RECOMMENDED PRACTICE FOR THE IMPLEMENTATION OF RENEWABLE ENERGY FORECASTING SOLUTIONS

- Part 3: Forecast Solution Evaluation -

2. EDITION

Draft for Review by Executive Committee of the International Energy Agency Implementing Agreement

Prepared by IEA Wind Task 36

Document Version: 2.0
November 7, 2021

Contents

Preface

This recommended practice document is the result of a collaborative work that has been edited by the undersigning authors in alignment with many discussions at project meetings, workshops and personal communication with colleagues, stakeholders and other interested persons throughout the phase 1 (2016-2018) and the phase 2 (2019-2021) of the IEA Wind Task 36 as part of workpackage 2.1 and 3.1.
The editors want to thank co-authors, participants, contributors and supporter of meetings, workshops and sessions that contributed to the discussions, provided feedback or other input throughout the past 6 years.

IEA Wind Task 36, November 7, 2021

Editors and Authos:
Dr. Corinna Möhrlen (WEPROG, DK) <com@weprog.com>
Dr. John Zack (UL AWS Truepower) <john.zack@ul.com>

Contributing Authors:
Dr. Jakob W. Messner (MeteoServe Wetterdienst, AT)
Dr. Jethro Browell (University of Glasgow, UK)

Contributions:
Dr. Craig Collier (Energy Forecasting Solutions, USA)
Dr. Aidan Tuohy, (EPRI, USA)
Dr. Stephan Vogt (Fraunhofer Institute IEE, DE)

Supported by:
Operating Agent Dr. Gregor Giebel (Danish Technical Univerity, DTU Wind, DK)

Chapter 1

Background and Objectives

1.1 BEFORE YOU START READING

This is the **third part** of a series of three recommended practice documents that deal with the selection, development and operation of forecasting solutions in the power market. It provides information and guidelines regarding effective evaluation of forecasts, forecast solutions and benchmarks and trials.

The **first part** *Forecast Solution Selection Process* deals with the selection and background information necessary to collect and evaluate when developing or renewing a forecasting solution for the power market. The **second part**, *Design and Execution of Benchmarks and Trials*, of the series deal with benchmarks and trials in order to test or evaluate different forecasting solutions against each other and the fit-for-purpose. The **third part**, *Forecast Solution Evaluation*, which is the current document, provides information and guidelines regarding effective evaluation of forecasts, forecast solutions and benchmarks and trials. The **fourth part**, *Meteorological and Power Data Requirements for real-time forecasting Applications* provides guidance for the selection, deployment and maintenance of meteorological sensors and the quality control of the data produced by those sensors with the objective of maximising the value of the sensor data for real-time wind and solar power production forecasting.

If your main interest is in (1) selecting a forecasting solution, (2) testing or evaluating different forecasting solutions against each other, or (4) setting up meteorological sensors or power measurements for real-time wind or solar power forecasting, please move on to part 1, 2 or 4 of this recommended practice guideline to obtain recommendations on any of these specific issues, respectively.

It is also recommended that the table of contents be actively used to find the most relevant topics.

1.2 Introduction

The evaluation of forecasts and forecast solutions is an essential task for any forecast provider as well as end-user of forecasts. It is important because economically significant and business-relevant decisions are often based on evaluation results. Therefore, it is crucial to allocate significant attention to the design and execution of forecast evaluations to ensure that the results are significant, representative and relevant. Additionally, forecast skill and quality has to be understood and designed in the framework of forecast value in order to evaluate the quality of a forecast on the value it creates in the decision processes. This second edition of the recommended practices guideline focuses on a number of conceptual processes to introduce a framework for evaluation of wind and solar energy forecasting applications in the power industry. A comprehensive outline of forecast metrics is not part of this guideline. There are a number of very useful and comprehensive publications available (e.g. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11**?**]) which will be specifically referenced. A state-of-the-art of forecast evaluation is also not part of this guidelines, as the process of standardization has only just started in the community. This topic will be covered in one of the next versions of this guideline.

This part of the recommended practices guideline focuses on:

1. *Impact of forecast accuracy on application*
   First, it's often difficult to define the impact of forecast accuracy impact on the performance of an application metric, because forecasts are just one of many inputs. Second, trials or benchmarks often last longer than anticipated or are too short to generate trustworthy results. Thus, the Forecast User is often under pressure to either wrap up the evaluation quickly or to produce meaningful results with too little data. As a consequence, average absolute or squared errors are employed due to their simplicity, even though they seldom reflect the quality and value of a forecast solution for the Forecast User's specific applications.

2. *Cost-Loss relationship of forecasts*
   A forecast that performs best in one metric is not necessarily the best in terms of other metrics. In other words, there is no universally best evaluation metric. Using metrics that do not well reflect the relationship between forecast errors and the resulting cost in the Forecast User's application, can lead to misleading conclusions and non-optimal (possibly poor) decisions. Knowing the cost-loss relationship of their applications and to be able to select an appropriate evaluation metric accordingly is important. This becomes especially important as forecasting products are becoming more complex and the interconnection between errors and their associated costs more proportional. Apart from more meaningful evaluation results, knowledge of the cost-loss relationship also helps the forecast service provider to optimize forecasts and develop custom forecast solutions that are tailored for the intended application.

Evaluation of forecast solutions is a complex task and it is usually neither easy nor

recommended to simplify the evaluation process. As a general recommendation, such a process needs to follow an evaluation paradigm that assigns the an appropriate level of attention to three core principles of an evaluation process:

1. **representativeness**
2. **significance**
3. **relevance**

The setup of an evaluation process that satisfactorily addresses these three principles is the focus of this recommended practice guideline.

In chapter 2 these three main principles are outlined and the general concept of evaluation uncertainty is explained as this should be the basis for any evaluation task. In chapter 3, the uncertainty of measurement data collection and reporting is explained as a basic issue of evaluation and verification tasks. If forecasts are evaluated against data that has inherit errors, results may still show some significance, but may no longer be considered relevant and representative. In chapter 4 metrics for evaluation and verification will be conceptualized and categorized in order to provide an issue-oriented guideline for the selection of metrics in a evaluation framework. The last chapter 5 introduces the concept of developing such an evaluation framework and provides practical information on how to maximize value of operational forecasts and also how to conduct an evaluation for benchmarks, trials and new forecasting techniques. Lastly, recommendations are made for a number of practical use cases for power industry specific applications.

Chapter 2

Overview of Evaluation Uncertainty

Key Points
All performance evaluations of potential or ongoing forecast solutions have a degree of uncertainty, which is associated with the three attributes of the performance evaluation process: (1) representativeness, (2) significance and (3) relevance.
A carefully designed and implemented evaluation process that considers the key issues in each of these three attributes can minimize the uncertainty and yield the most meaningful results.
A disregard of these issues is likely to lead to uncertainty that is so high that the conclusions of the evaluation process are meaningless and therefore decisions based on the results are basically random.

Uncertainty is an inherent characteristic of the forecast evaluation process. The objective of the design and execution of a forecast evaluation procedure is to minimize the uncertainty and thereby reduce its impact on the decisions association with forecast selection or optimization. In order to minimize forecast evaluation uncertainty it is useful to understand the sources of uncertainty within the evaluation process.

The sources of forecast evaluation uncertainty can be linked to three key attributes of the evaluation process: (1) representativeness (2) significance and (3) relevance. If any one of these are not satisfactorily addressed, than an evaluation will not provide meaningful information to the forecast solution selection process and the resources employed in the trial or benchmark will essentially have been wasted. Unfortunately, it may not be obvious to the conductor of a forecast evaluation or the user of the information produced by an evaluation whether or not these three attributes have been satisfactorily addressed. This section will present an overview of the key issues associated with each attribute. Subsequent sections of this document will provide guidance on how to maximize the likelihood that each will be satisfactorily addressed.

2.1 Representativeness

Representativeness refers to the relationship between the results of a forecast performance evaluation and the performance that is ultimately obtained in the operational use of a forecast solution. It essentially addresses the question of whether or not the results of the evaluation are likely to be a good predictor of the actual forecast performance that will be achieved for an operational application. These are many factors that influence the ability of the evaluation results to be a good predictor of future operational performance.

Four of the most crucial factors are:

1. size and composition of the evaluation sample,
2. quality of the data from the forecast target sites,
3. the formulation and enforcement of rules governing the submission of forecasts (sometimes referred to as "fairness"),
4. availability of a complete and consistent set of evaluation procedure information to all evaluation participants (sometimes referred to as "transparency")

2.1.1 Size and composition of the evaluation sample

The size and composition of the evaluation sample are the most important representativeness factors. Both the size and composition of the sample is a key factor in determining the extent to which the results are influenced by random variation, or noise, compared to true differences in forecast skill. The followinf considerations are recommended in order to ensure representative evaluation samples:

- **Data set representation and composition:**
  The selected data set should be representative for the application and forecasts should be compared with exactly the same data sets. Results of different locations, seasons, lead times etc. are in general not comparable. The composition should be constructed so that all significant modes of variation of the forecast variable (e.g. wind or solar power production) are included in the evaluation sample. For example, if there is a high wind season and a low wind season, then both should have a representative number of cases in the evaluation sample. Or, in the case of solar power forecasts, periods of cloudy weather should be included equally much than periods of clear sky periods. However, if this is not practical, then there should at least be a representative sample of the most important modes for the application.

- **Data set length:**
  The size of the evaluation sample is one of the most important representativeness and significance factors. The size of the sample is a key factor in determining to what extent results are influenced by random variation, or noise, compared to true predictive

performance. The use of a small sample increases the probability that any conclusions reached from the evaluation will be due to noise (random and unrepresentative events) in the sample. For example, the occurrence of very unusual weather events for a few days in a short sample may dominate the evaluation results.

That leads to the question of how large of a sample is adequate? A commonly used target sample size guideline when gathering data for statistical analysis is 30. If all the sample points are independent, then a sample of 30 provides a reasonable adequate minimization that sampling noise will impact the conclusions. But the key phrase is that the sample data points must be independent (uncorrelated) for this guideline to be valid. However, weather processes are typically highly correlated over time periods of 3 to 4 days. This means that an adequate sample from a continuous evaluation period should be 3 to 4 times larger than 30 or in other words, 90 to 120 days [10] (see also 5.1.2, 4.1.4 and 5.1.3.1).

