Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Evaluation of Wind Power Forecasts – An up-to-date view

Jakob W. Messner*[1] | Pierre Pinson[1] | Jethro Browell[2] | Mathias B. Bjerregård[1] | Irene Schicker[3]

[1]Technical University of Denmark (DTU), Lyngby, Denmark
[2]University of Strathclyde, Glasgow, UK
[3]Austrian Weather Service (ZAMG), Viena, Austria

Correspondence
*Jakob Messner. Email: jakob.messner@posteo.net

Abstract

Wind power forecast evaluation is of key importance for forecast provider selection, forecast quality control and model development. While forecasts are most often evaluated based on squared or absolute errors, these error measures do not always adequately reflect the loss functions and true expectations of the forecast user, neither do they provide enough information for the desired evaluation task. Over the last decade, research in forecast verification has intensified and a number of verification frameworks and diagnostic tools have been proposed. However, the corresponding literature is generally very technical and most often dedicated to forecast model developers. This can make forecast users struggle to select the most appropriate verification tools for their application while not fully appraising subtleties related to their application and interpretation. This paper revisits the most common verification tools from a forecast user perspective and discusses their suitability for different application examples as well as evaluation setup design and significance of evaluation results.

KEYWORDS:
Wind power, forecast evaluation, evaluation metrics, testing forecast performance

1 | INTRODUCTION

Wind power has become an important power source in many power systems. In Europe it already covers approx. 12% of the total electricity demand [1]. However, variability and limited predictability of its production challenges power systems and markets, making forecasts required for optimal operation (e.g. load balancing and maintenance) and trading. A lot of research has been carried out in the development of wind power forecasting models and a variety of models have been proposed for different applications and types of forecasts. These include deterministic point predictions, probabilistic forecasts of various forms, multivariate predictions or predictions for specific events such as ramps or gusts [2]. See e.g., [3] for a general state-of-the-art report on wind power forecasting or [4] for a recent coverage of challenges related to wind power forecasting (and extension to other renewable energy sources).

One of the current challenges, which is rarely covered and discussed, is forecast verification, maybe since many believe that verification frameworks are well-established and forecast users are content with their use. Forecast evaluation is crucial for model development, selection of the best forecast provider, or for quality control. Some of its main goals include estimation of future error statistics, comparison of the forecast accuracy of different forecasts, or finding flaws in a certain forecast model. Unfortunately, it is not the case that current knowledge in forecast verification and existing verification frameworks can give us the whole information about objective quality of forecasts and their value to forecast users. The original view on forecast quality and value (inspired by meteorological applications) was laid out in the 1980s by [5,6]. More recently, this aspect was discussed by [7] or [8] for the specific case of wind power forecasting.

Evaluation metrics are tools to summarize the characteristics of forecast errors but unfortunately there is no universal metric that can examine all forecast qualities. The best forecast in one metric can perform poorly with respect to another metric. Therefore, it is essential to select an evaluation criterion that well reflects the cost function of the forecast user. E.g., if the cost of an error is directly proportional to the error, the mean

absolute error is most appropriate. Selecting an inappropriate evaluation criterion can lead to wrong conclusions such as the selection of a forecast provider that is not the best for the intended application [9].

Just like the forecasts themselves, also forecast evaluation exhibits some degree of uncertainty and evaluation results do not always have to reflect future expectations. E.g., there might be performance differences between different years or if forecasts are evaluated only for the summer season the results do not have to be representative for the winter season. Therefore, it is important to design the evaluation setup appropriately and to be able to quantify and correctly interpret these uncertainties of the results.

In contrast to forecast model development, forecast evaluation has not received as much attention in wind power forecasting literature. Notable exceptions are [10], which proposes a standard protocol for forecast evaluation,[7], which examines the evaluation of ensemble forecasts [8], which discusses the relationship between forecast quality and value, or [11], which discusses evaluation approaches for wind power scenario forecasts. Nevertheless, performance evaluation has been an important tool in model development and nearly all publications ought to rely on some form of verification framework to benchmark their own approach. Beyond wind power only, one may find a number of reference works on forecast evaluation in the general forecasting literature. Examples include [12], [13], [14], [15], [16], [17], [18], or [19].

Traditionally, discussions of forecast evaluation techniques have mainly been considered by forecast model developers and therefore proposed evaluation approaches are often presented in a technical way and focused on specific problems. In this study we want to review the evaluation from the perspective of a forecast user, revisit some of the most important evaluation metrics for wind power forecasting and discuss their usability for different applications. Furthermore, the evaluation setup and the interpretation of evaluation results is discussed. Thus, this document intents to become a reference for forecast users when setting up a forecast evaluation procedure. It does not suggest specific procedures or metrics but rather critically examines the advantages and disadvantages of different approaches so that it enables forecast users to tailor solutions for their own specific application. As such it complements part 3 of the International Energy Agency (IEA) Recommended Practice on Forecast Solution Selection [9]

The remainder of this document is structured as follows. First, Section 2 demonstrates on a simple example forecast the importance of selecting a metric that fits to the forecast product, the difference between quality and value, and pitfalls when interpreting results from an inappropriate evaluation setup. Section 3 summarizes some of the most important evaluation metrics for different kinds of forecasts, including point forecasts, probabilistic forecasts of binary, multi-categorical, or continuous variables, and multivariate scenarios. Section 4 discusses approaches to set-up evaluation tasks and interpret their results. Finally, a conclusion can be found in Section 5.

2 | PRAGMATIC CONTEXT

In this section, we want to point out typical pitfalls of evaluation procedures on simple forecast example data. Two fairly simple examples are considered to illustrate the importance of loss functions, forecast verification framework, and the link between quality and value of forecasts. For this purpose we employ the openly available data set of the GEFCom 2014 wind power forecasting competition [20]. This data set consists of 2 years of hourly wind power measurements for 10 Australian wind farms (the exact locations have not been disclosed) and corresponding 25–48 hours numerical wind forecasts at 10 and 100 meter above ground obtained from the European Centre for Medium-range Weather Forecasts (ECMWF) high resolution model. For the current study we only used the 100 meter wind forecasts at one single wind park. The full data set is available as appendix to [20]. In order to facilitate reproduction and future work, the subset of these data and all the code used to generate the results of this paper can be downloaded at [21]

2.1 | A forecast benchmarking example

We first transform the 100 meter wind speed predictions into power generation forecasts using a simple local linear regression model (see e.g. [22]). If we denote the wind power measurement at time $t, t = 1, \dots, N$ as y_t and the corresponding day ahead wind speed predictions as $\hat{u}_t$ this model can be described by

$$y_t = \alpha_{i,0} + \alpha_{i,1}(\hat{u}_t - u_i) + \epsilon_t \tag{1}$$

where $u_i, i = 1, \dots, P$ are a number of fitting points, ϵ_t the forecast error, and $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}), i = 1, \dots, P$ are regression coefficients that are different for all P fitting points. Thus, separate regression equations are fitted for each fitting point, which are combined depending on the distance between the respective fitting points and the actual value of $\hat{u}_t$. A common choice for fitting points can e.g., be one point for each m/s. These

coefficients are estimated so as to minimize the weighted sum of a loss function $\rho()$ over the training data set

$$\hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\arg\min} \sum_{t=1}^{N} w_t \rho(y_t - \alpha_{i,0} - \alpha_{i,1}(\hat{u}_t - u_i)) \tag{2}$$

where the loss function ρ commonly is the squared (quadratic) loss but can be any loss function that ideally should reflect the intended application of the forecast. Clearly, therefore, if the end-user's preferred evaluation measure is suitable to be used directly as $\rho()$ then it should be, but this is not always possible.

The weights w_t are defined by a Kernel function,

$$w_t = K\left(\frac{|\hat{u}_t - u_i|}{h}\right) \tag{3}$$

where h is the bandwidth parameter controlling the smoothness of the fit and K can, e.g., be the tricube function

$$K(v) = \begin{cases} (1 - \mathsf{v}^3)^3 & \mathsf{v} \in [0, 1] \\ 0 & \mathsf{v} > 1 \end{cases} \tag{4}$$

We fit three different models of this kind with three different loss functions:

- *quadratic* loss $\rho(\epsilon) = \epsilon^2$,
- *absolute* loss $\rho(\epsilon) = |\epsilon|$
- 0.3 *quantile* loss $\rho(\epsilon) = \epsilon\,(0.3 - \mathbb{1}(\epsilon < 0))$

Figure 1 shows these loss functions. Compared to the absolute and quantile loss, the quadratic loss strongly penalizes larger errors and compared to the other loss functions the quantile loss is not symmetric and penalizes negative errors more than positive ones. Figure 2 shows example forecasts for a specific date. The absolute and quadratic error models provide rather similar forecasts with the absolute loss model predicting slightly lower power generation on average. The forecasts of the quantile loss model are even lower, which leads to less negative errors that are weighted higher than positive errors.

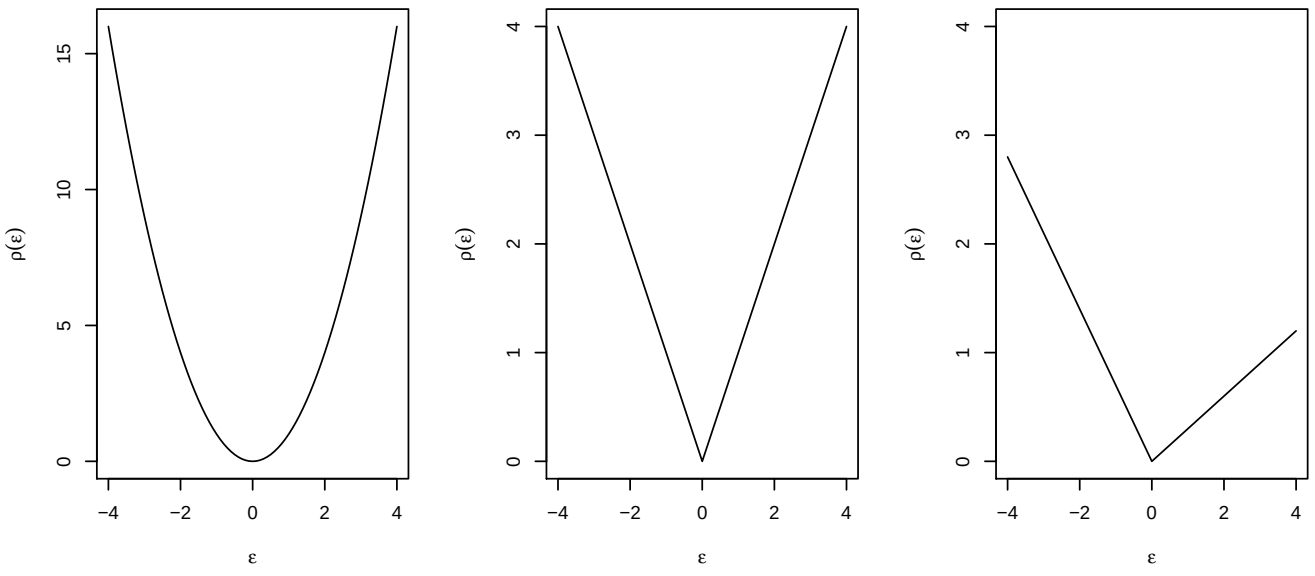

FIGURE 1 Quadratic (left), absolute (center), and quantile (right) loss functions $\rho(\epsilon)$. Note the different scale on the y-axes.

