See [16] and [15] for more details about those calibration tools. One can see that for both indicators the approach using adaptive generalized logit-normal distributions outperforms the other probabilistic forecasting methods. In Fig. 5 the climatology difference is not presented for being far bigger than zero.

Example probabilistic forecasts obtained from the adaptive generalized logit-normal approach over a 36 hour period of time are depicted in Fig. 6 by using prediction intervals with nominal coverage rates of 95 and 75%.

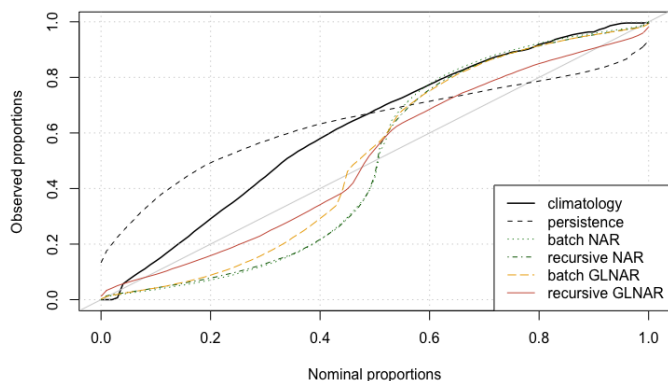


Fig. 4. Reliability diagram over the test set.

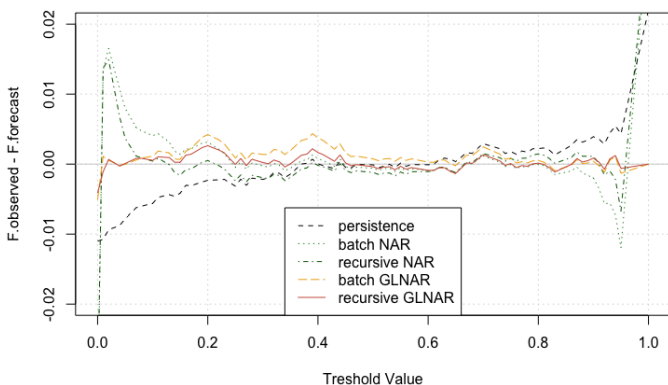


Fig. 5. Marginal calibration plot over the test set.

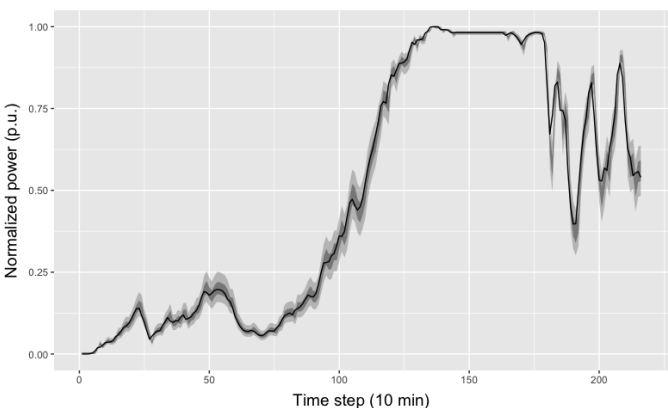


Fig. 6. Probabilistic forecasts from the recursive approach (Algorithm 2), based on prediction interval with nominal coverage rates of 95 and 75%, along with the power measurements (solid black line).

V. CONCLUSIONS

A generalized logit-normal distribution was considered for very short-term wind power forecasting, in order to adequately handle the double-bounded nature of the time series. All the parameters of the distribution were estimated from the data in a maximum likelihood framework, for both batch and online setups. The adaptive version of the distribution provides only a slight improvement in the accuracy of the point forecasts compared to approaches within a Gaussian framework, though

it substantially outperforms the other benchmarks when focusing on probabilistic forecasting (intervals and full predictive densities). This confirms that such a choice of distribution may be most appropriate. While it achieves better calibration and sharpness, there is still room for improvement. In particular, we have emphasized the importance of the double-bounded nature of the process, but in practice the upper bound may also change in time. Indeed, wind power generation is not always bounded by the nominal capacity of the wind farm, e.g. in case of curtailment. It should then be taken into account within the modelling and forecasting framework, by additionally adaptively estimating this upper bound from data.

Furthermore, the proposed framework could be applied for multi-step ahead forecasting, and makes it easy to assume other models for the conditional expectation of the transformed variable. In particular it is straightforward to add exogenous variables to the auto-regressive model, or to generalize it with a non-linear one. This may be a way to account for the individual productions of the wind turbines in order to improve the prediction of power generation for the whole wind farm. Finally, the δ hyper-parameter which handles the coarsened version of the distribution was selected upon cross-validation. It could instead enter a Bayesian or a likelihood inference as a parameter to be properly estimated.

REFERENCES

- [1] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang and H. Zareipour, "Energy forecasting: A review and outlook," *IEEE Open Access Journal of Power and Energy*, vol. 7, pp. 376-388, 2020.
- [2] V. Akhmatov, "Influence of wind direction on intense power fluctuations in large offshore wind farms in the North Sea," *Wind Energ.*, vol. 31(1), pp. 59-64, 2007.
- [3] C. Gilbert, J. Browell and D. McMillan, "Leveraging turbine-level data for improved probabilistic wind power forecasting," *IEEE Trans. Sust. Energ.*, vol. 11, no. 3, pp. 1152-1160, 2019.
- [4] A.P. Dawid, "Statistical theory: the prequential approach," *J. R. Statist. Soc. A*, vol. 157(2), pp. 278-292, 1984.
- [5] P. Pinson, "Very-short-term probabilistic forecasting of wind power with generalized logit-normal distributions," *J. R. Statist. Soc. C*, vol. 61(4), pp. 555-576, 2012.
- [6] F. Orabona. A Modern Introduction to Online Learning. Lecture notes, Boston University, 2020.
- [7] E. Lesaffre, D. Rizopoulos and R. Tsonaka, "The logistic transform for bounded outcome scores," *Biostatistics*, vol. 8(1), pp. 72-85, 2007.
- [8] D. Heijman and D. Rubin, "Ignorability and coarse data," *Ann. Stat.*, vol. 19(4), pp. 2244-2253, 1991.
- [9] D. Heijman, "Ignorability and coarse data: some biomedical examples," *Biometrics*, vol. 49(4), pp. 1099-1109, 1993.
- [10] P. Young, *Recursive estimation and time-series analysis: An introduction*. Springer-Verlag, Berlin, Heidelberg, 1984.
- [11] H. Madsen, *Time Series Analysis*. Chapman & Hall, Boca Raton, 2007.
- [12] P. Pinson and H. Madsen, "Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models," *J. Forecast.*, vol. 31, pp. 281-313, 2012.
- [13] L. Ljung and T. Söderström, *Theory and Practice of Recursive Estimation*, 1983.
- [14] G.W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78(1), pp. 1-3, 1950.
- [15] T. Gneiting, F. Balabdaoui and A.E. Raftery, "Probabilistic forecasts, calibration and sharpness," *J. R. Statist. Soc. B*, vol. 69(2), pp. 243-268, 2007.
- [16] P. Pinson, H.Aa. Nielsen, J.K. Møller and H. Madsen, "Non-parametric probabilistic forecasts of wind power: Required properties and evaluation," *Wind Energ.*, vol. 10(6), pp. 497-516, 2007.