- **Data set consistency:**
  For a fair evaluation of a forecast, whether against other forecasts, measurements or persistence, it is very important to use the same data set to derive the evaluation results. If a certain forecast is not available for a specific time, this time has to be disregarded for all the other forecasts or persistence as well. Else, if forecasts are for example 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. This also applies for curtailment data. It is important to evaluate a forecast against the weather related performance and remove all non-weather related impacts that are out of the forecasters control. Especially, if forecasts are evaluated against a persistence forecast, e.g. in minute- or hour scale forecasts, where models are adopted to measurements that may contain curtailment or failures due to turbine unavailability or communication issues. In such cases, the corresponding persistence need to be computed accordingly. If this is not done, the forecast performance of the persistence will be overestimated and the performance of the forecast underestimated.

2.1.2 Data Quality

The quality of the data used in the forecast evaluation process can be a major source of uncertainty. The data from the forecast target location is typically used for three purposes: (1) training data for the statistical components of each forecast system, (2) real-time input data to forecast production processes, which is especially important for very short-term forecast time horizons (minutes to a few hours ahead) and (3) forecast outcome data (i.e. the actual value used to compute forecast errors) for the evaluation of the forecast performance. If the data have many quality issues, then the representativeness of all three of these applications is compromised. The quality issues may include: (1) out of range or locked values, (2) biased values due to issues with measurement devices or location of measurement, (3) badly or not at all calibrated instruments and (4) power production data that are unrepresentative of

meteorological conditions because of undocumented generation outages or curtailments. If a substantial amount of data with these issues is used in the evaluation process for any of the three purposes, the results will likely not be representative of the true skill of the forecasting solutions that are being evaluated.

2.1.3 Forecast Submission Control

A third important factor is the formulation and enforcement of rules for the submission of forecasts in the evaluation process. This is sometimes noted as a "fairness" issue and it is indeed an issue of fairness to the forecast providers who are typically competing to demonstrate the skill of their system and thereby obtain an award of a contract for their services. However, from the user's perspective, it is a representativeness issue. If it is possible for some forecasting solution providers to provide forecasts with unrepresentative skill then the conclusions of the entire evaluation process are questionable. A couple of examples can illustrate this point. One example is a situation in which there is no enforcement of the forecast delivery time. In this case it would be possible for a forecast provider to deliver forecasts at a later time (perhaps overwriting a forecast that was delivered at the required time) and use later data to add skill to their forecast or even wait until the outcome for the forecast period is known. Although one might think that such explicit cheating is not likely to occur in this type of technical evaluation, experience has indicated that it is not that uncommon, if the forecast delivery protocol enables its occurrence.

A second example, illustrate how the results might be manipulated without explicit cheating by taking advantage of loopholes in the rules. In this example the issue is that the evaluation protocol does not specify any penalty for missing a forecast delivery and the evaluation metrics are simply computed on whatever forecasts are submitted by each provider. As a forecast provider it is not difficult to estimate the "difficulty" of each forecast period and to simply not deliver any forecasts during periods that are likely to be difficult and therefore prone to large errors. This is an excellent way to improve forecast performance scores. Of course, it makes the results unrepresentative of what is actually needed by the user. Often it is good performance during the difficult forecast periods that is most valuable to a user.

2.1.4 Process Information Dissemination

A fourth key factor is the availability of a complete and consistent set of information about the forecast evaluation process to all participants. Incomplete or inconsistent information distribution can occur in many ways. For example, one participant may ask a question and the reply is only provided to the participant who submitted the inquiry. This can contribute to apparent differences in forecast skill that are not associated with true differences in the skills of the solution. This of course results in unrepresentative evaluation of the true differences in forecast skill among the solutions.

2.2 Significance

Significance refers to the ability to differentiate between performance differences that are due to noise (quasi-random processes) in the evaluation process and those that are due to meaningful differences in skill among forecast solutions. Performance differences that stem from noise have basically no meaning and will not represent the performance differences that a user will experience in a long-term operational application of a solution. *Real* performance differences on the other hand should be stable and should not change if an evaluation process is repeated, e.g., one year later. A certain degree of noise is inevitable in every evaluation task but both, minimization of noise and awareness of the uncertainty it causes are crucial in order to make reliable decisions on the evaluation results.

As mentioned above, repeatability is a good practical indication of significance in evaluation results. The highest potential for achieving repeatability is the use of a representative evaluation sample. This means the sample should cover as many potential weather events, seasons, and perhaps forecast locations as possible. Otherwise, there is a high probability that the results will be different for features that are not well represented in the evaluation sample. Thus, significance is highly related to representativeness and very much depends on the evaluation sample size and composition.

2.2.1 Quantification of Uncertainty

In addition to noise minimization through the use of representative evaluation data sets, it is also very useful to quantify the significance (i.e. the uncertainty) of the evaluation results. Quantification of the uncertainty is important for decision making. For example, if a number of forecast solutions are evaluated with a specified metric, but their differences are much smaller than the uncertainty in the result due to e.g. measurement uncertainty, the meaning of their ranking is actually very limited and should not be used for important decisions.

2.2.1.1 Method 1: Repeating the evaluation task

The simplest approach to estimate evaluation uncertainty would be to repeat the evaluation task several times on different data sets. This approach is often effective, because the variation or uncertainty of the evaluation results is typically attributable largely to their dependence on the evaluation data set and therefore results often vary among different evaluation data sets. However, since evaluation data sets are usually very limited, this is often not a feasible approach.

2.2.1.2 Method 2: Bootstrap Resampling

A simple alternative method is to simulate different data sets, through the use of a bootstrap resampling process. In this approach an evaluation data set of the same length as the original data set is drawn from the original data set with replacement and the evaluation results are derived on this set. By repeating this "N" times, "N" different evaluation results become

available and their range can be seen as the evaluation uncertainty. Alternatively, parametric testing can also provide information on the significance of evaluation results. Typically two sample paired t-tests applied on the sets of error measures for each event provide a good estimate of the significance of the results. Diebold et al. [**?**] proposed a variation of this t-test to account for temporal correlations in the data and can therefore provide a more accurate significance quantification. Messner et al. [10] also describes different parametric testing or bootstrap resampling approaches that can be employed to quantify the evaluation uncertainty.

If it is found, that the forecast that is identified as the "best" from an evaluation process does not exhibit significantly better performance than some of the other benchmark participants, the final selection of forecast solutions should only consider differences among forecast solutions that are significant. For example, if there is a group of forecast solutions that are at the top of the metric-based performance ranking list, but there are no significant differences in performance among them, the selection process should treat them as equivalent in terms of forecast accuracy and the differentiation among them should be based on other factors.

2.3 Relevance

Relevance refers to the degree of alignment between the evaluation metrics used for an evaluation and the true sensitivity of a user's application(s) to forecast error. If these two items are not well aligned then even though an evaluation process is representative and the results show significant differences among solutions, the evaluation results may not be a relevant basis for selecting the best solution for the application. There are a number of issues related to the relevance factor.

1. Best Performance Metric
   First, the selection of the best metric may be complex and difficult. The ideal approach is to formulate a cost function that transforms forecast error to the application-related consequences of those errors. This could a monetary implication or it might be another type of consequence (for example a reliability metric for grid operations). However, if it is not feasible to do this, another approach is to use a matrix of performance metrics that measure a range of forecast performance attributes.

2. Multiple Performance Metrics
   If there is a range of forecast performance attributes that are relevant to a user's application, it most likely will not be possible to optimize a single forecast to achieve optimal performance for all of the relevant metrics. In that case, the best solution is to obtain multiple forecasts with each being optimized for a specific application and its associated metric.

3. Multiple Forecast Solutions
   Another type of issue arises when the user intends to employ multiple (N) forecast

solutions and create a composite forecast from the information provided by each individual forecast. In this case it may be tempting to select the best N performing forecasts in the evaluation according to the metric or metrics identified as most relevant by the user. However, that is not the best way to get the most relevant answer for the multiple provider scenario. In that case the desired answer is to select the N forecasts that provide the best composite forecast for the target metric(s). This may not be the set of N forecasts that individually perform the best. It is the set of forecasts that best complement each other. For example, the two best forecasts according to a metric such as the RMSE may be highly correlated and provide essentially the same information. In that case, a forecast solution with a higher (worse) RMSE may be less correlated with the lowest RMSE forecast and therefore be a better complement to that forecast.

Chapter 3

Measurement Data Processing and Control

Key Points

- *Measurements from the forecast target facilities are crucial for the forecast production and evaluation process and therefore much attention should be given to how data is collected, communicated and quality controlled*
- *Collection and reporting of measurement data requires strict rules and formats, as well as IT communication standards in order to maximize its value in the forecasting process; information about standards and methods for collecting and reporting data is provided in Chapter 4 of Part 1 of this RP.*
- *An effective quality control process is essential since bad data can seriously degrade forecast performance; standard quality maintenance and control procedures have been documented and some are noted in this section*

In any evaluation the measurements or observations are alpha and omega for trustworthy results. For this reason, this section is dedicated to the importance of data collection, verification and the identification of the measurement uncertainty. In the evaluation of wind power forecasts, power data is most important but also meteorological measurements are often provided to the forecast providers as input to improve their forecast models. Furthermore, failure of generation assets (unscheduled outages), service periods of generation assets (scheduled outages), curtailment and other disturbances in the power measurements can have significant impact on the results of an evaluation. The following section deal with these aspects and provide recommendations for a correct handling of such data for the evaluation phase.

3.1 Uncertainty of instrumentation signals and measurements

All data are derived from different measurement devices and depending on the quality of these devices the measurements can deviate from the reality to a certain degree. In fact, measurement errors can never be avoided completely and can potentially affect the representativeness or significance of evaluation results. Therefore, it is crucial to establish and maintain specific quality requirements for the measurement devices to obtain data of good quality and thus keep the measurement uncertainty to a low level. This will not only improve the significance and representativeness of the evaluation results, but also assure an optimum quality of forecasts that use the measurements as input.

For power data, the measurement quality is usually ensured by existing grid code standards that are verified in the commissioning phase and are serviced as part of the turbine's SCADA system maintenance.