These three models are fit on the first 10000 entries of the GEFCom2014 data set and are used to generate forecasts for the remaining 6789 entries. These forecasts are evaluated using 3 different evaluation metrics which are the mean over the test data set of the 3 loss functions listed

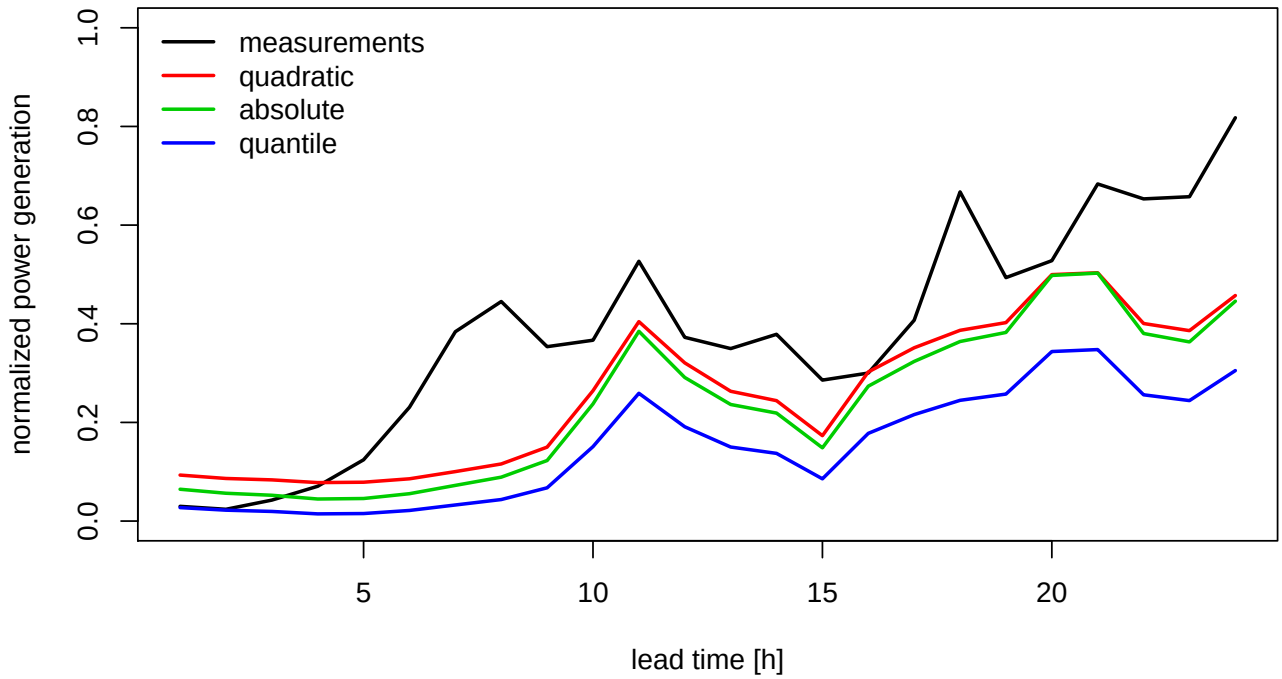

FIGURE 2 Example 24 hour forecast from different forecast models for 2013-11-30.

	MSE	MAE	QS
quadratic	**3.46**	14.27	7.27
absolute	3.46	**13.54**	6.61
quantile	4.63	15.29	**5.67**

TABLE 1 Different evaluation measures for the three local linear models with quadratic, absolute and quantile loss function. All scores are in their normalized version, hence expressed in percentage of nominal capacity. The best model for each score is highlighted in bold.

above: the mean squared error (MSE), the mean absolute error (MAE) and the quantile score (QS) – to be introduced and thoroughly discussed in Section 3. Table 1 summarizes these evaluation results.

Thinking about how models were fitted and based on the intuitive match between loss functions for model fitting and verification, it is not surprising that each model performs best in the metric that was used in the model fitting. Nevertheless, these results show three important aspects of forecast evaluation:

1. the ranking of forecasts clearly depends on the chosen metric and based on a single metric it is not possible to define a forecast that is best for all possible applications

2. to achieve the best possible results it is important for forecast providers to know the actual loss function

3. it is important for the forecast users to know their loss function of forecast errors. First, the forecast providers can only then optimize their models to this loss function and when evaluating different providers, a wrong metric could lead to choosing not the most suitable one for a specific application.

In the above example, the forecast performance is measured on a rather big data set (i.e., a test dataset with 6789 forecast-observation pairs). However, often not as many data are available and performance has to be measured on smaller data sets. Table 2 shows the same performance

measures as Table 1 but only using the first 200 time steps of the test data set. Since forecasts are updated hourly, 200 time steps translates to approximately 8 days. However, if forecasts were updated daily or twice daily, this would translate to period of 6 months and 3 months, respectively.

	MSE	MAE	QS
quadratic	**3.14**	**12.93**	5.87
absolute	3.68	13.97	5.87
quantile	4.55	15.97	**5.35**

TABLE 2 Same as Table 1 but only derived from the first 200 time steps of the test data set

It can be seen that this clearly changes the ranking of the different models so that the quadratic loss function ranks best in terms of MSE and MAE, even though we know from construction that the absolute loss function models should be preferred by the latter metric. The problem here is that a data set length of 200 is not sufficient to draw final conclusions based on score differences, especially for a highly temporally correlated data set such as the one used here, which is typical of wind power data. Evaluation results based on a finite data set are always subject to some degree of uncertainty and the best ranked forecast does not necessarily have to be the truly best one. Depending on the actual setup, e.g., in a benchmarking exercise to hire a forecaster, it should be remembered that even periods of several months may still yield uncertainty in terms of who the best forecaster truly is.

2.2 | A maintenance planning example

Let us now assume these forecasts are used for turbine maintenance planning for which an hour with zero production or wind speeds below cut in speed (e.g., 3 m/s) is required. Additional to the models above, we want to use a forecast directly based on the 100 meter wind speed numerical prediction, which forecasts conditions suitable for maintainance when the numerical prediction falls below 3 m/s.

Table 3 shows the contingency tables (to be introduced and thoroughly discussed in Section 3.2.1) for this simple model and the absolute loss function model from the previous subsection. Since the absolute loss model predicts zero generation very rarely (only four times in the whole test data set) it is not of much value for this application and only predicts one event, suitable for maintenance, correctly. Thus, even though the local linear model is clearly more advanced and predicts the correct outcome (correct positive and negative) more often (6192+1=6193 versus 5890+206=6096), it is not of much value for this specific application and in most practical applications easily outperformed by the direct numerical model output. This example shows that the value of a forecast clearly depends on the intended application and that not always the forecast with the best quality is the one that has the highest value.

	absolute loss model		direct model output	
	FALSE	TRUE	FALSE	TRUE
FALSE	6192	3	5890	305
TRUE	593	1	388	206

TABLE 3 Contingency tables for forecasting zero wind power with a local linear regression model with minimized absolute loss (left) and with the direct numerical model output (right). Rows are for observations (TRUE or FALSE) and columns for forecasts (TRUE or FALSE)

3 | EVALUATION METRICS

Forecast evaluation is often used to test if forecasts are reasonable and to analyse their performance in various situations, which can help to improve the forecast models. This is often referred to as forecast verification and is usually done by employing different metrics or graphical representations. Furthermore, forecast evaluation is necessary to compare different forecasts to each other, for example to select the best forecast provider for a specific application. In principle, the same metrics as for verification can be used, however, usually single valued metrics or scoring rules are preferred to graphical devices. Since this paper mainly focuses on forecast comparison, we will mainly regard single valued metrics but also cover a few useful important graphical verification devices.

In the following we list a number of scoring rules. This is clearly only a selection of the most widely used metrics and is not a comprehensive list. We also omit to describe theory about desired properties of scoring rules, such as the importance of being proper and refer to e.g.,[23] for more details.

This section is divided into subsections for different forms of forecasts. The first subsection focuses on deterministic point forecasts (single valued forecats), the second subsection treats probability forecasts for binary events and the last two subsections present metrics for distributional probabilistic forecasts in the uni- and multivariate case respectively.

3.1 | Single valued wind power forecasts

This subsection compares a set of single valued forecasts $\hat{y}_t, t = 1, \ldots, N$ to corresponding observations $y_t, t = 1, \ldots, N$. Clearly, a good forecast $\hat{y}_t$ should be as close to y_t as possible. Here, various approaches are listed to measure the distance between forecasts and observations, i.e. the quality of a forecast.

3.1.1 | Bias

The bias (i.e., mean or systematic error) is defined as

$$Bias = \frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t) \tag{5}$$

and measures the average difference between the forecast and observations, which can easily be seen when reformulating (5) as

$$Bias = \frac{1}{N}\sum_{t=1}^{N}\hat{y}_t - \frac{1}{N}\sum_{t=1}^{N}y_t, \tag{6}$$

As an illustration, Figure 3 shows example observations, two different forecasts and their averages. Forecast 1 has very little correlation to the observations (correlation coefficient $<$0.02) but has the same average as the observations and thus a very small bias of 0.01. In contrast, Forecast 2 predicts the evolution of the observations perfectly accurately but is always 0.2 too low, which results in a bias of -0.2.

Thus, the bias only measures the ability of a forecast to predict the right average level but does not give any information about the forecasts ability to predict specific events (commonly referred to as resolution or discrimination ability). Since a known bias can easily be corrected by adding a constant, a low bias should be more seen as a necessary condition than a forecast quality measure.

3.1.2 | (Root) mean squared error - (R)MSE

The mean squared error is defined as

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2 \tag{7}$$

and measures the mean squared distance between forecasts and observations. The root mean squared error

$$RMSE = \sqrt{MSE} \tag{8}$$

contains the same information but has the same physical unit as the observations and forecasts (e.g. kW for wind power).

Since errors contribute to the MSE quadratically, larger errors are penalized strongly (see also Figure 1). Therefore, this error measure is particularly useful for applications where large errors are related to high costs while small errors lead to relatively low costs. Despite the popularity of this error metric, there actually exist almost no examples in wind power applications that follow such a cost function. One example could be the cost of reserve energy available to power system operators, which typically becomes more expensive the more is required. In this case, the costs incurred as a result of wind power forecast errors will not be in proportion to the size of the errors; however, it will likely not be symmetric or quadratic either, and will change over time. In general it is far more common for costs to be in proportion to the size of a forecast error (perhaps asymmetrically, as in quantile loss), or discrete based on thresholds, than in proportion to the squared error.

3.1.3 | Mean absolute error - MAE

The mean absolute error is defined as

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|\hat{y}_t - y_t| \tag{9}$$

and measures the mean absolute distance between forecasts and observations. In contrast to the RMSE errors are penalized proportionally (see also Figure 1). Hence, it is well suited for applications where the cost of errors is directly related to its magnitude. For example, the economic consequence of a forecast error may be the product of forecast error and some price-per-unit. This is more common in contractual arrangements between forecast vendors and their customers (or regulators) than in energy markets.

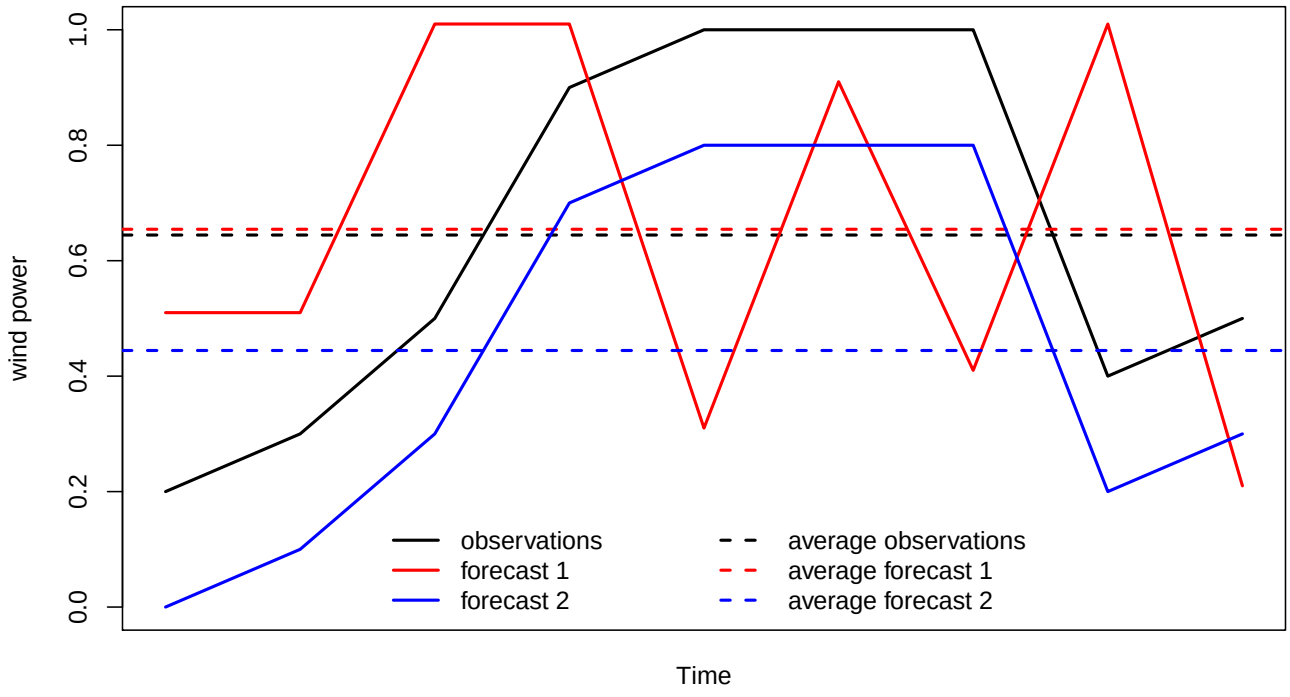

FIGURE 3 Example time series of observations and two different forecasts (solid lines) and their averages (dashed lines)

3.1.4 | Quantile score - QS

The quantile score also measures the absolute error but weights it differently whether the error is positive or negative. It is defined as

$$QS(p) = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{y}_t - y_t\right)\left(\mathbb{1}\left(y_t \leq \hat{y}_t\right) - p\right) \tag{10}$$

where $0 < p < 1$ is the weighting of positive errors (i.e. $y_t > \hat{y}_t$) while negative errors are weighted with $1 - p$. The right panel in Figure 1 shows the contribution of errors exemplary for $p = 0.3$.