Recommendations on minimum technical requirements beyond the scope of this recommended practice guideline. For anyone intending to collect and process bankable wind measurements, the following standards and guidelines provide a basis for the adaptation of those measurements for real-time operational applications :

1. the International Electrotechnical Committee (IEC)
2. the International Energy Agency (IEA)
3. the International Network for Harmonised and Recognised Wind Energy Measurement (MEASNET)
4. United States Environmental Protection Agency (EPA)

If these requirements are fulfilled, the measurement error is usually negligible compared to other sources of uncertainty in the evaluation procedure.

3.2 Measurement data reporting and collection

Once wind farms are operational and the production data are measured it is important to collect, store and report them properly, which requires strict rules and formats, as well as IT communication standards. Standard protocols for collecting and reporting power data are usually enforced by jurisdictional grid codes. There are however a number of aspects that are not covered in the grid codes that are essential for the evaluation of forecasting tools. This section will discuss the main aspects to be considered for any measurement data collection and archiving. In the following, the description is limited for the purpose of evaluation of forecasts in a real-time operational framework or a forecast test framework.

3.2.1 Non-weather related production reductions

Raw power production data contains a number of non-weather related reductions that need consideration in the collection or archiving of measurement data, such as

- failure of turbines in a wind park (availability)
- scheduled and non-scheduled maintenance
- curtailment
- reductions due to environmental constraints (noise, birds, ...)

The so-called "Net to Grid" signal is often disturbed by such technical constraints that are usually not part of the wind power forecasting task. Therefore, to evaluate the actual forecast quality, such events have to be filtered in the evaluation. Especially in the case of curtailment, the forecast user needs to decide whether the target parameter is the real power production or available power. If it is the latter, data with curtailment should be removed from the evaluation data set, because errors are not meaningful for the forecast performance, unless the curtailments are predicted as well. Ideally, direct signals from the turbines on their available active power, even if in retrospective manor, are used to filter such data for evaluation and verification purposes.

- To receive *relevant* results, remove events from the evaluation data set that are affected by non-weather related production constraints unless these are to be predicted as well.

3.2.2 Aggregation of measurement data in time and space

Often, temporally or spatially aggregated data (averages, sums) are more useful in power applications than instantaneous signals. The aggregation level of the data should always be communicated to the forecast provider to assure optimum forecast performance for the intended application. This also applies in the absence of any aggregation over time, for example, it should be explicitly specified, if hourly values are provided that are not hourly averages of higher resolution data, but instantaneous values taken at the start of the hour. Furthermore, it is strongly recommended that the measurement data should be aggregated according to the intended applications before comparing, analysing and verifying forecasts. Otherwise, the evaluation results might not be relevant for the forecast user.

When aggregating measurement data over parks, regions, control zones or other aggregation levels, it is important to consider non-weather related events as discussed in Section 3.2.1. In particular:

- Non-reporting generation units
- IT communication failures or corrupt signals

have to be identified and reported, and the aggregated data should be normalized accordingly. Such failures are impossible to predict by the forecast vendor and should therefore not be part of the evaluation process.

- For *relevant* results, average the measurement data over a time frame that is also useful for the intended application.
- For *representative* results, non-weather related events should be identified and the aggregated signals normalized accordingly.

3.3 Measurement data processing and archiving

In any real-time environment, measurements should be delivered as is, but flagged, if they are considered wrong (1) at the logger level and (2) after a quality control before employing measurements in a forecast process.

The optimal data archival structure is dependent on the plans for the further processing of the data. In most cases, it is useful to archive data in a database. There are many structures of databases available today. Such structural decisions are out of the scope of this guideline. Nevertheless, there are general considerations when planning and designing a database for operational data. For example, a data point may get "stuck" on a set value. Monitoring of the incoming data is therefore an important feature to ensure correct measurement data, where this is possible.

While measurements are available only at one specific time, forecast data have overlapping time periods and need to be separated from measurement data. At the design level, it is necessary to consider the following aspects.

1. single or multiple time points per measurement signal in database
2. flagging at each data point and
   (a) possibility to overwrite corrupt data in database
   (b) possibility to add correct data point in database
   (c) knowledge of time averaging level of data signal
3. single or multiple measurement points per wind farm
4. ability to expand and upscale the database: expansion with increasing number of measurement points/production units
5. importance of access to historical data

The database dimensions and setup of tables has to take such decisions and requirements into consideration.

3.4 Quality assurance and quality control

Quality of data is a crucial parameter for any real-time forecasting system. If the data that real-time forecasts are based on are corrupt or misleading, the result can be worse than not having measurements or observations at all. Therefore, any real-time system using measurements needs a quality control mechanism to discard bad data. However, reasons for bad, corrupt or misleading data signals are almost unlimited, which means that specific limits, operating ranges and validity checks need to be established when dealing with observational data. Is it also worth mentioning that as well as the quality, the latency of the data, i.e. the lag between live time and data being available for forecast use, is critical for live applications. This will affect what time lags can practically be used in any forecast model.
While all this is critical in real-time environments, the quality of measurement data in the verification phase is equally important. For example, if a wind power forecast is verified against observations from a wind farm and a maintenance schedule or a curtailment from the system operator is not filtered out or marked in the data time series, then the result may be bad for the wrong reason. Trustworthiness in data can only be a result of control and maintenance of both the hardware and the corresponding software and data archiving. The following sections outline the most important parts of a quality control that should be carried out regularly in real-time environments and prior to verification or evaluation exercises.

Key Points:
For relevant evaluation results, the data has to be of high quality, and faulty or corrupt data has to be detected, flagged and disregarded for the evaluation process. A detailed description of quality assurance and control processes can be found in chapter 5 of "Meteorological and Power Data Requirements for Eeal-time Forecasting Applications", which is Part 4 of this Recommended Practice Series.

Chapter 4

Assessment of Forecast Performance

Key Points

- *All performance evaluations of potential or ongoing forecast solutions have a degree of uncertainty*
- *The uncertainty is associated with three attributes of the performance evaluation process: (1) representativeness, (2) significance and (3) relevance*
- *A carefully designed and implemented evaluation process that considers the key issues in each of these three attributes can minimize the uncertainty and yield the most meaningful results*
- *A disregard of these issues is likely to lead to uncertainty and/or decisions based on unrepresentative information*

The relevance of different aspects of forecast performance depends on the user's application. For instance, one user may be concerned with the size of *typical* forecast errors, while another may only be concerned with the size and frequency of particularly large errors. There are a wide range of error metrics and verification methods available to forecast users, but their relationship to different attributes is not always clear. This chapter deals with the issues around evaluating specific attributes of forecast performance, including metric selection and the verification and the use of some specific metrics in forecast optimization.

4.1 Forecast Attributes at Metric Selection

Forecast users may be interested in either a single attribute, or a range of forecast performance attributes. When evaluating forecasts to either track performance changes or discriminate between different forecasts, it is important to consider those attributes relevant to the forecasts

intended use. Where a forecast is used in multiple applications, there is no guarantee that these attributes will be aligned and it may be necessary to compromise or procure multiple forecast products. Selecting an appropriate metric, or set of metrics, is a key requirement in order to produce an evaluation of forecast performance that is relevant to the forecast's end use.

Quantitative evaluation methods are usually the core of the evaluation framework, since they allow an evaluator to objectively rank different forecast solutions. Typical choices of quantitative metrics are the (root) mean squared error, the mean absolute error or the quantile score (see [10] for details) for continuous forecasts and various quantities derived from contingency tables for categorical or binary forecasts.

As emphasized in Section 5.1.5, the selection of metrics should be informed by the forecast user's intended use, and if a forecast is intended to be used for multiple applications, different basic metrics may be applied and merged into a weighted sum. Below, a range of forecast attributes and their relation to different evaluation metrics are discussed.

4.1.1 Typical Error Metrics

The most common error metrics used in renewable energy applications summarise 'typical' errors by averaging the absolute value of errors, or squared errors, often normalized by installed capacity. Such metrics are simple to produce and give a high-level view of forecast performance. They give equal weighting to all errors included, which may be appropriate if the forecast is used to inform decisions at any time, as opposed to only when a particular event is predicted.

In energy trading, for example, the forecast is used to inform decisions for every trading period and the cost implication of a forecast error is usually proportional to the error. In this case, the absolute value of the error is directly related to the forecast's end-use, so mean squared error would not be as informative as mean absolute error.

However, average error metrics hide some information which may be of interest. For example, a forecast with mostly small errors and occasional large errors could return a similar mean score to one with all moderate errors. In some cases this may not be an issue, but some users may prefer to experience fewer large errors even if that also means fewer small errors.

Examples of typical error metrics are discussed in section 5.1 and especially in section 5.1.2.

4.1.2 Outlier/Extreme Error

Another important attribute is the prevalence of large errors. Some applications are primarily sensitive to large errors, such as managing reserve energy or other risk management. Calculating metrics based on large errors is more challenging than for 'typical' errors, as large errors are more effected by specific situations. It is recommended that different root causes of large errors are considered separately, and that positive and negative errors are treated separately.

For example, large errors at a single wind farm during a period of high wind speed may be caused by high speed shut down, but are unlikely if the wind speed is only just above rated. If considering aggregated production from multiple wind farms, large errors may be caused by wind speed forecast errors in the vicinity of large areas of concentrated capacity.

4.1.3 Empirical Error Distribution

The empirical distribution of past forecast errors gives a detailed picture of how frequent errors of different sizes have been. It can be useful to examine the distribution of errors for specific situations, such as when power was forecast to be 70±2%, as the shape of the distribution will depend on power level, particularly for individual wind farms.

4.1.4 Binary or Multi-criteria Events

Some attributes of forecast performance relate to the prediction of events such as ramps (or particular rate and duration) which may span multiple lead-times and spatial scales. Furthermore, events typically have multiple attributes, such as timing and magnitude. Different attributes may be of more or less interest depending on the use case for the forecast. In these cases, average error metrics may not be representative of the desired forecast attribute.

For example, ramp rate may be of most importance to one user, whereas the timing or ramp magnitude may be of more importance to another. This effect is illustrated in Figure 4.1. Timing or *phase* errors are penalized heavily by mean absolute error so the forecast which best predicted both the ramp rate and magnitude appears worse by this measure. A similar principal applies to events such as the duration of high or low power periods. In general, average error metrics favour 'smooth' forecasts rather than those which capture the precise shape of specific events.