This metric is called quantile score because it can be shown that it is minimized by the p-quantile of the predictive distribution. The quantile score should be used in situations where it is known that the costs of positive and negative errors differ such as in dual-price electricity markets, where the economic cost of over-contracting is usually less than for under-contracting. In this situation the expected cost is minimized by deliberately over-contracting in order to reduce exposure to large costs at the expense of increasing exposure to small costs, as in [24].

3.1.5 | Economic value and decision making

As already noted in the description of the different metrics above, there are certain situations or applications that suit certain metrics very well. Before selecting a metric to base a forecasting model on, it is therefore important to know the expected costs related to inevitable forecast errors. Clearly, in many situations the cost function is more complex and cannot be directly described by any of the above metrics but if it is known, it can directly serve as a metric and so directly reflect the economic value of a forecast. Where the economic cost takes the form of a cost-loss ratio, the optimal decision is a quantile, and Murphy diagrams may be used to evaluate and visualise the range of all economic scenarios [25].

However, in many situations the cost function is not clear, is effected by many other factors, can vary over time or a forecast may be used for different applications with different cost functions. In such a situation, decisions should be based on a combination of different metrics such as Bias, MAE, RMSE, quantile scores for different values of p, and potential other single valued metrics.

TABLE 4 Contingency table

Forcast \ Observation	yes	no
yes	hits	false alarms
no	misses	correct negatives

3.2 | Forecasts of binary events

Often forecast users are interested in the occurrence of specific events and want accurate forecasts of them. Examples could be ramps or cut-outs. Modern forecast systems usually provide probabilistic forecasts for such events, e.g., the probability of cutting-out between 10am and 11am tomorrow. The forecast users then have to decide for themselves at what probability threshold they want to take action. This threshold should be related to the costs of an action and the loss in case no action has been taken and it can easily be shown that usually the expected revenue is maximized when action is taken whenever the predicted probability exceeds the cost loss ratio.

There are two main approaches to evaluate such forecasts. First, metrics such as the Brier score or the area under the receiver operating characteristic curve (ROC, see below) can be used to directly measure the accuracy of the probabilistic forecast. Alternatively, the forecasts can be evaluated based on the actions that have been taken, thus directly reflecting the economic value of the forecast.

In the following let $\hat{z}_t, t = 1, \ldots, N$ be a probability forecast ($0 \leq \hat{z}_i \leq 1$) for the observation z_t, which has the value 1 when the considered event occurs and 0 if not.

3.2.1 | Contingency table and derived metrics

Let us consider the cost loss function is well known and thus a threshold th can be defined to take action. Then the forecast probabilities $\hat{z}_t$ can be transformed into binary forecasts

$$\hat{z}_t^* = \begin{cases} 1 & \text{if } \hat{z}_t > th \\ 0 & \text{else} \end{cases} \tag{11}$$

A contingency table summarizes the quality of the forecast by displaying the number of

- hits – forecast event to occur, and did occur
- misses – forecast event not to occur, but did occur
- false alarms – forecast event to occur, but did not occur
- correct negatives – forecast event not to occur, and did not occur

Table 4 illustrates the construction of a contingency table and Table 3 shows 2 examples.

Contingency tables can give a nice overview over the forecast performance but are difficult to use for forecast comparison. Therefore several different single valued metrics can be derived from it. Examples are the hit rate (HR)

$$HR = \frac{\text{hits}}{\text{hits} + \text{misses}} \tag{12}$$

or the false alarm rate (FAR)

$$FAR = \frac{\text{false alarms}}{\text{false alarms} + \text{correct negatives}} \tag{13}$$

Similarly, scores such as accuracy, bias score, threat score, Peirce's skill score, or Heidke skill score can be derived from the entries in the contingency table. For more details see e.g., [12].

If the cost of action (C) and the loss in case of no action (L) are known, they can be used to directly derive the costs related to a forecast.

$$\frac{C(\text{hits} + \text{false alarms}) + L(\text{misses})}{N} \tag{14}$$

where $N = \text{hits} + \text{false alarms} + \text{misses} + \text{correct negatives}$.

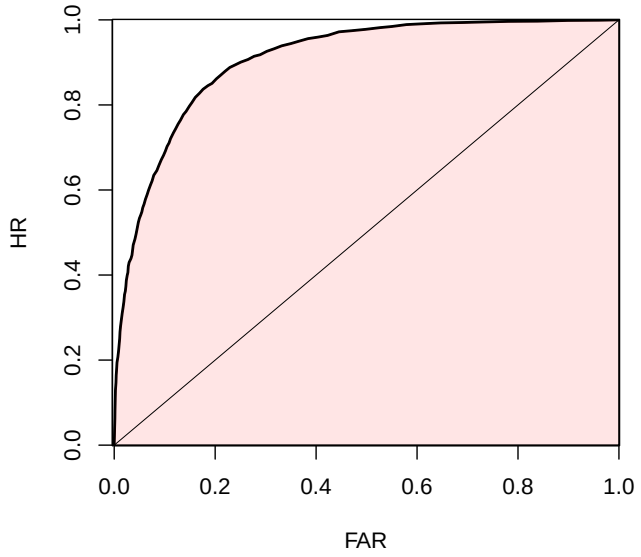

FIGURE 4 Example ROC plot. The thick black line shows the ROC curve while the diagonal thin line shows the ROC of a forecast with no skill. The area under the curve is shown in red shading.

3.2.2 | Receiver operating characteristic (ROC)

If the cost loss function is not known or not constant over time it can be better to directly evaluate the probabilistic forecast. One common approach to do so is the receiver operating characteristic (ROC). The ROC is a plot of the hit rate (HR; Equation 12) versus the false alarm rate (FAR; Equation 13) and by connecting a number of points for different probability thresholds th a curve is drawn that starts at (0,0) and ends at (1,1). Figure 4 shows an example ROC curve.

A well performing forecast should have a high hit rate and a low false alarm rate so that the curve should lie as much in the upper left corner of the plot as possible. Randomly forecasting probabilities between 0 and 1 (forecast with no skill) would lead to a diagonal ROC curve.

To compare forecast models to each other, it is common to derive the area under the ROC curve which summarizes in a single value how far the ROC curve is away from the no-skill diagonal. However, it should be noted that when evaluating probabilistic forecasts, ROC curves and AUC do not consider reliability and therefore should be accompanied by reliability diagrams [26].

3.2.3 | Brier score - BS

The Brier score is given by

$$BS = \frac{1}{N}\sum_{t=1}^{N}\left(\hat{z}_t - z_t\right)^2, \tag{15}$$

which is equivalent to the mean squared error in Equation 7 but for probability forecasts $\hat{z}_t$ and binary observations z_t instead of continuous variables.

The Brier score can take values between 0 and 1 with smaller values indicating better forecasts. [27] showed that the Brier score can be decomposed into reliability (REL), resolution (RES), and uncertainty (UNC)

$$BS = REL - RES + UNC \tag{16}$$

Reliability denotes the property of a forecast to be in line with the conditional relative frequencies of the observations, i.e., in the long run an event should occur in 40% of the cases the probability forecast is 40%. Resolution is the property of a forecast to discriminate between situations, i.e., a forecast that has almost the same value every day has a bad (low) resolution. Uncertainty is the base uncertainty in the outcome of the considered event and is independent from the forecast. This decomposition can be very useful to examine the forecast performance and find where forecast models have problems.

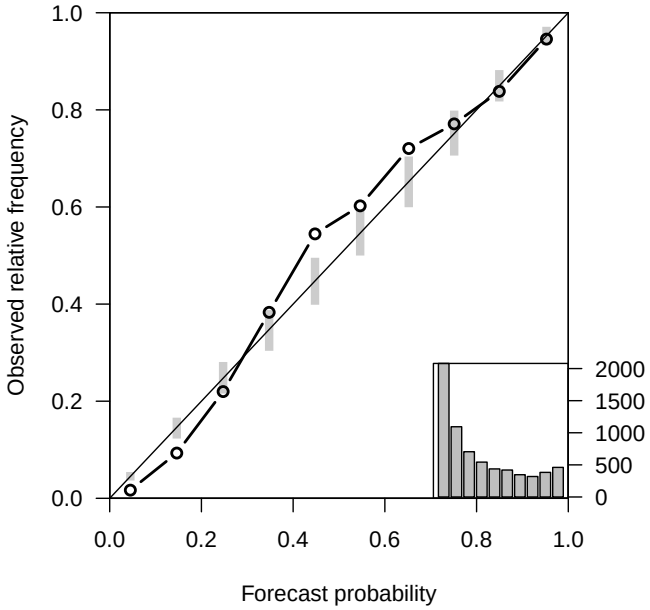

FIGURE 5 Example reliability diagram. The gray bars show consistency bars as in [29] and the refinement distribution is plotted in the lower right corner.

3.2.4 | Reliability diagram

As written above, reliability is the property of a probabilistic forecast to predict probabilities that fit to the relative frequencies in the data. A probabilistic forecast that is not reliable can lead to wrong decisions when the predicted probabilities are interpreted directly. As such, it should be seen as a necessary condition of a good probabilistic forecast, similar to the bias for deterministic forecasts.

As shown in Section 3.2.3 the reliability can be assessed as a part of the Brier score. Alternatively, reliability diagrams are related graphical devices that can be used for assessing the reliability of binary probabilistic forecasts. In reliability diagrams, the observed frequencies are plotted against the predicted probabilities. Therefore the interval $(0, 1)$ is divided into several subintervals and relative frequencies conditional on forecasts falling in each of these intervals are plotted against the interval center or median. For reliable forecasts, observed and predicted frequencies should be similar so that their reliability diagram should be close to a diagonal line.

Traditionally, reliability diagrams also contain a refinement distribution subplot which show histograms of the predicted probabilities e.g., [28]. These show the confidence of a forecaster, which is high if probabilities close to 0 and 1 occur frequently and is low when the predicted probabilities are always similar. The refinement distribution can also be used to estimate the expected sampling variation of the reliability diagram. If there are only few data in one subinterval this variation is expected to be higher than for well populated intervals. [29] proposed another approach to estimate this sampling variability, based on consistency bars that show the potetnial deviation of actually perfectly reliable forecasts due to limited sampling. This concept of consistency bars was then generalized by [30], arguing that it is not only limited sampling, but also correlation, that affect estimates of reliability. This ought to be accounted for when estimating and visualizing consistency bars.

Figure 5 shows an example reliability diagram with consistency bars and refinement distribution. In this example, the reliability diagram is close to the diagonal but falls outside the bootstrap confidence intervals in some of the bins.

3.3 | Probabilistic forecasts of continuous variables

Probabilistic forecasts have been shown to be beneficial for various decision making processes in wind power applications e.g., [31,24,32,33] and therefore are becoming more and more popular. Thus, nowadays many forecast providers offer probabilistic wind power forecasts in the form of quantiles (perhaps in the form of prediction intervals, which are just specific quantiles), ensembles (set of possible scenarios), or full parametric distributions. The advantage of probabilistic forecasts is that they provide information about the forecast uncertainty and allow to take this into account for decision making. Sometimes, probabilistic forecasts are used optimally by taking specific quantiles as point forecasts, which maximize the revenue. If the

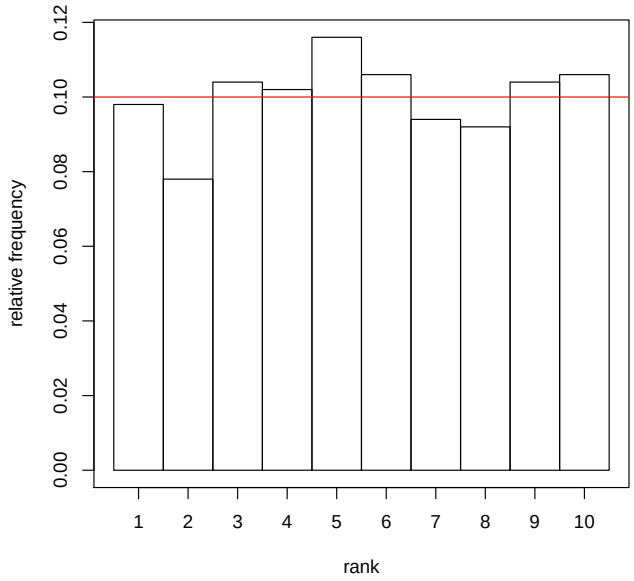

FIGURE 6 Example verification rank histogram for a 10 member ensemble based on a data set length of 500. Perfect reliability is shown as red line.

required quantiles are not provided directly, they can be easily derived from full continuous probabilistic distributions, by interpreting an ensemble as a set of quantiles, or by interpolating between quantiles. In such a case, a straightforward way to evaluate the accuracy of the forecast is to use the quantile score (see Section 3.1).