Contingency tables provide a framework for quantifying the prediction of categorical events, which can be defined to match the user's decision making process. For example, the user may define a particular ramp event with some tolerance for phase and amplitude error and then evaluate the performance of a particular forecast solution at predicting such events. There are four possibilities for each predicted and/or actual event: a true positive (hit), true negative (correct negative), false positive (false alarm) or false negative (miss). From these, a range of metrics can be calculated and used for comparison with other forecast systems. Furthermore, if the cost implications of decisions based on the forecast are known (or can be estimated) then the relative value of forecasting systems may be calculated.

Examples on how to verify outliers can be found in section 5.1, and 5.5.2.1.

4.2 Prediction Intervals and Predictive Distributions

Prediction intervals may be supplied to provide situational awareness or to information or quantitative risk management. These intervals predict an upper and lower bound which the observation will fall between with some probability. It is therefore an important attribute that

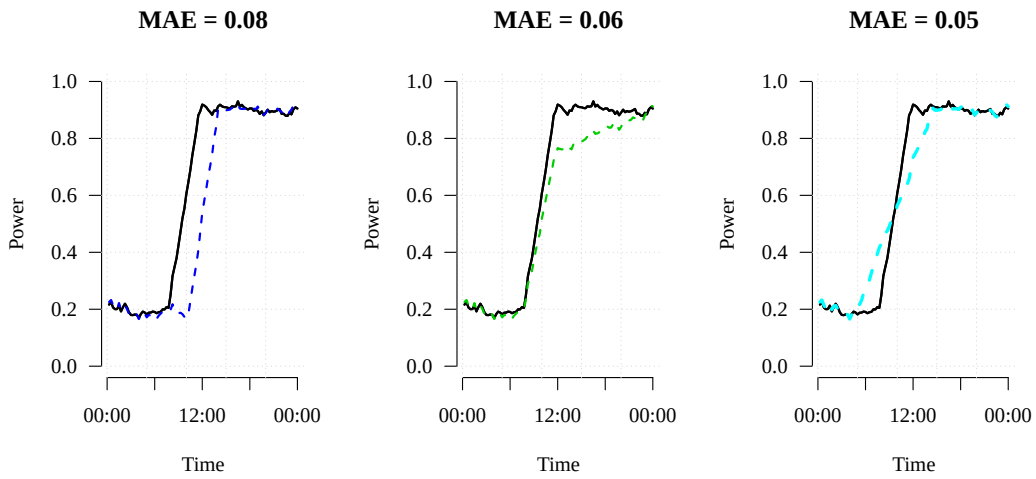

Figure 4.1: Examples of different types of ramp forecast error. Actual power is shown as solid black lines, forecasts are colored dashed lines. From left to right: phase or timing error, amplitude error and ramp rate error. The mean absolute error (MAE) for each forecast is shown above the plots. Despite being the only forecast the correctly predict the ramp rate and duration, the forecast with a phase error has the largest MAE.

observations do in fact fall between the interval with the prescribed frequency. This property is call 'reliability' and can by evaluated by simply counting the frequency of observations within and outside the interval. A more accurate forecasts with a narrower interval is said to be 'sharp' and provides greater confidence than a wide interval, but must be reliable in order to inform risk-based decision making. Therefore, prediction intervals should be evaluated following the principal of *sharpness subject to reliability*.

A predictive distribution is a smooth probability density function for the future value. It provides full information about probability of all possible value ranges rather than a single interval. In this case the principal of *sharpness subject to reliability* still applies, but sharpness and reliability needs to be evaluated for a range of probability levels.

In quantitative decision-making under uncertainty, the optimal decision is often a *quantile*, i.e. the value that is forecast to be exceeded with some probability. For example, if the cost of taking precautionary action is C to protect against an uncertain adverse effect with potential loss L, then the precautionary action should be take in the probability of the adverse effect happening is greater than the cost-loss ratio C/L.

In applications of wind power forecasting, the adverse event could be exposure to imbalance costs, or holding insufficient energy reserves. In most cases, the values of C and L will be changing continuously and the decision maker will be aiming to select a future value of energy production which will be achieved with some probability $p = C/L$. Therefore, it is necessary to have access to the full predictive distribution in order to make an appropriate decision. Where the cost-loss ratio is known, the relative economic value of different forecasting systems can be calculated.

4.3 Probabilistic Forecast Assessment Methods

Probabilistic forecast evaluation is a complex topic. There are a number of classical metrics, just like for deterministic forecasts. However, the evaluation of probabilistic forecasts places greater importance on an end-user's knowledge of a cost function that provides a good indication of how well the forecast performance has met the requirements of the user's application. (see 5.1.5).

The considerations from chapter 2 on the performance evaluation and its inherent uncertainty are even more important here. The three attributes *(1) representativeness, (2) significance* and *(3) relevance* are equally important to consider when setting up evaluation of probabilistic forecasts.

In the same cases, it might be best to only use a graphical inspection of how well observations lie within forecast intervals. This can then be extended to an interval evaluation to provide objective values to the visual impression from the graph. This is similar to the "dichotomous event evaluation" described in 5.1.3.1 for predefined events. These scores can also be used for probabilistic/uncertainty forecasts, if the application is about how well the probabilistic forecasts or forecast intervals have captured the observations.

It is therefore important to follow the recommendations in the "best practice recommendations" 5 on how to build up an evaluation platform that reflects the purpose of the forecasts and provides an incentive to the forecast provider to match these criteria with the appropriate methods.

Therefore, the following description of metrics only provide a set of possible tools that can be used for the evaluation of probabilistic forecasts and the user must select the most appropriate set depending on the characteristics of the user's application and objectives of the forecast evaluation.

4.3.1 Brier Scores

The Brier score [12] is probably the most prominent and widely-used probabilistic forecast performance metric. It is a useful measure for a general assessment of the performance of probabilistic forecasts. However, the formulation of the basic Brier makes it suitable only for the evaluation of probabilistic forecasts of binary events (i.e. occurrence or non-occurrence of a defined event)

The Brier Score (BS) is the equivalent of the mean-squared error (MSE) for probabilistic forecasts with the same limitations as for deterministic forecasts. That means, the Brier Score is sensitive to the climatological frequency of events in the sense that the rarer an event, the easier it is to get a good BS without having any real skill. The BS is defined as,

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{4.1}$$

where p_i is the *forecast probability* at time i, o is the observation at time i, and N is the

number of forecasts. The forecast probabilities range in value between 0 and 1, the observed values are either 0, if the event does not occur, or 1, if the event occurs.
Equation 4.1 is bound mathematically to values between 0 and 1. A lower Brier score, similar to the MSE, indicates greater accuracy. The maximum squared error is 1, because all squared errors will lie between 0 and 1. A *perfect accuracy* is reflected in the Brier score with 0, i.e. there is no difference between scores of an event and someone's probabilistic predictions for those events. The opposite, i.e. a Brier score of 1, reflects perfect inaccuracy, which means that there are probabilities of 0 given to events that occur and probabilities of 1 to events that do not occur.

In order to gain further insight into the behaviour of the Brier score, it can be decomposed algebraically into three components:

$$BS = CAL\breve{} - RES + UNC \tag{4.2}$$

where, the CAL is the Calibration and is also sometimes referred to as the "reliability" , RES is the resolution and UNC is the uncertainty.

In [8], these three components are explained in a way that is easy to understand and relate to applications:

- *CAL* is a squared function of forecasted probability (f_p) and the mean probability ($\overline{x_p}$) and measures whether the forecasted values consistently represent the frequencies with which events occur (i.e., is the forecasted probability too large or too small on average?). For example, does the event occur 30% of the time when a forecast of 0.30 is issued? Specifically, CAL measures the difference between the actual frequency of occurrence and the forecast prediction. This is also referred to as the "reliability" of a probabilistic forecast.

- *RES* is a squared function of ($\overline{x_p}$) and ($\overline{x}$) and measures how much the frequency of event occurrence varies among the forecasts. It measures the ability of the forecast to distinguish between event and non-event. For example, if the average frequency of event occurrence across all forecasts is 0.50, the relative frequency of occurrence ($\overline{x_p}$) should be much smaller for events, when the forecast is 0.10 (low likelihood of event) and much larger when the forecast probability is 0.90 (high likelihood of event). Higher RES scores indicate more skill and therefore appears in equation (4.1) with a negative sign. In the worst case, when the same probability (for example, the climatological probability) is always forecasted, the resolution is zero.

- *UNC* is a function of ($\overline{x}$) only and does not specifically measure how well the forecasts predict the event. Instead, UNC is an important measure of the difficulty of the forecasting situation. Large values of UNC (e.g., when the event is very rare) indicate that the forecasting situation is more difficult. It is inappropriate to compare forecasts for systems with significantly different UNC values.

In [8] a very useful list of specific questions that the Brier score answers have been listed:

1. ***Brier Score (BS)*** answers how accurate the probability forecasts are
2. ***Calibration (CAL)*** answers how well does the conditional relative frequency of occurrence of the event match a situation?
3. ***Resolution (RES)*** answers how well does the forecast separate events according to whether they occur or don't occur
4. ***Uncertainty UNC)*** answers how difficult/uncertain is the forecast situation

4.3.2 Ranked Probability (Skill) Score (RP(S)S)

The Ranked Probability Score (RPS) and Ranked probability Skill Score [13] (RPSS) is widely used for multi-category probability forecasts that have a magnitude order to them (such as generation forecasts). The RPS is the multi-category extension of the Brier score, and the "Skill" part refers to a comparison of the RPS of a specified forecast to the RPS of a reference forecast (such as climatological probabilities).

In other words, the RPS measures cumulative, squared error between categorical forecast probabilities and the observed categorical probabilities, and the RPSS measures the error relative to a reference (or standard baseline) forecast (climatology, persistence or other reference forecast). The observed categorical probabilities are 100% in the observed category, and 0% in all other categories [13].

$$RPS = \sum_{cat=1}^{Ncat} (P_{cumF(cat)} - P_{cumO(cat)} \tag{4.3}$$

Where $Ncat = 3$ for tercile forecasts. The "cum" implies that the summation is done first for cat 1, then cat 1 and 2, then cat 1 and 2 and 3 [13].

The higher the RPS, the poorer the forecast. RPS=0 means that the probability given to the category that was observed was 100%. The RPSS is based on the RPS for the forecast compared to the RPS For a reference forecast such as one that simply gives climatological probabilities.

RPSS > 0 when RPS for actual forecast is smaller (i.e. better) than RPS for the reference forecast.

$$RPSS = 1 - -\frac{RPS_{forecast}}{RPS_{observation}} \tag{4.4}$$

The RPSS is made worse by three main factors [14]:

(1) Mean probability biases

(2) Conditional probability biases (including amplitude biases)

(3) The lack of correlation between forecast probabilities and observed outcomes

(1) and (2) are calibration factors and (3) involves discrimination. The tercile category system can be seen as a two category system if the two tercile boundaries are considered one at a time: below normal vs. not below normal above normal vs. not below normal.

4.3.2.1 The Continuous Ranked Probability Skill and Energy Score

As described above in 4.3.1, the Brier Score (BS) is useful for binary events (e.g. critical ramp/not critical ramp). When analysing discrete multiple-category events (e.g. below critical/critical/above critical ramping) the Ranked Probability Score (RPS) (see 4.3.2) is preferably used. In the continuous case, where there are an infinite number of predictand classes of infinitesimal width, the RPS is extended to the Continuous Ranked Probability Score (CRPS) [6]. Alternatively, it can be interpreted as the integral of the Brier score over all possible threshold values for the parameter under consideration. For a deterministic forecast system, the CRPS reduces to the mean absolute error [1].

For an ensemble prediction system, the CRPS can be decomposed into a reliability part and a resolution/uncertainty part, in a way that is similar to the decomposition of the Brier score. The reliability part of the CRPS is closely connected to the rank histogram of the ensemble (see section 4.3.3.1), while the resolution/uncertainty part can be related to the average spread within the ensemble and the behaviour of its outliers[1].
In [1] it is noted that the definition of CRPS makes it well suited to explain the relationship between the Brier Score and the usefulness of CRPS for comparisons between deterministic and probabilistic forecasts and is also decomposed for the use of ensemble prediction systems. Hersbach [1] defined the relationships in the following way:
if we consider the parameter of interest being denoted by x. For instance, x could be the 100-m wind speed or power output. Suppose that the forecast by an ensemble system is given by $r(x)$ and that x_a is the value that actually occurred. Then the continuous ranked probability score expressing some kind of distance between the probabilistic forecast r and truth x_a, is defined as

$$CRPS = CRPS(P,xa) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 \, dx \tag{4.5}$$

Here, P and P_a are cumulative distributions:

$$P(x) = \int_{-\infty}^{\infty} P_y(y) \, dy \tag{4.6}$$

and

$$P_a(x) = H(x - x_a), \tag{4.7}$$

where is the well-known Heaviside function, defined to

$$H(x) = \begin{cases} 0 & \text{for x < 0} \\ 1 & \text{for } x \geq 0 \end{cases} \tag{4.8}$$

So, $P(x)$ is the forecasted probability that x_a will be smaller than x. Obviously, for any cumulative distribution, $P(x) \in [0,1]$ $P(-\infty) = 0$, and $P(\infty) = 1$. This is also true for parameters that are only defined on a subdomain of R. In that case $r(x) = 0$, P is constant

outside the domain of definition. The CRPS measures the difference between the predicted and occurred cumulative distributions.
Its minimal value of zero is only achieved for $P = P_a$, that is, in the case of a perfect deterministic forecast.
In this definition, the CRPS has the dimension of the parameter x (which enters via the integration over dx). In practice the CRPS is averaged over an area and a number of cases:

$$CRPS = \sum_k CRPS(P_k, x_k^a) \tag{4.9}$$

where k labels the considered grid points and cases.
It is here, where the CRPS can be seen as the limit of a ranked probability score with an infinite number of classes, each with zero width and it is not difficult to see that if $k = i$ and $p_k = P_k(x_t)$ and $O_k = P_k^a(x_t)$ the CRPS is directly connected to the Brier score by

$$CRPS = \int_{-\infty}^{\infty} BS(x_t)\, dx \tag{4.10}$$

For a deterministic forecast, that is, $x = x_d$ without any specified uncertainty, $P(x) = H(x^2\ x\ d$. In that case, the integrand of Eq.4.5 is either zero or one. The non-zero contributions are found in the region where $P(x)$ and $P_a(x)$ differ, which is the interval between x_d and x_a . As a result,

$$\overline{CRPS} = \sum_k |x_d^a - x_a^b| \tag{4.11}$$

which is the mean absolute error (MAE).
Although the CRPS is widely used (see e.g. [3, 5, 11]), in most real-world applications, it is no longer appropriate, when assessing multivariate forecasts[1]. To assess probabilistic forecasts of a multivariate quantity, Gneiting et al.[5] therefore proposes the use of the so called "energy score", which is a direct generalization of the CPRS 4.5, to which the energy score reduces in dimension $d = 1$. Gneiting and Raftery [15] showed its propriety and noted a number of generalizations. If $P = P_{ens}$ is an ensemble forecast of size m, the evaluation of the energy score is straightforward, i.e., the predictive distribution P_{ens} places point mass 1/m on the ensemble members $x1, ..., x_m \in {}_d$, and the "energy score" is defined as:

$$es(P_{ens}, x) = \frac{1}{m}\sum_{j=1}^{m} \| x_j x \| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} m \| x_i - x_j \| \tag{4.12}$$

If $P = \delta\mu$ is the point measure in $\mu \in {}_d$, that is, a deterministic forecast, the energy score reduces to

$$es(\delta\mu, x) = \| \mu - x \| \tag{4.13}$$

[1]A forecast is multivariate, when it consists of multiple variables, which typically refer to multiple time-steps, multiple sites or multiple parameters

Thus, the energy score provides a direct way of comparing deterministic forecasts, discrete ensemble forecasts and density forecasts using a single metric that is proper[2] If such closed form expressions for the expectations in 4.13 are unavailable, as is often the case for density forecasts, it is recommended to employ Monte Carlo methods [5].

4.3.2.2 Logarithmic and Variogram Scoring Rules

In addition to CRPS, other recently investigated scoring rules for probabilistic forecasts in the energy sector are the Logarithmic Score (LogS) and Variogram Score (VarS). Details of these scores, examples of basic calculations and suggestions for software implementations can be found in Bjerregård, Møller and Madsen [11], where the authors also distinguish forecasting evaluation for energy systems between univariate and multivariate forecasts.
A forecast is considered multivariate, when it consists of multiple variables, which typically refer to multiple time-steps, multiple sites or multiple parameters. Obviously, for multivariate and multi time-step forecasts the errors typically show some inertia or auto-correlation[3], and a proper modelling of these auto-correlation is important for many applications, such as use of battery facilities associated with wind farms.

Also with these more sophisticated metrics, it has become clear and can be considered a common understanding [5, 11] that no scoring rule performs optimally in all aspects. A proper evaluation of multivariate forecasts is mainly of interest when the auto-correlation structure of the forecast is assumed to be important, i.e. a high degree of similarity between a given time series and a lagged version of itself over successive time intervals. Such multivariate forecasts can be evaluated using the VarS score [11]. However, in energy systems it is often crucial to have well-calibrated univariate forecasts, and these have to be evaluated by applying the univariate LogS or CRPS score.

4.3.3 Reliability Measures

There are a number of reliability measures that measure or depict the same attribute: the agreement between forecasted probabilities and observed frequencies.
The differences and similarities of the various measures, rank histogram, reliability diagrams and calibration diagrams are explained and discussed in the following sections, so that the use and benefits of combining some of these measures become clear.

It is also worth noting that the *CAL* term in the Brier Score (BS) is basically a quantification of what is seen in these diagrams.

[2]defined here according to [5]: A proper scoring rule is designed such that it does not provide any incentive to the forecaster to digress from her true beliefs of the forecaster's best judgement.

[3]representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals

4.3.3.1 Rank Histogram

Rank histograms measure the consistency and reliability and assumes that the observation is statistically indistinguishable from the ensemble members.

The rank histograms are developed by ranking the N ensemble members from lowest to highest and identifying the rank of observation with respect to the forecasts. Figure 4.3 show typical distributions and their characteristics with respect to their skill.

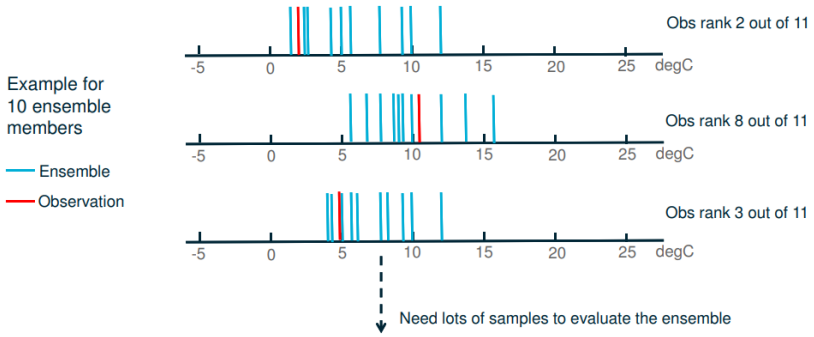

Figure 4.2: One rank histograms ©[14]

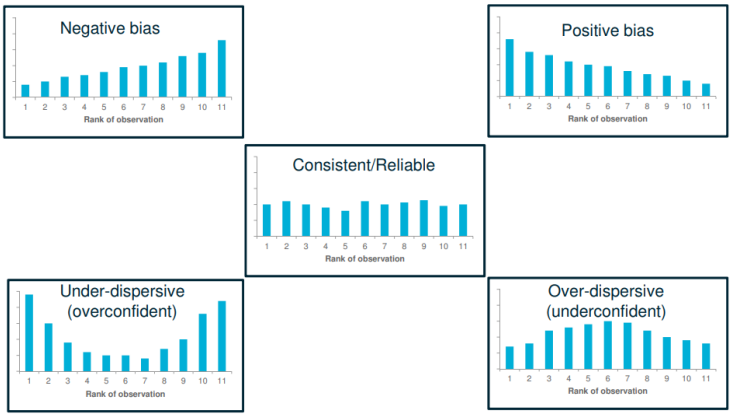

Figure 4.3: Examples of a rank histograms ©[14]

It is important to note that the flat rank histogram does not necessarily indicate a skillful forecast. Rank histograms show conditional/unconditional biases, but does not necessarily provide a full picture of the skill, because it[14]:

- only measures whether the observed probability distribution is well represented by the ensemble
- does NOT show sharpness – for example, climatological forecasts are perfectly consistent (flat rank histogram) but not useful

•

4.3.3.2 Reliability (Calibration) Diagram

The reliability (calibration) diagrams tell how well predicted probabilities of an event correspond to their observed frequencies and provides insight into how well calibrated a probabilistic forecast is and is a complementary metric to the Brier scores (4.3.1 and the Relative Operating Characteristics (ROC) curve (4.3.4).
The reliability diagram plots the observed frequency against the forecast probability, where the range of forecast probabilities is divided into K bins (for example, 0-5%, 5-15%, 15-25%, etc.). The sample size in each bin is often included as a histogram or values beside the data points [16].