Unfortunately, decision making processes based on probabilistic forecasts are often much more complex and sometimes made manually and based on various inputs, not only the wind power forecast. In such a case, the full forecast distribution should be evaluated. There has been a number of metrics proposed for probabilistic forecast evaluation and below we list the most important ones.

3.3.1 | Verification Rank histogram and Probability integral transform (PIT) histogram

The Verification Rank histogram and PIT histogram are closely related graphical devices that are commonly used to examine the reliability of probabilistic forecasts. Reliability again denotes the property of a probabilistic forecast to be in line with the relative frequencies of observations, i.e., in the long run 20% of the data should fall below the 20% quantile.

The verification rank histogram is used to examine the reliability of ensemble forecasts by counting the number of observations falling in the different intervals that are specified by the ensemble forecasts. This is equivalent to a histogram of the ranks of the observations within the ensemble forecasts thus the name verification rank histogram. If the ensemble forecast is reliable, the verification rank histogram should be flat. Figure 6 shows an example verification rank histogram. Note that here the deviations from perfect reliability are most probably an effect of sampling variations and that the forecast here can be regarded as reliable. With a longer data set the histogram would become more and more flat.

A similar plot can also be drawn for forecasts that are given as a set of quantiles. Though, depending on which quantiles are given, the histogram does not have to be flat but should follow the nominal probabilities of the different intervals.

PIT histograms are the continuous counterpart of verification rank histograms and show the distribution the probability integral transform, which is

$$PIT_t = \hat{F}_t(y_t) \tag{17}$$

where $\hat{F}_t(y_t)$ is the predicted cumulative distribution function. If the forecasts are well calibrated and reliable, the PIT histogram should be flat as well. Note that when discrete cumulative distribution functions are derived from ensemble forecasts, the resulting PIT histogram is almost identical to the verification rank histogram only with a different scale on the x-axis.

Reliability is a crucial property of probabilistic forecasts. Unreliable forecasts can lead to not ideal decisions and thus to financial loss. Looking at rank or PIT histograms should therefore be one of the first steps in evaluating probabilistic forecasts and if they deviate significantly from uniformity the forecasts should be calibrated or only be used with care.

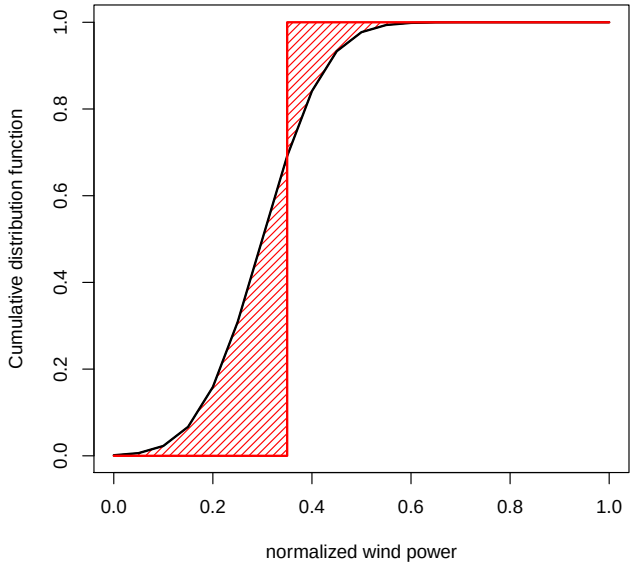

FIGURE 7 Schematic plot for the derivation of the continuous ranked probability score. The black curve shows the predicted cumulative distribution function and the red curve indicates the step function $\mathbb{1}(x \leq y_t)$. The difference between these two lines is shown as red shaded area.

Reliability should also be seen more like a property a forecast has or does not have and the reliability of different forecasts should in general not be ranked.

3.3.2 | Continuous ranked probability score

The continuous ranked probability score is one of the most common single value scores to evaluate the accuracy of probabilistic forecasts of continuous variables. It evaluates the quality of the predicted cumulative distribution function and is defined as

$$CRPS = \frac{1}{N}\sum_{t=1}^{N}\int_{-\infty}^{\infty}\left[\hat{F}_t(x) - \mathbb{1}(x \leq y_t)\right]^2 dx \tag{18}$$

where $\mathbb{1}(x \leq y_t)$ is the indicator function that is 1 if $x \leq y_t$ and 0 otherwise. Figure 7 shows a schematic plot for the derivation of the CRPS. The CRPS for a specific forecast occasion is the integral of the squared distances between the cumulative distribution function and the step function defined by the observed value. Therefore it is not directly the shaded area in Figure 7 but related to it.

Note that the integrand in Equation 18 can be interpreted as a Brier score (Equation 15) so that the CRPS can be seen as the integral over the Brier score. There are also other equivalent definitions of the CRPS, e.g.[34],

$$CRPS = \int_0^1 \hat{F}_t^{-1}(\tau) - y_t)\left(\mathbb{1}\left(y \leq \hat{F}^{-1}(\tau)\right) - \tau\right)\mathrm{d}\tau, \tag{19}$$

which shows that the CRPS is also closely related to the quantile score (Equation 10), which is equal to the integrand in Equation 19. Another definition proposed by[23] is

$$CRPS = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{1}{2}E|\hat{Y}_t - \hat{Y}_t'| - E|\hat{Y}_t - y_t|\right] \tag{20}$$

where $\mathrm{E}|\cdot|$ denotes the expected value and $\hat{\mathrm{Y}}_t$ and $\hat{\mathrm{Y}}_t'$ are independent copies of a random variable with distribution function $\hat{\mathrm{F}}_t$. From this definition, a formula for forcasts given as ensembles or quantiles can be easily derived as

$$\mathrm{CRPS} = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{1}{M}\sum_{m=1}^{M}|\hat{y}_t^{(m)} - y_t| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}|\hat{y}_t^{(m)} - \hat{y}_t^{(l)}|\right] \tag{21}$$

where $\hat{\mathrm{y}}_t^{(\mathrm{m})}$, $\mathrm{m} = 1, \ldots, \mathrm{M}$ M are ensemble members or predicted quantiles.

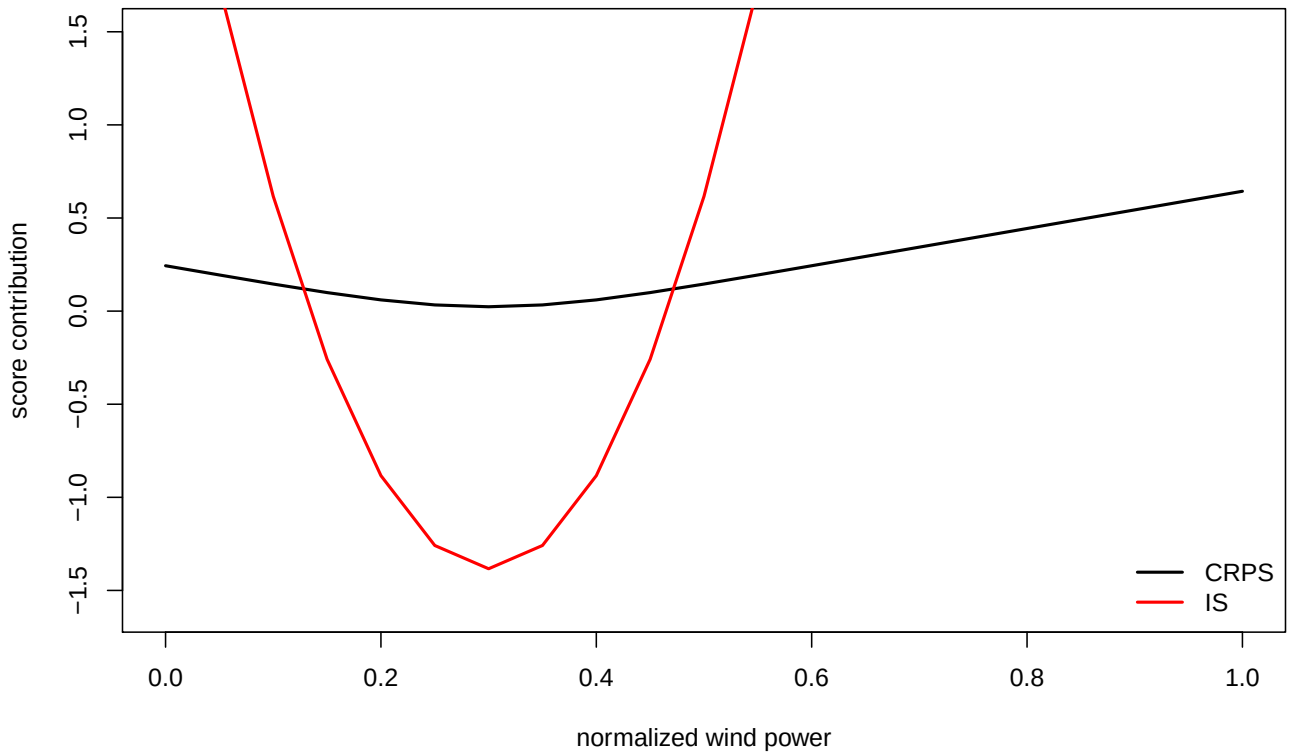

FIGURE 8 Score contributions of observations to the same example forecast distribution as shown in Figure 7. The black and red curves shows contributions to the CRPS and ignorance score, respectively.

[35] showed that the CRPS, similar to the Brier score, can be decomposed into reliability, resolution and uncertainty.

Figure 8 shows the CRPS contributions of different observations. It can be seen that, except close to the distribution mean, deviations from the distribution mean contribute almost linear to the CRPS. This is comparable to the mean absolute error (see Figure 1) and in fact, for a deterministic forecast (i.e., the predictive cumulative distribution function is a step function as well), the CRPS and the mean absolute error are equivalent.

3.3.3 | Ignorance (logarithmic) score

The ignorance score, also called logarithmic score is defined as

$$IS = \frac{1}{N}\sum_{t=1}^{N} \log(\hat{f}_t(y_t)) \tag{22}$$

where $\hat{f}_t(y_t)$ is the predicted probability density function evaluated at the value of the observation y_t. Since a probability density function can not easily be derived from quantiles or ensembles, the ignorance score is only applicable for full continuous distribution forecasts.

As it can be seen in Figure 8 the ignorance score penalizes deviations from the distribution center much more heavily than the CRPS. In the case of a normal predictive distribution the ignorance score is, up to a factor, equivalent to the squared loss. Similar to the choice between mean absolute error and (root) mean squared error, the ignorance score should be preferred if large forecast errors are related to very high costs.

3.4 | Multivariate probabilistic forecasts

Multivariate forecasts are usually provided as a set of scenarios that are consistent in time and/or space and consider the spatio-temporal correlations. E.g., these could be a set of possible realizations for the 24 hours of the next day. Multivariate forecasts are important in short-term wind

power forecasting and therefore have become popular in the wind power literature. For example, [36] showed that when considering forecasts for a set of wind power production sites, properly accounting for spatio-temporal inter-dependence between neighbouring sites results in a reduction in prediction errors compared to simply issuing independent forecasts of individual sites.

Similar to other forecast formats, multivariate forecasts are, depending on the application, used for decision making. Multivariate forecasts could e.g., be used to estimate the probability that a threshold is exceeded within a certain time period or that the accumulated wind power in a region exceeds a certain threshold. In such situations, these derived forecasts can be evaluated directly with evaluation metrics from the previous subsections [11]. However, it is also possible to evaluate multivariate scenarios directly using e.g., one of the metrics presented below.

In the following we present some of the most popular multivariate scoring rules.However, multivariate forecast evaluation is still a very active research field and it is possible that other, perhaps better evaluation metrics will become more popular in the near future. We denote multivariate observations as vectors $\mathbf{y}_t$, which can contain a set of forecasts for different locations, different lead times, or both. Multivariate forecasts are usually provided as set of M scenarios in K dimensions $\hat{\mathbf{y}}_t^{(m)} = (\hat{y}_{t,1}^{(m)}, \hat{y}_{t,2}^{(m)}, \dots, \hat{y}_{t,K}^{(m)})^\top, m = 1, \dots, M$.

3.4.1 | Multivariate ignorance or Dawid-Sebastiani score

Similar to the univariate case, multivariate forecasts could be evaluated based on the algorithm of their multivariate density function $\hat{f}_t(\mathbf{y}_t)$

$$IS = \frac{1}{N}\sum_{t=1}^{N} \log(\hat{f}_t(\mathbf{y}_t)) \tag{23}$$

However, usually multivariate forecasts are not provided in parametric form but rather as a set of possible multivariate scenarios. In such a case, the closely related multivariate Dawid-Sebastiani score [37] can be used

$$DS = \frac{1}{N}\sum_{t=1}^{N} \left[\log(\det \hat{\mathbf{\Sigma}}_t) + (\mathbf{y}_t - \overline{\hat{\mathbf{y}}}_t)^\top \hat{\mathbf{\Sigma}}_t^{-1}(\mathbf{y}_t - \overline{\hat{\mathbf{y}}}_t)\right] \tag{24}$$

where $\overline{\hat{\mathbf{y}}}_t$ is the mean and $\hat{\mathbf{\Sigma}}$ the covariance matrix of the forecasts $\hat{\mathbf{y}}_t^{(m)}$ and $\det \hat{\mathbf{\Sigma}}_t$ is the determinant of $\hat{\mathbf{\Sigma}}_t$. The Dawid-Sebastiani score is equivalent to the ignorance score for a predicted multivariate normal distribution with mean $\overline{\hat{\mathbf{y}}}_t$ and covariance $\hat{\mathbf{\Sigma}}$. Thus, it is the ignorance score assuming the multivariate scenarios are samples from a multivariate normal distribution and estimating the distribution parameters with mean and covariance matrix. For wind power this assumption might not always hold.

Similar as the univariate ignorance score, its multivariate version penalizes unlikely observations, i.e. misidentified tails, very hard, which may or may not be desired depending on the problem of consideration.

3.4.2 | Conditional likelihood and censored likelihood score

In order to maintain the nice properties of the multivariate ignorance score while damping the penalty associated with unlikely observations (cf. above), [38] proposed two scores that accomplishes exactly that. Let A be a subset of the sample space of the forecast, such that observations that fall outside A, i.e. in A^c are denoted "unlikely observations". The simplest of the two scores is the conditional likelihood score,

$$\text{CDLS} = \frac{1}{N}\sum_{t=1}^{N} \mathbb{I}(\mathbf{y}_t \in A) \log\left(\frac{f_t(\mathbf{y}_t)}{\int_A f_t(\mathbf{u})\mathrm{d}\mathbf{u}}\right) \tag{25}$$

which is the ignorance score only evaluated for observations within A. Hence, this can be used to exclude unlikely observations from the forecast evaluation. The other score in question is the censored likelihood score,

$$\text{CSLS} = \frac{1}{N}\sum_{t=1}^{N} \mathbb{I}(\mathbf{y}_t \in A) \log f_t(\mathbf{y}_t) + \mathbb{I}(\mathbf{y_t} \in A^c) \log\left(\int_{A^c} f_t(\mathbf{u})\mathrm{d}\mathbf{u}\right) \tag{26}$$

Under this score, observations that fall outside A are still evaluated. The penalty for each unlikely observation is then based on the total probability mass on A^c rather than on the probaility of the unlikely observation itself (as is the case for the ignorance score). Hence, unlikely observations are penalized in a more robust manner than in the ignorance score.

3.4.3 | Multivariate continuous ranked probability or energy score

As for the ignorance score, the CRPS can also be extended to cover multivariate scenarios, which has been proposed under the name *energy score* by [23]

$$\text{ES} = \frac{1}{N}\sum_{t=1}^{N} \left[\int_{-\infty}^{\infty} (\hat{F}_t(\mathbf{x}) - \mathbb{1}(\mathbf{x} \geq \mathbf{y}_t))^2 \mathrm{d}\mathbf{x}\right] \tag{27}$$

where $\hat{\mathrm{F}}_t()$ is the predicted multivariate cumulative distribution function.

If the forecasts are given as a set of scenarios, the formula

$$\mathrm{ES} = \frac{1}{N}\sum_{t=1}^{N}\left[\frac{1}{M}\sum_{m=1}^{M}||\hat{\mathbf{y}}_t^{(m)} - \mathbf{y}_t|| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{l=1}^{M}||\hat{\mathbf{y}}_t^{(m)} - \hat{\mathbf{y}}_t^{(l)}||\right] \tag{28}$$

can be used where $||\mathbf{d}||$ is the Euclidean norm.

Similar to the univariate case, the CRPS does not penalize unlikely observations as strongly as the ignorance score.

3.4.4 | Variogram score

[11] showed that the energy score is not very sensitive to misspecification in the multivariate correlation structure and puts most weight on the quality of the marginal distributions. In applications where the correlation structure is important this can be undesirable. As an alternative score that puts more weight on the correlation structure, [39] proposed the variogram score

$$\mathrm{VS}_p = \frac{1}{N}\sum_{t=1}^{N}\left[\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}(|y_{t,i} - y_{t,j}|^p - E[|\hat{Y}_{t,i} - \hat{Y}_{t,j}|^p])^2\right] \tag{29}$$

where $\mathrm{y_{t,i}}, \mathrm{i} = 1, \ldots, \mathrm{K}$ are the components of the multivariate observations $\mathbf{y_t} = (\mathrm{y_{t,1}}, \mathrm{y_{t,2}}, \ldots, \mathrm{y_{t,K}})^\top$, $\hat{\mathrm{Y}}_{\mathrm{t,i}}, \mathrm{i} = 1, \ldots, \mathrm{K}$ are components of a random vector $\hat{\mathbf{Y}}_\mathbf{t}$ that are distributed according to a forecast distribution $\hat{\mathrm{F}}_\mathrm{t}(\mathbf{y_t})$, and $\mathrm{w_{i,j}}$ are nonnegative weights that can be assigned if desired. p is the order of the variogram score and affects how closely the distribution of $|\hat{\mathrm{Y}}_{\mathrm{t,i}} - \hat{\mathrm{Y}}_{\mathrm{t,j}}|^\mathrm{p}$ attains symmetry. [39] thus found $\mathrm{p} = 0.5$ to be optimal for model separation. If the forecasts are provided as scenarios, $\mathrm{E}[|\hat{\mathrm{Y}}_{\mathrm{t,i}} - \hat{\mathrm{Y}}_{\mathrm{t,j}}|^\mathrm{p}]$ can be replaced by $\frac{1}{\mathrm{M}}\sum_{\mathrm{m}=1}^{\mathrm{M}}|\hat{\mathrm{y}}_{\mathrm{t,i}}^{(\mathrm{m})} - \hat{\mathrm{y}}_{\mathrm{t,j}}^{(\mathrm{m})}|^\mathrm{p}$.

The scores ability to distinguish between models in terms of their correlation structure becomes more apparent with increasing dimensions, and the computation time is quadratic, making it relatively fast and applicable for high-dimension scenarios compared to the available alternatives such as the ignorance score. The main downside of the score is that it does not cover calibration at all, i.e. different models with different expectations but identical correlation structures will be scored equally. Therefore, use of the variogram score may be supplemented by univariate CRPS or ignorance scores to make sure calibration and sharpness of the marginal distribtuions are addressed as well.

4 | EVALUATION SETUP

As pointed out in Section 2 an appropriate setup is required to get meaningful evaluation results and since these results are subject to uncertainty it is important to know how to interpret them. This section first regards different aspects and approaches for setting up an evaluation task such as data preparation or data set size. Subsequently, different approaches are presented to estimate the significance of evaluation results, which, for many decisions, can be as important information as the results themselves. An even more practical oriented discussion on this topic can also be found in [9].

4.1 | Data preparation/missing data/corrupt data

Evaluation results are highly dependent on the data set on which the evaluation is performed. Therefore it is important to use an appropriate data set for evaluating wind power forecasts. First, it is crucial that the selected data set is representative for the application the forecasts are supposed to be used for. E.g., the data set should cover all seasons, times of day, locations, etc. that they are planned to be used for or at least to a subset of these that is known to be representative. Second, the data set should be long enough for the results to be meaningful. Evaluation results are always subject to uncertainty, which increases with smaller data sets. In the case of small data sets it can therefore be difficult to see significant differences between competing forecast models. For limited data sets, cross validation approaches (see Section 4.2) can help to obtain more meaningful results.

Another aspect to consider is the aggregation of lead times. If forecast users are interested in the overall performance of a forecast model they may choose to evaluate all lead times at once. If forecasts for different lead times are used for different applications (e.g., trading in intraday and day ahead markets), forecast errors at different lead times are related to different costs, or users have the possibility to use different forecast models for different lead times it makes sense to evaluate forecast performance on lead times or subsets of lead times separately.

When comparing different forecasts to each other it is crucial to use exactly the same data sets. Results of different locations, seasons, lead times etc. are in general not comparable. If a certain forecast is not available for a specific time, this time has to be disregarded for all the other forecasts as well. Else, if e.g., forecasts are missing for days that are particularly difficult to predict, they would in total perform much better than forecasts that are expected to have high errors at these days.

Another important decision to be made is whether curtailment data should be kept or removed from the data before evaluation. Again this decision should be made based on the intended application. If the forecast user is interested in the available power and not in the real power production,

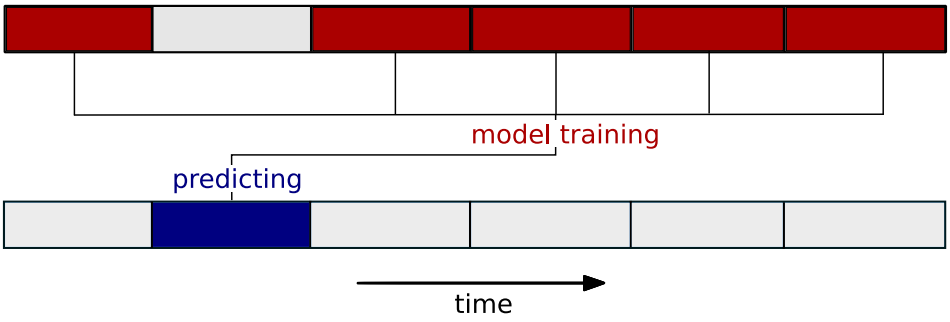

FIGURE 9 Schematic illustration of 6-fold cross validation with temporally contiguous blocks. The top box illustrates the full data set where the red blocks show the part of the data that is used for training the forecast model. The bottom row illustrates the forecasts where the blue block is the one that is predicted by the model that has been trained on the red blocks above. By repeatedly leaving out different blocks, independent predictions for the full time series can be derived.

data with curtailment should be removed from the evaluation data set since errors when not predicting these cases are not meaningful for the forecast performance. If periods of curtailment are retained, it may be instructive to separate errors that resulted from unforeseen curtailment from those that resulted from others as average scores will conflate these effects.

4.2 | Cross validation

In all evaluation tasks, it is of crucial importance to have independent training and test data sets, meaning that the data on which forecast models are evaluated should never be used in the model development. This is also reflecting a real forecasting task where the forecasted data is not available for developing the model. Violating this important condition can lead to very wrong conclusions. Often, only a limited data set is available on which the forecast models have to be trained and evaluated. Simply separating these data into two sets can on the one hand limit the training data such that the forecast models loose accuracy and on the other hand limit the test data such that the evaluation results are less meaningful and might be influences by few unusual events.

Cross validation is a frequently used tool to assure independence but still make efficient use of the available data. There are different cross validation approaches but all of them use the basic idea of repeatedly training the models on a major part of the data and evaluate them on the remaining part. By repeating this for different subsets, the evaluation results become less variable even if the actual test data part is small.

Cross-validation types for a data set of length N :

- k-fold cross validation is probably the most frequently used approach for wind power forecast evaluation. The original data set is split into k equally sized subsets. Then forecasts for each of the subsets are derived from models trained on all data leaving out the subsets that are to be forecasted. After repeating this for all partitions, independent predictions for the full data set are available for evaluation.
- leave-one-out-cross validation: derive independent forecasts for all N data points by fitting N models on the data set, leaving out the data point that is to be predicted. Similar to k-fold cross validation this results in independent forecasts for the full data set but requires N times fitting the models.
- leave-p-out cross validation: similar to leave-one-out but derive forecasts for a set of p events by leaving out those in fitting the model. Usually this is repeated on all ways to cut the full data set, so that the model has to be fitted $\binom{N}{p}$ times where $\binom{N}{p}$ is the binomial coefficient. Different to k-fold cross validation and leave-one-out cross validation each data point is predicted multiple times.
- random subsampling: randomly assign data to a train and a test data set and repeat this several times.

Since in wind power forecasting evaluation, model fitting often is rather computationally expensive, k-fold cross validation is usually preferred to leave one-out or leave-p-out cross validation. Another advantage of k-fold cross validation is, that temporal blocks can be selected as partitions thus avoiding problems with temporal correlations (see below). For the same reason also random subsampling is usually avoided. Figure 9 shows the cross validation procedure schematically.

4.2.1 | Temporal correlation

Cross validation assumes that the statistical properties of the dataset stay constant with time so that using future data for training is equivalent to using past data. However, wind power data is usually temporally correlated, which often implies that data that is temporally close to each other often behave similar. Thus, if the data just before and after a specific data point is used for training, the forecasts are not entirely independent and can lead to wrong conclusions. Therefore, leave-one-out cross validation can be problematic and in k-fold cross validation the partitions should be selected in temporally connected blocks and not randomly sampled.

When a sufficiently large dataset is available, it may be preferable to simulate operational forecasting and model re-training on a rolling basis. For example, training a model on the first 12 months of data and predicting the 13th month, and then re-training the model using the first 13 months and predicting the 14th, and so on. This structure is inherent to some forecasting methodologies that are explicitly *adaptive* [22].