The characteristics of the reliability is indicated by the proximity of the plotted curve to the diagonal. The deviation from the diagonal gives the conditional bias. If the curve lies below the line, this indicates over-forecasting (probabilities too high); points above the line indicate under-forecasting (probabilities too low). The flatter the curve in the reliability diagram, the less resolution it has. A forecast of climatology does not discriminate at all between events and non-events, and thus has no resolution. Points between the "no skill" line and the diagonal contribute positively to the Brier skill score. The frequency of forecasts in each probability bin (shown in the histogram) shows the sharpness of the forecast [16]. Figure 4.4 show this principle and Figure 4.5 show typical examples of reliability diagrams for various forecast flaws.

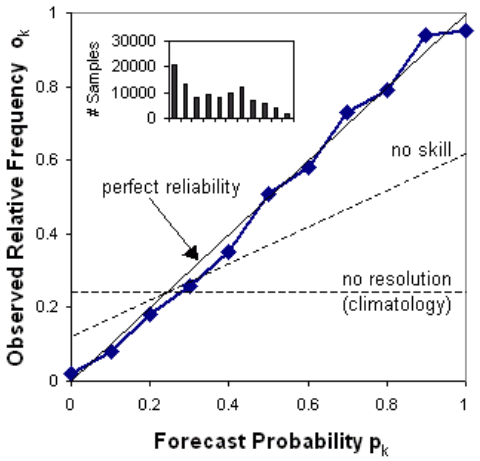

Figure 4.4: Connection between rank histograms and reliability diagrams ©[16]

The reliability diagram is conditioned on the forecasts (i.e., given that an event was predicted, what was the outcome?), and can be expected to give information on the real meaning of the forecast. It is a good partner to the ROC, which is conditioned on the

observations. Some users may find a reliability table (table of observed relative frequency associated with each forecast probability) easier to understand than a reliability diagram.

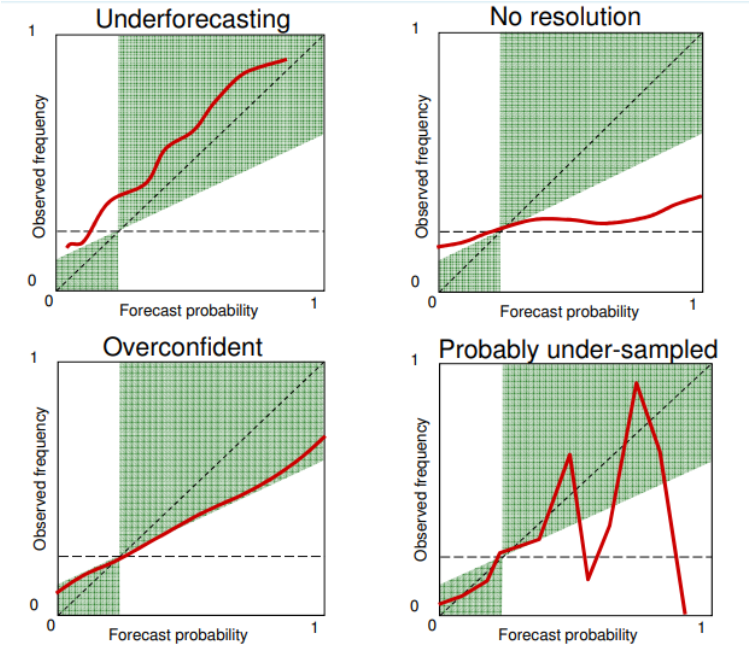

Figure 4.5: Examples of reliability diagrams. The left upper and lower figure correspond to the histograms for over- and under-dispersive distributions in Figure 4.3. ©D. Hudsen, "Ensemble Verification Metrics" Presentation at ECMWF Annual Seminar[14]

4.3.4 Event Discrimination Ability: Relative Operating Characteristic (ROC)

This metric shows a probabilistic forecast's ability to predict the occurrence of events and non-occurrence of non-events.

In the *ROC* diagram the performance of forecasts at different probability thresholds is visualised. One important aspect of the ROC is that it ignores calibration of the forecasts. That is, a poorly calibrated forecast will not be penalized by the ROC. Thus, it is important to pair the ROC evaluation with an evaluation of forecast calibration, such as the calibration diagram, which is discussed in the previous section.

The ROC is based on computing two categorical statistics (see 5.1.3.1):

1. the Probability of Detection (POD), Hit Rate (HR) or true positive rate (TPR)
2. the False Alarm Rate (FAR) or False Positive Rate (FPR)

The ROC curve is created by plotting the true positive rate (TPR) or the probability of detection (POD) against the false positive rate (FAR) or false alarm rate (FAR) at various

thresholds. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as probability of false alarm and can be calculated as (1 *specificity*) [17].

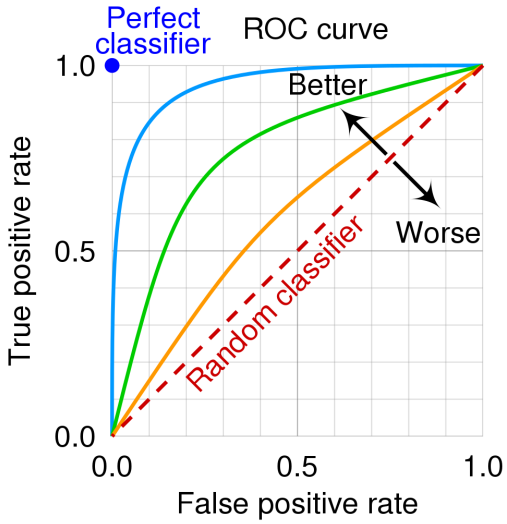

Figure 4.6: Example of a "Relative Operating Curve" (ROC) curve ©Wikipedia [17]

Figure 4.6 depicts examples of ROC curves. The orange line represents a forecasting system with little skill, the green with moderate (better) skill and the blue line a forecasting system with reasonable skill.
As shown in Figure 4.6, when the ROC curve falls below the diagonal line the forecasts are random classifiers, or in other words have no skill according to this metric. The blue line shows a good, or better forecast skill, where the curve is pushed up towards in the upper left corner (TPR = 1.0). The area under the ROC curve provides a useful measure of forecast skill[17]).

It can also be thought of as a plot of the power as a function of the Type I Error of the decision rule (when the performance is calculated from just a sample of the population, it can be thought of as estimators of these quantities).
The ROC curve is thus the sensitivity or recall as a function of fall-out.
In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function (area under the probability distribution from $-\infty$ to the discrimination threshold) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability on the x-axis[17].

4.3.5 Uncertainty in Forecasts: Rény Entropy

General forecast metrics such as MAE and RMSE do not measure the uncertainty of the forecast and are only considered unbiased, if the error distribution is Gaussian, which is seldom the case. In order to define this, and compare it with uncertainty forecasts, it is recommended to use the Rény entropy, defined as the variation of wind or solar forecast errors in a specified time period [8] (chapter 6).
The Rényi entropy is defined as:

$$H_\alpha(X) = \frac{1}{1-\alpha} log_2 \sum_{i=1}^{n} p_i^\alpha$$

where α (where $\alpha > 0$ and $\alpha \neq 1$) is the order of the Rényi entropy, which enables the creation of a spectrum of Rényi entropies with p_i being the probability density of the i discrete section of the distribution. Large values of α favour higher probability events, while smaller values of α weigh all instances more evenly. The value of α is specified by the metric user.

4.4 Metric-based Forecast Optimization

Once the most important attributes of a forecasting system and an evaluation metric or matrix has been decided, it may be possible to optimize the forecasting system to have desirable properties. Many forecasting solutions are tuned/optimized for specific performance criteria either at the post-processing stage (conversion of weather forecasts to power forecasts) or even in the numerical weather models themselves. For example, many statistical post-processing techniques allow the user to specify whether to minimize (root) mean squared error or mean absolute error. The former is implicit in *ordinary lest squares*, a widely used method for estimating the parameters of linear models or methods that are based on maximum likelihood estimation assuming Gaussian (or 'Normally') distributed errors. The latter has no closed form solution for estimating linear models so requires the application of numerical methods to solve.

It is recommended that the desired properties of a forecasting solution be considered from the outset and communicated to those responsible for the solution's development and implementation.

Chapter 5

Best Practice Recommendations

Key Points
The recommendations in this section are based on the idea that the verification framework or scoring rules chosen to evaluate forecasts shall assist end-users and forecast providers in the determination of which aspects of the forecast should be the focus of forecast improvement efforts. To achieve this, the following set of principles shall be considered:

- ***Verification is subjective***
  it is important to understand the limitations of a chosen metric
- ***Every verification has an inherent uncertainty***
  due to its dependence on the evaluation data set
- ***Evaluation should contain a set of metrics or scoring rules (framework)***
  in order to measure a range of forecast performance attributes
- ***Evaluation should reflect a "cost function"***
  i.e. the metric combinations should provide an estimate of the value of the solution

In this last chapter, the principles developed in the previous chapters are brought to the application level. In other words, the somewhat theoretical considerations from the previous chapters are now applied to real-world problems. In the second chapter 2, the concept of forecast evaluation uncertainty was introduced with the three attributes "representative", "significant" and "relevant" to help minimize this type of uncertainty in the evaluation. The following chapter 3, introduced the concept of measurement uncertainty with the associated uncertainty in the evaluation process and how to minimize the errors in the evaluation due to this type of uncertainty. In the previous chapter 4 the performance assessment was described

in general terms and with examples that are relevant for all types of evaluation in the power sector.