4.3 | Comparing forecast performance

Most of the time forecast evaluation is used to compare different forecast models to each other, e.g., to select the best model for the intended application. Clearly one could simply compare one or several of the performance measures presented in Section 3 and rank the forecast models accordingly. However, evaluation results are always subject to uncertainty and should therefore interpreted carefully. Figure 10 shows mean squared error results for the example forecasts in Section 2 from different subsets of the test data set. Even though, the forecast model with quantile loss optimization seems to perform slightly worse there are subsets where it shows better mean squared errors than some of the mean squared errors of the other models. The right panel in Figure 10 shows that the sampling variation becomes a bit lower for larger subsamples.

The remainder of this section presents different approaches to estimate the evaluation result uncertainty and the significance of performance differences.

4.3.1 | Skill scores

Before regarding the uncertainty of evaluation results we want to introduce skill scores. In the boxplots in Figure 10 a number of mean squared errors are shown for different subsets of the data. The mean squared errors of the quadratic and absolute models are not always lower than that of the quantile model but in fact we cannot say from the figure whether the quantile model is expected to be always worse or not. Possibly, there are subsets where all models perform equally bad and the variation we see is not caused by variation in the ranking of the model but by the variation of the subset data.

To investigate model differences, one should therefore regard error differences or skill scores. A skill score of a metric M is defined as

$$\frac{M_{ref} - M}{M_{ref} - M_{perf}} \tag{30}$$

with M_{ref} the score of a reference method and M_{perf} the score of a perfect forecast. Skill scores show the score improvement of a forecast model compared to a reference model where positive values indicate an improvement. Often, basic forecast models such as the long term (climatological) mean or persistence are use as reference but when e.g., a new forecast model should be tested against the one currently in use it makes sense to use the current model as reference.

For many metrics the perfect score is 0, so that often the form

$$1 - \frac{M}{M_{ref}} \tag{31}$$

is used. Note also that for some metrics such as the logarithmic score, the perfect score is not finite so that no skill score can be derived.

Figure 11 left shows the same results as the right panel in Figure 10 but as skill scores with the quadratic loss model as reference. Clearly the quadratic model has skill score 0 itself but compared to Figure 10 it can be clearly seen that the quantile loss model performs worse than the quadratic in all evaluation subsets, which is in the median even around 30% worse.

4.3.2 | Bootstrapping

Analyses such as shown in Figure 10 or 11 left can be very useful to estimate the significance of an evaluation result. However, most of the time evaluation data sets are limited and as shown by Figure 10 the sampling variation increases with the usage of smaller evaluation subsets. Bootstrapping [40] is a popular resampling approach that reveals similar information but without sacrificing the accuracy of the results. Therefore, for an evaluation data set of length N random samples with replacement (each data point can be sampled several times) of size N are drawn repeatedly and the average scores are derived on these random samples. After repeating this k times, k different average score values are available that reflect the sampling variation of the average score. Similarly, bootstrap averages of score differences can give a good indication of significance of these

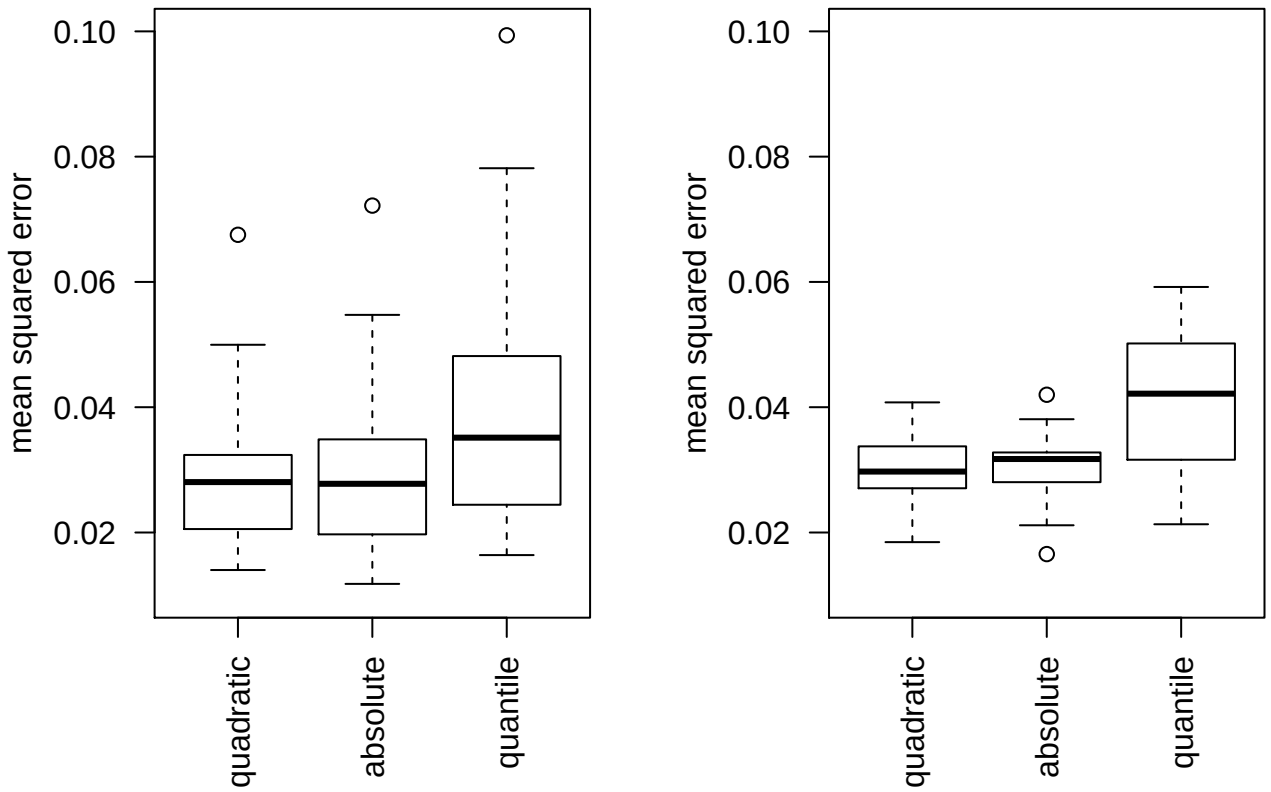

FIGURE 10 Sampling variation of mean squared error for different forecasting models. The boxplots in the left panel show the mean squared errors of 20 different samples of length 200 and the boxplots in the right panel show mean squared errors for 10 different samples of length 400.

differences. However, it is important to note, that the bootstrapping approach assumes serial independence of forecast errors so that for possible positive serial correlation in wind power data the bootstrapping approach can be too confident.

The right panel of Figure 11 shows the mean squared error skill score variation from bootstrap sampling. Compared to the 10 subsets in the left panel the average results are very similar but because the average skill scores are derived on larger samples their sampling distribution is much lower. Even the difference between the quadratic and absolute loss models becomes apparent.

Additionally to the larger samples the skill scores are derived on, the differences in the variations can also partly be caused by serial correlation in the forecast errors. A more quantitative approach to estimate the significance of results that also takes into account these correlations is presented in the next subsection.

4.3.3 | Diebold-Mariano test

[41] proposed a statistical test to test for differences in performance of two forecasts. In the following let $S(\hat{y}_t, y_t)$ be a scoring rule such as the squared error or the absolute error and $d_t = S(\hat{y}_t^1, y_t) - S(\hat{y}_t^2, y_t)$ be the score difference between two different forecasts $\hat{y}_t^1$ and $\hat{y}_t^2$. Furthermore, $\bar{d} = \frac{1}{N}\sum_{t=1}^{N} d_t$ is the mean loss difference and $\gamma_k = \frac{1}{N}\sum_{t=k+1}^{N}(d_t - \bar{d})(d_{t-k} - \bar{d})$ its autocovariance at lag k. Then the Diebold-Mariano test statistic is

$$DM = \frac{\bar{d}}{\sqrt{\frac{\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k}{N}}} \tag{32}$$

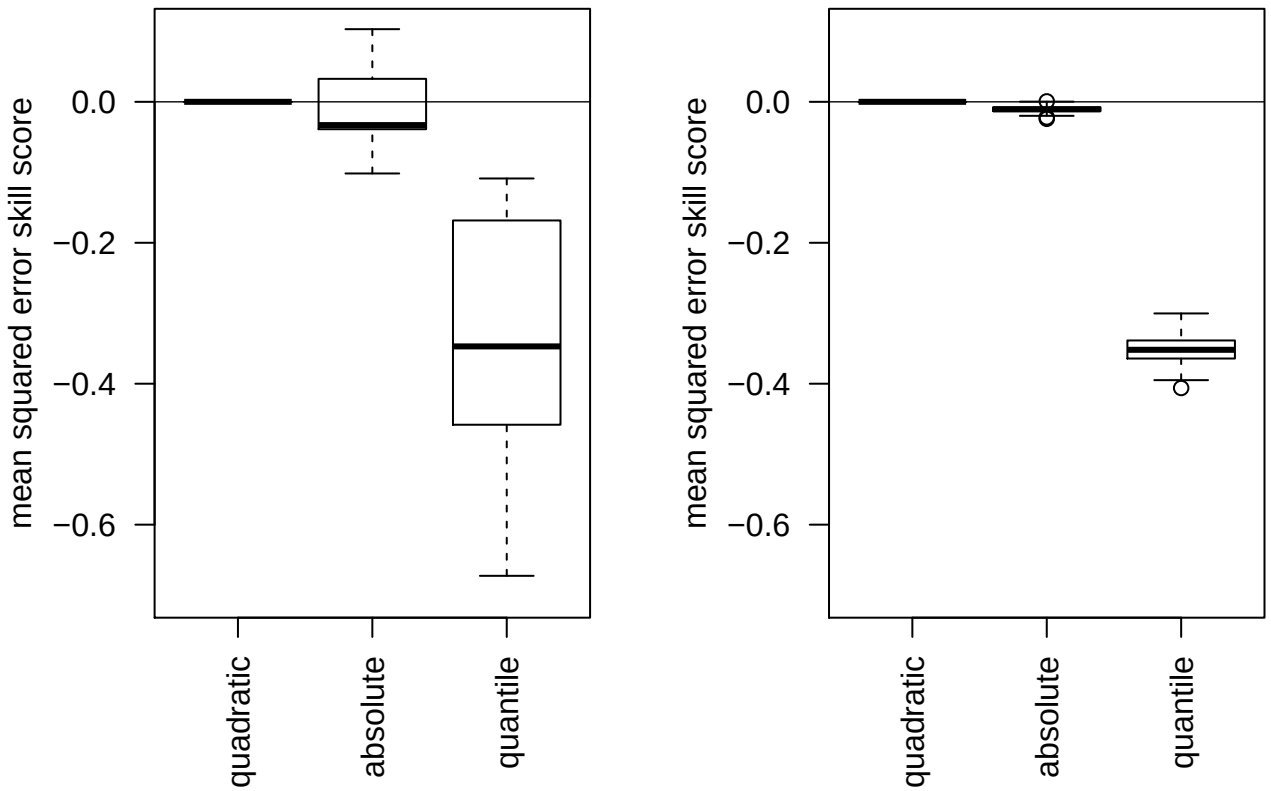

FIGURE 11 Mean squared error skill score with the quadratic loss function model as reference. The boxplots show the distribution of the average skill scores for 10 different subsamples (left) and 250 bootstrap samples (right).

where h is the number of considered lags and should be selected large enough to not miss any autocorrelations in the forecast errors. Under the null hypothesis of equal performance the Diebold Mariano statistic asymptotically follows a standard normal distribution

$$DM = \mathcal{N}(0, 1) \tag{33}$$

so that, for a two sided test, the null hypothesis can be rejected when

$$|DM| > z_{\alpha/2} \tag{34}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution to the desired α level of the test.

Note that in the case of serial independence, the Diebold-Mariano statistic in Equation 32 becomes

$$DM = \frac{\overline{d}}{\sqrt{\frac{\gamma_0}{N}}} \tag{35}$$

and thus the Diebold-Mariano test becomes asymptotically equivalent to a paired sample Student-t test. Care must be taken in the case of over-lapping forecasts and it is suggested in[41] that different lead-times should be tested separately, though an appropriate modification to the denominator of 32 is also a possibility.

Table 5 shows results for the Diebold-Mariano test for the squared errors of the absolute and quantile loss models compared to the squared loss models from the example of section 2. These results show clearly, that the difference between the absolute and squared loss models is not significant on this data set whereas the difference between quantile and squared loss model clearly is (i.e., p-value clearly below a typical $\alpha/2 = 0.025$ confidence level).

	absolute	quantile
quadratic	0.98	9.7e-09

TABLE 5 p-values from two sided Diebold-Mariano tests for equality of the squared error for the absolute and quantile loss models compared to the quadratic loss model. See Section 2 for details on the test setup. Lower values signify more significant differences.

4.3.4 | Variation in Forecast Performance

The performance of forecasting methodologies will vary according to the predictability of specific situations; however, different methodologies may exploit the various sources of predictability to different degrees. This is particularly relevant to the set-up of underlying numerical weather prediction models which can differ in spatial and temporal resolution, observations available for assimilation and the specific assimilation scheme, parameterisation of physical process that cannot be resolved directly, and other factors [42]. When comparing forecasts that draw on different sources of predictability, their relative performance will also vary with the prevalence of those sources. Similarly, statistical post-processing methods risk being biased by conditions that are abundant in training data but not explained by specific features.

Examples relevant to wind power include boundary layer mixing and low level jets. Differences in model performance during these events may manifest in diurnal and/or seasonal variations in forecast performance. Therefore, evaluating and comparing forecast performance based on time-of-day, time-of-year, or weather-type (if such information is available) may reveal valuable information relevant to forecast model selection, model mixing/blending and routes to forecast improvement.

4.4 | Practical Demonstration of Forecast Evaluation

In this section we outline the practical steps required to evaluate wind power forecasts. The data and code that accompany this paper [21] serve as a practical demonstration of this procedure. This approach is aligned with the IEA Wind Recommended Practice for Selecting Renewable Power Forecasting Solutions [9], which focuses on the selection and evaluation of services from commercial forecast providers.

Consider the following practical example: a modification to an operational forecasting system has been proposed but has to be evaluated to determine if it is an improvement or not. We will use the example from Section 2.1 as the 'Benchmark', and compare it to a regression spline model [43] which incorporates wind direction. The 'Proposed' model is given by

$$y_t = f_1(\hat{u}_t) + f_2(\hat{d}_t) + \epsilon_t \tag{36}$$

where $f_1(\hat{u}_t)$ is a smooth function of predicted wind speed $\hat{u}_t$ estimated by a cubic regression spline , and $f_2(\hat{d}_t)$ is a smooth function of wind direction $\hat{d}_t$ estimated by a cyclic cubic regression spline (i.e. with value and first 2 derivatives of $f_2(\theta)$ matching at θ and $\theta + 2\pi$ for all θ).

In the following subsections, the practical steps taken to evaluate the new system and compare it's performance are illustrated in an offline environment using historic data, and an online environment where only 'live' data is available and accumulates over time.

4.4.1 | Offline Evaluation Example

Offline or *ex-post* evaluation is the practice of evaluating the forecasts that would have been produced in the past using a given process. For example, a forecast producer may want to evaluate changes to their forecasting system without having to build up a record of new operational forecasts, or as is common in research, the sole objective may be to study forecasting methodology. In practice, *ex-post* evaluation may not be possible or desirable. For instance, re-forecasts may be prohibitively expensive to produce, or if running a trial or competition it may be prudent to remove the possibility of participants cheating by using using the historic data they are supposed to be predicting.

In our example, the forecast user has determined that absolute error best reflects the cost of forecast errors (as opposed to mean squared/quadratic or a particular quantile), so reducing MAE is the objective. Using the GEFcom2014 data again, suppose the present time is one year since the start of that dataset, so one year of historic data is available. The data is divided into $k = 6$ folds for cross-validation and the folds are stratified to mitigate seasonal effects. Each fold comprises two distinct periods of data separated by half a year, e.g. fold one comprises January and July, fold two comprises February and August, and so on. The exact number of folds k is somewhat arbitrary but should be large enough that $\frac{k-1}{k}\%$ of available data is sufficient for model training, and small enough that the sections of test data are not highly correlated with the surrounding data used in training. Computational burden may also be a consideration. Ten months of training data is reasonable here for our relatively simple models with only tens of parameters to estimate. And auto-correlation of the wind power time series is very low beyond one day (around 5% of test data) so the risk of correlation impacting the results is low.

CV Fold	Benchmark	Proposed
1	0.153	0.151
2	0.158	0.152
3	0.126	0.124
4	0.124	0.123
5	0.127	0.123
6	0.130	0.126
All	0.136	0.133

TABLE 6 Mean Absolute Error (as a proportion of installed capacity) for each cross-validation fold of the offline example, and for all folds combined. The Proposed model has a lower score in each fold and overall, but further analysis is required to determine the significance of this difference.

The Benchmark model from Section 2.1 is implemented with the absolute loss function in order to minimise the target metric, MAE. Similarly, the parameters of the Proposed model (36) are estimated by minimizing the absolute loss. The MAE is calculated for each fold and overall, and the significance tests described in Section 4.3.2 are performed. The results are presented in Table 6 and Figure 12. The results indicate that the Proposed model offers an improvement of around 2% in terms of MAE, and that this is not a result of sampling variability, shown in Figure 12. To further support this conclusion, the Diebold-Mariano test has been performed using the out-of-sample predictions produced by the cross-validation exercise and returned a p-value of $p = 0.002$, well below the standard threshold of $p \leq 0.05$, allowing us to conclude that the MAE of the Proposed method is lower than that of the Benchmark.

Based on this analysis it is reasonable to conclude that switching to the Proposed method will reduce MAE by around 2% in the long run. Great! Now it becomes a business decision as to whether this improvement justifies any costs that may be associated with its implementation. In the next section, we'll see whether this improvement was realised once implemented.

4.4.2 | Operational Evaluation Example

Operational forecasts, those produced to inform decision-making *ex-ante*, may be evaluated for reasons ranging from providing live feed-back to forecast users, to complying with regulation or contractual agreements (e.g. where a financial reward/penalty is attached to forecast performance). In addition, trials of forecast products are special cases of operational forecast evaluation. Operational evaluation is characterised by relatively short evaluation periods and the possibility of interaction with decision-making processes.

To continue our example, consider the situation where forecasts are being supplied by a third party in the case of e.g., a product trial or procurement exercises. We begin at the same point, one year into the GEFcom2014 dataset with forecast models trained on this data. Operational forecast from both the Benchmark and Proposed methods are collected as the trial progress into the future. The end-use of the forecasts is unchanged so we use the same metric as in the offline example, MAE, but now can only evaluate the forecasts that have been collected since the beginning of the trial period. After one week, there is only one week of data to analyse, and so on.

Rather than leaving the evaluation until the end of the trial at some arbitrary point in time, lets consider a periodic evaluation where each week the MAE calculated for the most recent seven days and the aggregate of all data available to date. The MAE for each week of the trail are tabulated in Table 7. From this information, it is very difficult to determine whether one method is superior to the other (in terms of MAE) as both methods have weeks when they out-perform the other, and the magnitude of the difference is variable. This uncertainty is illustrated in the left panel of Figure 13 showing the large uncertainty introduced by sampling variation when only considering individual weeks of the trail.

The evolution of the MAE skill score throughout the trial is illustrated in the right panel of Figure 13. At each week the skill score and uncertainty due to sampling variation has been calculated using all available data up to that point in the trial. This analysis suggests that the results are inconclusive before week 10 at the very earliest, but as noted in Section 4.3.2, the bootstrap approach can be over-confident as it does not consider serial correlation. It is prudent, therefore, to also consider the Diebold-Mariano test, the results of which are presented in Table 8. We see that the difference in forecast performance is not significant at the 0.05 level (i.e. there is a greater than 5% chance that the performance of both methods is the same) at week 10 of the trial. In general the significance of the overall result increases with the trial length, but not exclusively as the auto-correlation of the score differential and the mean differential are also factors. There may also be seasonal variation in forecast performance (and relative forecast performance) which we will not explore in this example. IEA Wind Recommended Practice for Selecting Renewable Power Forecasting Solutions [9] recommends a minimum trial length of three months and ideally one year for these reasons.

After three to four months enough data has been collected to conclude that the Proposed method has a lower MAE than the Benchmark, and that this reduction is of around 2%. This aligns well with our expectations from the offline study, verifying that significant performance differences observed on historic data transfer into operational experience. However, this example also serves as a reminder that an switching methods to reduce

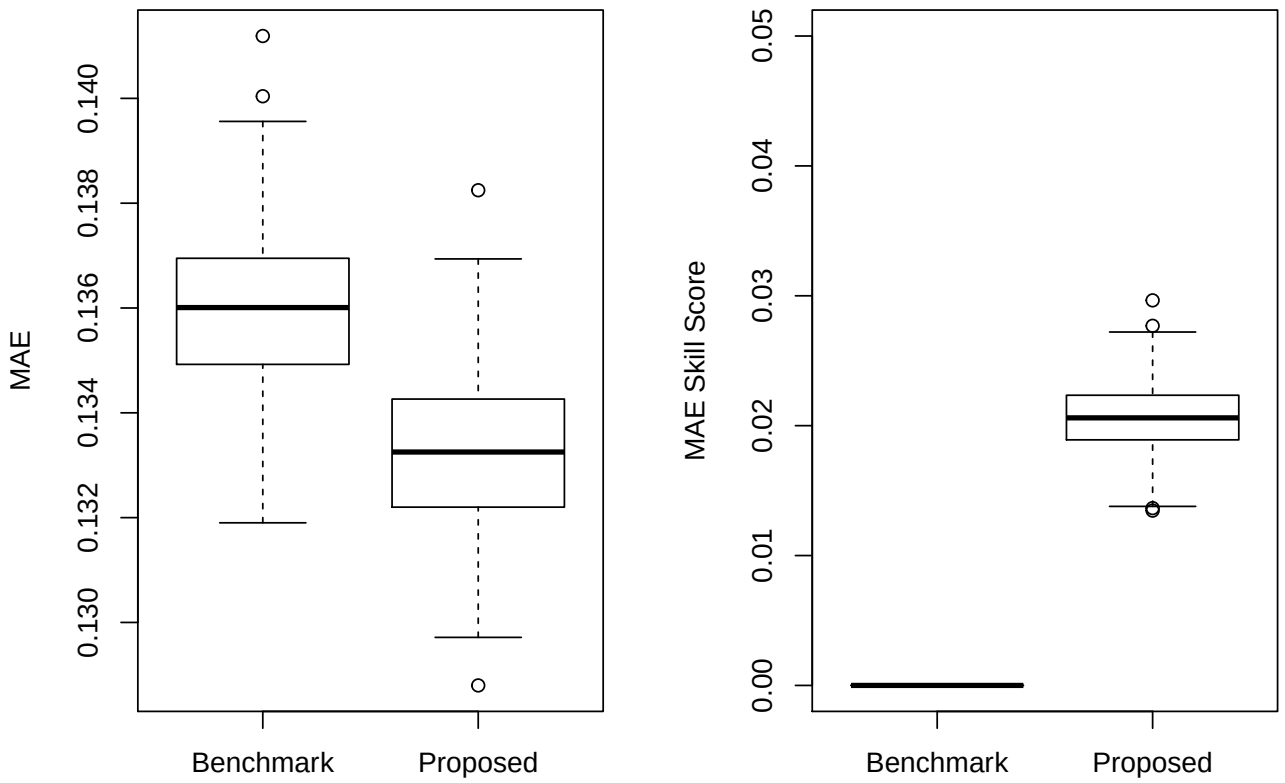

FIGURE 12 Mean Absolute Error (left) and Mean Absolute Error skill score (relative to *Benchmark*, right) for the Benchmark and Proposed models across the combined cross-validation folds. Box plots illustrate the sampling variation estimated using the bootstrapping approach described in Section 4.3.2 with 250 bootstrap samples. The right plot indicates that the median skill score is 0.021 with an interquartile range of 0.004, therefore uncertainty due to sampling variation is low and we can be confident that the long-run skill score will be close to the median.

an average score does not mean that each individual prediction will be more accurate, and that other qualities not captured by scores may be of interest and value to some users.

5 | CONCLUSION

Forecasting has become an important part of the successful integration of wind power in energy systems and markets. Evaluating of these forecasts is very important for selecting a forecast provider, quality control, or forecast model development. Most of the time, forecast errors can be related to some kind of costs and ideally the evaluation should provide information about these expected costs. Since wind power forecast users can be very different such as wind park operators, distribution system operators, transmission system operators, or traders, the forecasts are also used for different applications with different error costs. Therefore, it is important to adjust the forecast evaluation setup to the specific needs of the forecast user. Nevertheless, often just standard evaluation protocols are used and therefore the drawn conclusion might not always be ideal. Furthermore, with the advent of new advanced forecast products such as probabilistic or multivariate predictions also new less intuitive evaluation techniques have been proposed and the risk of selecting inappropriate evaluation approaches has even increased.