5.1 Developing an Evaluation framework

> ***Key Points***
> *The construction of a comprehensive evaluation framework is an alternative to a one-metric forecast evaluation approach and can be an effective way to mitigate the "relevance" issues associated with the tuning (optimization) of forecasts to target metrics that are not optimal indicators of value for an end user's application.*

The "typical forecasting task" is defined in this context as forecasts generated to fulfill operational obligations in electric system operation, trading and balancing of renewable energy in power markets. There are certainly many other tasks and applications of weather and power forecasts in the energy industry that can also benefit from the following best practice recommendations. However, the primary target for the following recommendations is the evaluation of forecasts for these particular applications. In this section we define the evaluation framework and its components and considerations. Section 5.2 deals with the evaluation to maximize value from operational forecasts, section 5.3 with the evaluation of trials and benchmarks and in the use cases section 5.5 there are example evaluations for energy trading and balancing, power ramps and reserve allocation.

5.1.1 Scoring Rules for comparison of Forecast Types

Scoring rules can be defined as summary measures in the evaluation of deterministic or probabilistic forecasts, by which a numerical score based on the predictive distribution and the event or value that materializes (i.e. the outcome) is assigned [5]. In this sense, scoring rules are negatively oriented penalties that a forecaster wishes to minimize and are often also referred to as "loss functions".

A scoring rule also needs to be proper which means that a forecaster maximizes the expected score for an observation drawn from the distribution F, if the forecaster issues the probabilistic forecast F, rather than $G \neq F$. In prediction problems, proper scoring rules encourage the forecaster to make careful assessments and to be honest. In estimation problems, strictly proper scoring rules provide attractive loss and utility functions that can be tailored to the problem at hand [15].

In recent years, where the possibilities for more profound evaluations, also due to more available open source software, have become possible, a common understanding has been established (e.g. [5, 6, 7, 11, 15]) that no scoring rule performs optimally in all aspects. The end-user therefore needs to determine which aspects of the forecast should be the focus of forecast improvement efforts over time.

If there are more than one aspect to consider, or even contradicting aspects due to the use of one forecast, that is used for different applications, a framework will (1) assist in identifying contradicting forecast performance objectives and (2) allow giving weight to different aspects of the forecast for an overall evaluation.

With the energy score described in section 4.3.5, it is also possible to directly compare deterministic forecasts, discrete ensemble forecasts and density forecasts using a single metric that is proper.

5.1.2 Analyses of Forecasts and Forecast errors

In this discussion, forecast errors are defined as forecast minus observation ($fc - obs$). Errors in forecasting are inevitable. The primary objective is, of course, to minimize the magnitude of the error. However, a secondary objective may be to shape the error distribution in ways that are beneficial to a specific application. A direct and deep analysis of the prediction errors can provide considerable insight into the characteristics of forecast performance as well as information that can allow users to differentiate situations in which forecasts are likely to be trustworthy from those that are likely to produce large errors.

The construction of a frequency distribution of errors (also referred to as density functions or probability density functions) is an effective way to obtain insight about forecast error patterns. These are created by sorting errors and visualizing their distribution as e.g.,

- (probability) density curve
- histogram (frequency bars)
- box plot

All of these chart types show the same basic information but with different degrees of detail. Density curves provide the most detail since they depict the full probability density function of the forecast errors. Histograms provide an intermediate level of detail by showing the frequency of a specified number of error categories. Box plots condense this information into several quantiles (see 5.1.3.2). Errors of a well calibrated forecast model should always be scattered around zero. A frequency distribution that has a center, that is shifted from zero indicates a systematic error (also known as a bias).

For power forecasts one will often see positively skewed error distributions, which are due to the shape of the power curve which has flat parts below the cut-in wind speed and at wind speeds that produce the rated power production. The skewed distribution is often the result of the fact that forecasts close to zero cannot have large negative errors. The inverse is true for forecasts of near rated power (i.e. large positive errors cannot occur) but forecasts of rated power are often less frequent than near zero forecasts and hence have less impact on the error distribution.

5.1.3 Choice of Deterministic Verification methods

There is not a single best metric that can be effectively used for all applications. The definition of "best metric" highly depends on the user's intended application and should be based on a quantification of the sensitivity of a user's application to forecast error. For example, if a user has to pay a penalty for forecast errors that are proportional to the squared error, a mean squared error metric is well suited for evaluation.

However, if the penalty is proportional to the absolute error, a mean absolute error metric would be a better choice. If the user is interested in predictions of specific events such as high wind shutdown or large wind ramps, the mean squared or absolute error metrics are not good choices, because they do not provide any information about the ability of a forecast to predict these events due to their averaging characteristics. In this case, an event-based metric should be employed. An example of this type of metric is the critical success index (CSI), which measures the ratio of correct event forecasts to the total number of forecasted and observed events.

5.1.3.0.1 "Loss function:" In order to get forecast performance information that is relevant for a user's application, it is crucial to carefully select the evaluation metrics and ideally they should be based on the so-called "loss function" for the user's application. The "loss function" is also often referred to as a "cost function", especially when related to costs that can be associated with specific forecast errors. Conceptually, a well-formulated "loss" or "cost" function measures the sensitivity of a user's application to forecast error. If one forecast is used for different applications with different loss functions, a set of metrics should be derived. If a single metric is desired, then a composite metric can be constructed by weighting the individual application-based metrics by the relative importance. More details on how to develop such loss functions and evaluation matrices can be found in 5.1.5 .

5.1.3.1 Dichotomous Event Evaluation

One may quantify desirable qualities of a forecast by considering a range of of dichotomous (yes/no) events, such as high-speed shut-down or ramps. A forecast might imply that "yes, a large ramp will happen" and trigger the user to take action, but the ability of a forecasting system to make such predictions is not clear from the average error metrics. Therefore, one should employ a quantitative verification approach to assess this ability by analysing the number of correct positive, false positive, correct negative and false negative predictions of particular events [18], [9]. Table 5.1 provides an example table to carry out such categorical evaluations.

Recommendation for applications with (Extreme) Event Analyses:
Categorical statistics that can be computed from such a yes/no contingency table. The list below is an excerpt of a comprehensive list of categorical statistical tests published by the Joint World Weather Research Program (WWRP) and Working Group Numerical Experimentation on Forecast Verification (WGNE) and provides the most commonly used

Table 5.1: Example of a dichotomous evaluation table

		Observations	
		YES	NO
Fore-cast	YES	a correct event forecast	b false alarm
	NO	c surprise events	d no events

metrics and their characteristics that are relevant for forecast applications in the power industry. Details, equations and a more comprehensive explanation on the use of these as well as references can be found (online) in [9]. It is recommended to apply these categorical statistics in particular for applications, where standard "typical error" metrics do not provide a measure of the true skill of a forecast to predict a specific event. In renewable power forecasting applications this is particularly important for extreme event analysis, ramping and high-speed wind turbine shutdown forecasting, etc. In such applications, it is important to distinguish between *quality of a forecast* (the degree of agreement between the forecasted and observed conditions according to some objective or subjective criteria) and *value of a forecast*(the degree to which the forecast information helps a user to achieve an application objective such as improved decision-making). Wilks [19] and Richardson [20] present concepts for the value versus skill for deterministic and probabilistic forecast evaluation of that type, respectively.

- **Accuracy**
  Answers the question: Overall, what fraction of the forecasts were correct?
  Range: 0 to 1. Perfect score: 1

- **Bias score**
  Answers the question: How did the forecast frequency of "yes" events compare to the observed frequency of "yes" events?
  Range: 0 to 1. Perfect score: 1

- **Probability of detection (POD)** Answers the question: What fraction of the observed "yes" events were correctly forecast?
  Range: 0 to 1. Perfect score: 1

- **False alarm ratio (FAR)**
  Answers the question: What fraction of the predicted "yes" events actually did not occur (i.e., were false alarms)?
  Range: 0 to 1. Perfect score: 0

- **Probability of false detection (POFD)**
  Answers the question: What fraction of the observed "no" events were incorrectly

forecast as "yes"?
Range: 0 to 1. Perfect score: 0

- **Success ratio**
  Answers the question: What fraction of the forecast "yes" events were correctly observed?
  Range: 0 to 1. Perfect score: 1

- **Relative value curve (versus skill)** for deterministic forecast
  Answers the question: For a cost/loss ratio C/L for taking action based on a forecast, what is the relative improvement in economic value between climatological and perfect information? Range: -1 to 1. Perfect score: 1.

5.1.3.2 Analysing Forecast Error Spread with Box and Wiskers Plots

The box-and-whiskers plot is a visualization tool to analyse forecast performance in terms of the error spread when comparing forecasts with different attributes such as forecast time horizons, vendors, methodologies. Figure 5.4 shows the principle of a box and whiskers plot. This type of charts can be used to illustrate the spread of forecast performance in each hour of the day-ahead horizon can be visualized. It can also show that some forecasts in some hours have very low errors compared to the average error in that hour, as well as occasionally very high errors. In section 5.4.2, a use case for the application of box plots is demonstrated to verify significance of results.