This paper revisited different forecast evaluation approaches with a specific focus on selecting the right methods for the specific needs of a forecast user. In the first part of the paper a simple example case showed that the selection of the right metric is crucial to find the best forecast system for the application these forecasts are needed. Furthermore, it is very important to use an appropriate evaluation setup (e.g., to use a large

Week	Benchmark	Proposed	Week	Benchmark	Proposed	Week	Benchmark	Proposed
1	0.133	0.133	11	0.131	0.118	21	0.080	0.082
2	0.120	0.119	12	0.143	0.140	22	0.110	0.110
3	0.129	0.133	13	0.110	0.101	23	0.105	0.122
4	0.158	0.143	14	0.129	0.114	24	0.136	0.127
5	0.139	0.136	15	0.092	0.091	25	0.093	0.089
6	0.113	0.116	16	0.124	0.124	26	0.104	0.109
7	0.175	0.170	17	0.118	0.120	27	0.177	0.173
8	0.187	0.192	18	0.077	0.081	28	0.107	0.107
9	0.152	0.148	19	0.195	0.195	29	0.153	0.145
10	0.116	0.112	20	0.148	0.142	30	0.120	0.112

TABLE 7 Mean absolute error for each week of trial. Weekly MAE is volatile, and it is difficult to determine if one method is consistently outperforming the other without further analysis.

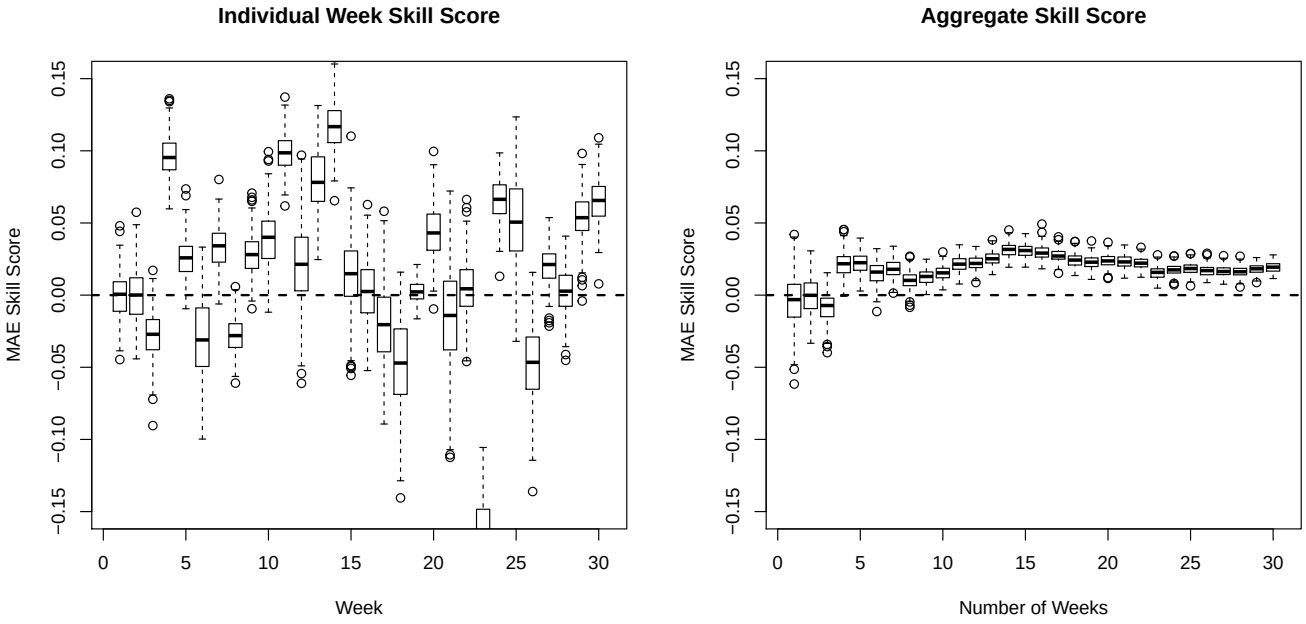

FIGURE 13 Mean absolute error skill score (*Proposed* relative to *Benchmark*) for individual weeks of online evaluation (left) and for aggregation of available data as the trial progresses (right). Box plots illustrate the sampling variation estimated using the bootstrapping approach described in Section 4.3.2 with 250 bootstrap samples. Results for individual weeks provide inconclusive evidence, but once data for multiple weeks is aggregated the proposed method emerges as significantly more skillful.

enough data set) and know how to interpret the results. The second part then presented and discussed a number of metrics that can be useful in wind power applications and the third part discussed the evaluation setup and interpretation of results.

To facilitate reproduction and future work, all the data and code that were used to generate the results are available for download at [21]

ACKNOWLEDGEMENTS

This manuscript was created as part of the International Energy Agency (IEA) Wind Task 36 on Forecasting for Wind Energy and we thank in particular Markus Abel, Jan Dobschinski, Gregor Giebel, Gianni Goretti, Sue Ellen Haupt, Alexander Kann, Henrik Madsen, Corinna Möhrlen, Will Shaw,

Week	Benchmark	Proposed	Skill Score	DM p-value
1	0.133	0.133	-0.2%	0.850
5	0.136	0.133	2.3%	0.149
10	0.142	0.140	1.5%	0.183
15	0.135	0.131	3.1%	0.005
30	0.129	0.127	1.9%	0.027
45	0.136	0.133	2.4%	0.002

TABLE 8 Mean Absolute Error, skill score of Proposed method relative to Benchmark, and Diebold-Mariano test statistic for aggregation of available data as the trial progresses (right). After 10 weeks the results are not significant using the standard threshold for $p \leq 0.05$.

Thorsten Simon, Aidan Tuhoy, Stephan Vogt, and John Zack for their comments and participation in various discussions. Jakob W. Messner was supported by EUDP-64015-0559. Jethro Browell is supported by an EPSRC Innovation Fellowship (EP/R023484/1). Research Data Statement: all data associated with this paper, including code to reporduce all results, can be found at [21] .

References

1. WindEurope . Wind in Power 2017. 2018.
2. Sheridan P. Current Gust Forecasting Techniques, Developments and Challenges. *Advances in Science and Research* 2018; 15: 159–172. doi: 10.5194/asr-15-159-2018
3. Giebel G, Brownsword R, Kariniotakis G, Denhard M, Draxl C. *The State-Of-The-Art in Short-Term Prediction of Wind Power: A Literature Overview, 2nd Edition.* ANEMOS.plus . 2011.
4. Kariniotakis G. *Renewable Energy Forecasting – From Models to Applications.* Elsevier . 2017.
5. Murphy A, Winkler R. A General Framework for Forecast Verification. *Monthly Weather Review* 1987; 115(7): 1330-1338. doi: 10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2
6. Murphy A. What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting* 1993; 8(2): 281-293. doi: 10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2
7. Landberg L, Giebel G, Badger J, et al. *Ensemble Predictions: Understanding Uncertainties*. 2. ; 2007.
8. Bessa RJ, Miranda V, Botterud A, Wang J. 'Good' or 'Bad' Wind Power Forecasts: A Relative Concept. *Wind Energy* 2010; 14(5): 625-636. doi: 10.1002/we.444
9. Möhrlen C, Zack J, Lerner J, Messner JW, Browell J. Recommended Practice on Forecast Solution Selection. 2019.
10. Madsen H, Pinson P, Kariniotakis G, Nielsen HA, Nielsen TS. Standardizing the Performance Evaluation of Short-Term Wind Power Prediction Models. *Wind Engineering* 2005; 29(6): 475-489. doi: 10.1260/030952405776234599
11. Pinson P, Girard R. Evaluating the Quality of Scenarios of Short-Term Wind Power Generation. *Applied Energy* 2012; 96: 12-20. doi: 10.1016/j.apenergy.2011.11.004
12. Jolliffe IT, Stephenson DB. *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Wiley . 2012.
13. Murphy AH, Katz RW, Winkler RL, Hsu WR. Repetitive Decision Making and the Value of Forecasts in the Cost-Loss Ratio Situation: A Dynamic Model. *Monthly Weather Review* 1985; 113(5): 801-813. doi: 10.1175/1520-0493(1985)113<0801:RDMATV>2.0.CO;2
14. Gneiting T. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association* 2011; 106(494): 746-762. doi: 10.1198/jasa.2011.r10138
15. Richardson DS. *Economic Value and Skill*ch. 9: 167-184; Wiley-Blackwell . 2012

16. Roulston MS, Smith LA. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review* 2002; 130(6): 1653-1660. doi: 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2

17. Thorarinsdottir T, Schuhen N. Verification: Assessment of Calibration and Accuracy. In: Vannitsem S, Wilks DS, Messner JW., eds. *Statistical Postprocessing of Ensemble Forecasts*Elsevier. 2018.

18. Wilks DS. A Skill Score Based on Economic Value for Probability Forecasts. *Meteorological Applications* 2001; 8(2): 209-219. doi: 10.1017/S1350482701002092

19. Katz R, Murphy A. *Economic Value of Weather and Climate Forecasts.* Cambridge University Press . 1997.

20. Hong T, Pinson P, Fan S, Zareipour H, Troccoli A, Hyndman RJ. Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 and Beyond. *International Journal of Forecasting* 2016; 32(3): 896 - 913. doi: 10.1016/j.ijforecast.2016.02.001

21. Messner J, Browell J. Supplementary material for "Evaluation of Wind Power Forecasts – An up-to-date view". https://doi.org/10.15129/3cf00c9b-e891-433e-afe3-732577aa74d2;

22. Pinson P, Nielsen HA, Madsen H, Nielsen TS. Local Linear Regression with Adaptive Orthogonal Fitting for the Wind Power Application. *Statistics and Computing* 2008; 18(1): 59–71. doi: 10.1007/s11222-007-9038-7

23. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 2007; 102(477): 359-378. doi: 10.1198/016214506000001437

24. Pinson P, Chevallier C, Kariniotakis GN. Trading Wind Generation From Short-Term Probabilistic Forecasts of Wind Power. *IEEE Transactions on Power Systems* 2007; 22(3): 1148-1156. doi: 10.1109/TPWRS.2007.901117

25. Ehm W, Gneiting T, Jordan A, Krüger F. Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations and Forecast Rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2016; 78(3): 505–562. doi: 10.1111/rssb.12154

26. Gneiting T, Vogel P. Receiver Operating Characteristic (ROC) Curves. *arXiv e-prints* 2018: 1809.04808.

27. Murphy AH. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology* 1973; 12(4): 595-600. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

28. Wilks DS. *Statistical Methods in the Atmospheric Sciences.* Academic Press. 3 ed. 2011.

29. Bröcker J, Smith LA. Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting* 2007; 22(3): 651-661. doi: 10.1175/WAF993.1

30. Pinson P, McSharry P, Madsen H. Reliability Diagrams for Nonparametric Density Forecasts of Continuous Variables: Accounting for Serial Correlation. *Quarterly Journal of the Royal Meteorological Society* 2010; 136(646): 77-90. doi: 10.1002/qj.559

31. Bremnes JB. Probabilistic Wind Power Forecasts using Local Quantile Regression. *Wind Energy* 2004; 7(1): 47-54. doi: 10.1002/we.107

32. Dobschinski J, Bessa R, Du P, et al. Uncertainty Forecasting in a Nutshell: Prediction Models Designed to Prevent Significant Errors. *IEEE Power and Energy Magazine* 2017; 15: 40-49. doi: 10.1109/MPE.2017.2729100

33. Bessa RJ, Möhrlen C, Fundel V, et al. Towards Improved Understanding of the Applicability of Uncertainty Forecasts in the Electric Power Industry. *Energies* 2017; 10(9). doi: 10.3390/en10091402

34. Laio F, Tamea S. Verification Tools for Probabilistic Forecasts of Continuous Hydrological Variables. *Hydrology and Earth System Sciences* 2007; 11(4): 1267–1277. doi: 10.5194/hess-11-1267-2007

35. Hersbach H. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting* 2000; 15(5): 559-570. doi: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

36. Tastu J, Pinson P, Madsen H. Multivariate Conditional Parametric Models for a Spatiotemporal Analysis of Short-Term Wind Power Forecast Errors. In: ; 2010: 77–81.

37. Dawid AP, Sebastiani P. Coherent Dispersion Criteria for Optimal Experimental Design. *Annals of Statistics* 1999: 65–81.

38. Diks C, Panchenko V, Van Dijk D. Likelihood-Based Scoring Rules for Comparing Density Forecasts in Tails. *Journal of Econometrics* 2011; 163(2): 215–230. doi: 10.1016/j.jeconom.2011.04.001

39. Scheuerer M, Hamill TM. Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities. *Monthly Weather Review* 2015; 143(4): 1321–1334. doi: 10.1175/MWR-D-14-00269.1

40. Efron B. Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and other Methods. *Biometrika* 1981; 68(3): 589-599. doi: 10.1093/biomet/68.3.589

41. Diebold FX, Mariano RS. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 1995; 13(3): 253-263. doi: 10.1080/07350015.1995.10524599

42. Magnusson L, Källén E. Factors Influencing Skill Improvements in the ECMWF Forecasting System. *Monthly Weather Review* 2013; 141(9): 3142–3153. doi: 10.1175/MWR-D-12-00318.1

43. Hastie TJ, Tibshirani RJ. *Generalized Additive Models.* Chapman and Hall . 1990.