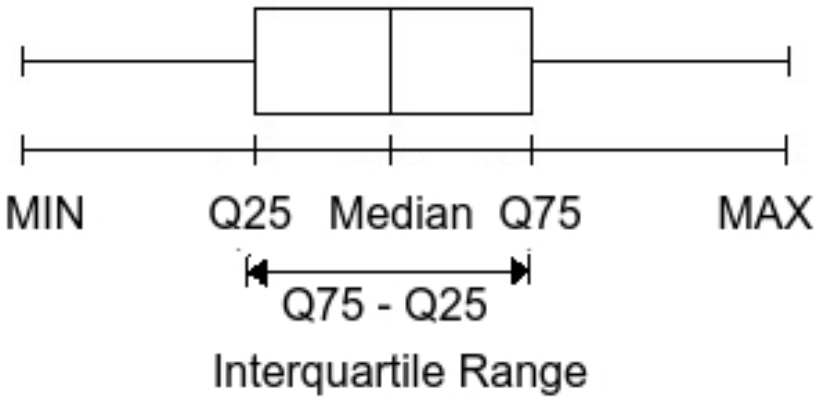

Figure 5.1: Principle of a box-and whiskers plot. The plot displays a five-number summary of a set of data, which is the minimum, first quartile, median, third quartile, and maximum. In a box plot, a box from the first quartile to the third quartile is drawn to indicate the interquartile range. A vertical line goes through the box at the median.

5.1.3.3 Visualising the error frequency distribution with histograms

Histograms allow one to (1) quantify the frequency of occurrence of errors below or above a specified level or (2) visualise the forecast error distribution for specified error ranges. In case (1) a graphical or tabular presentation can be directly used to derive a metric that indicates that errors are less than x% of the installed capacity in y% of the time. In this way, histograms function as a metric providing the percentage of time that errors are within a specified margin [[2]]. In case (2) the error distribution of a forecast can be derived from the graphical or tabular presentation of the histogram information. This enables an easy identification of the frequencies of large errors and provides the possibility to analyse and possibly modify the forecast system to minimize these errors. In summary, histograms visualize two main attributes:

- Robustness of a forecast
- Large Errors in an error distribution

In Madsen et al. [2] an example can be found for the way histograms help to interpret statistical results and error distributions. In their example, they directly determined that a 1 hour-ahead prediction contained errors less than 7.5% of the available capacity in 68% of the time, while a 24 hour-ahead prediction showed errors of that size only in 24% of the time. For large errors, they determined from the histogram that the same 1 hour-ahead prediction's largest errors were 17.5% of available capacity in only 3% of the time.

Figure 5.2 provides two example histograms with typical frequency distribution of errors for a 2-hour forecast horizon (left) and a day-ahead horizon (right) as described in [2].

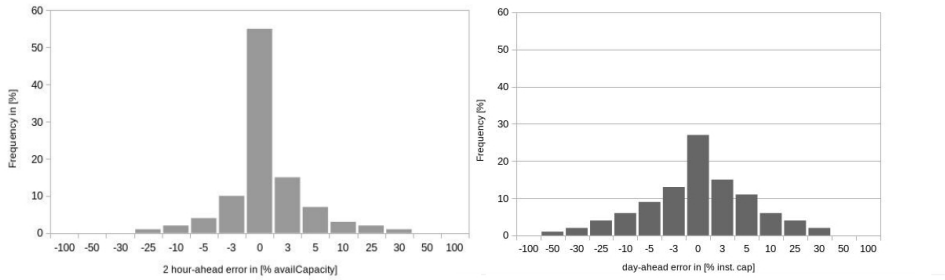

Figure 5.2: Examples of two histograms showing typical frequency distribution of errors for a 2-hour forecast horizon (left) and a day-ahead horizon (right).

Recommendation: If the application requires that specified error sizes should occur less than a certain, specified percentage of the time, a histogram analysis should be used to directly identify, whether or not a forecast's performance fulfills the specified criteria.

5.1.4 Specific Probabilistic Forecast Verification

As in the case of the verification of deterministic forecasts, it is recommended that multiple verification scores be employed for the evaluation of probabilistic forecasts. A well-chosen

set of probabilistic forecast evaluation metrics will provide an indication of several key aspects of forecast skill, and thus provide a more comprehensive representation of forecast performance than a single metric.
This also holds for recently investigated, more complex metrics such as the "Logarithmic Scores" (LogS) or "Variogram Scores" (VarS) (see section 4.3 and 4.3.2.2), when the so-called auto-correlation structure[1] of multivariate[2] forecasts is assumed to be important, or in energy systems, where it may be crucial to have well-calibrated univariate forecasts. These can then, for example, be evaluated by applying the univariate *LogS* or CRPS score to all the marginal densities, and depending on whether the shapes of the tails are considered important, a *LogS* could additionally be used or if not, a CRPS may additionally be more appropriate [11].
In the optimal case, the verification framework or scoring rules assist users and providers in the determination of which aspects of forecast should be the focus of forecast improvement efforts.

An evaluation of probabilistic forecasts should ideally be made of three components:

1. a metric that measures overall quality (discrimination and calibration together), such as the Brier Score (BS) or Ranked Probability Score (RPS)

2. a metric that measures discrimination alone such as the ROC

3. a metric or chart that provides an indication of the reliability (calibration) such as the ranked histogram, reliability diagram or CAL component of the Brier Score.

This combination of metrics will provide a broad perspective on forecast performance and also can assist in the identification of forecast performance issues. For example, when discrimination is good but calibration (biases) issues are degrading the overall quality, a reliability diagram can reveal the nature of the calibration problems [21].

Details about how to compute or construct each of these metrics and diagrams can be found in section 4.3.

5.1.5 Establishing a Cost Function or Evaluation Matrix

Due to the complexity of the task and the fact that the objectives of forecast users are not the same, the following section is an introduction to the concept of a evaluation framework in which structured procedures for the evaluation and verification of forecasts are established. The structure may be shortened and adapted depending on the size of the forecasting system and the importance in the overall business processes.

[1] representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals

[2] consisting of multiple variables, which typically refer to multiple time-steps, multiple sites or multiple parameters

Best practice in this context is to follow a procedure, where the evaluation/verification reflects the importance of forecasts in the role of the business processes and provides incentives for the forecast service provider to generate forecasts that fit the specified purpose.

As a minimum requirement when establishing such an evaluation framework, the following set of procedures should be considered:

1. **Definition of the forecast framework**
   It is important to exactly define the forecast applications, the key time frames and a ranking of the relative importance of each application.

2. **Base performance evaluation on a clearly defined set of forecasts**
   The base performance should contain "typical error" metrics in order to monitor an overall performance level.

   - time frame: minimum 3 months, ideally 1 year
   - "typical error" metrics: nMAE, nRMSE, BIAS

3. **Quality assessment of the evaluation sample data**
   The detection of missing or erroneous data and a clear strategy how to deal with such missing data needs to be made at the outset of any evaluation period to ensure that verification and forecasting is fair and transparent.

4. **Specific Performance evaluation on a set of error metrics**

   - Visual Inspection
   - Use of more specific metrics:
     (a) deterministic: SDE, SDBIAS, StDev, VAR, CORR
     (b) probabilistic: Brier Score, ROC curve, Probability Interval Forecast Evaluation (4.3)
   - Use of histogram or boxplot for evaluation of outliers
   - Use of contingency tables for specific event analysis
   - Use of improvement scores relative to a relevant reference forecast for comparisons
   -

Note, details on the framework and evaluation metrics can be found in [2] and [10], specific metrics and explanation of metrics can be found in [22], [23] for deterministic forecasts and for probabilistic forecast metrics in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11**?**]. Significant tests can be found e.g. in [24].

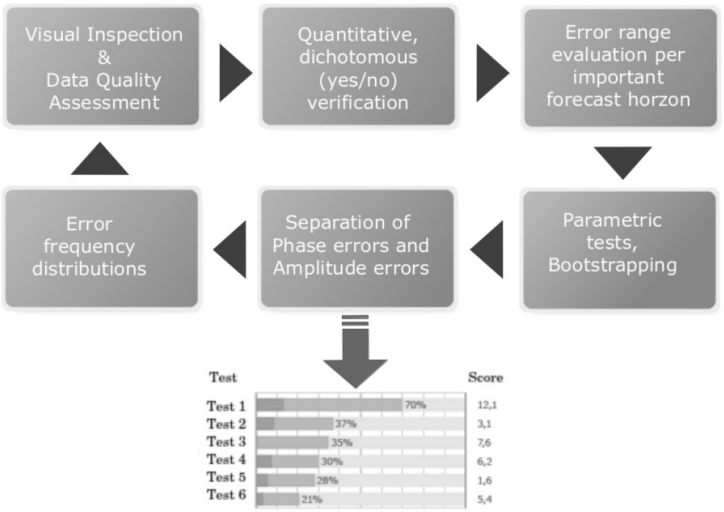

Figure 5.3: Example of an evaluation matrix that verifies forecasts against 6 test metrics and displays the scores for a holistic overview of the forecast performance.

5.1.5.1 Evaluation Matrix

Establishing an evaluation matrix is complex, but can be straightforward if the principles of forecast uncertainty and choice of appropriate metrics are incorporated into the evaluation strategy.

Best practice for the establishment is to go through the various steps outlined in section 5.1.5 to choose the components for the evaluation framework. The core concept is to use this framework to define a formal structure and then add multiplication factors to weight each of the selected individual metrics according to their relative importance.

The matrix can be setup in a spreadsheet environment with macros or within a database environment, where all data is available and metrics may even be directly computed though the database software. The key point of the matrix is that the forecast performance results can be collected, multiplied with an “importance factor”, normalised and transferred into the summary table to visualize the scores. For example the scores can be visualized with a bar chart that indicates the performance in a scale from e.g. 0 to 1 or 0 to 100 as shown in 5.3.

Such a evaluation matrix provides important information in a comprehensive way and can be applied for comparisons of forecast solutions and for the analysis of the potential for forecast improvement.

5.2 Operational Forecast Value Maximization

Key Points

- *Once operational forecasts have been established it is important to monitor the quality of generation facility data supplied to the forecast system(s) and used for forecast evaluation; often attention to this diminishes after a benchmark is completed*
- *Ongoing "deep analysis" of forecast performance and effective provider user communication is critical for maintaining and refining forecast performance*
- *Focus should be on maximizing forecast value for the application and not on maximizing performance of standard metrics; this may include identifying or refining the "cost" function for a user's application and/or working with the provider to optimize forecasts for the application(s)*
- *A plan should be developed to motivate and reward providers to continually refine forecast methods and adapt new approaches from the latest research; this may include financial incentive schemes*

Operational forecasts should be evaluated in the context of their end-use. Different use cases will have different cost functions, some of which may be complex or virtually impossible to define. Organizations evaluate operational forecasts for a variety of reasons and on a wide range of scales, from individual wind farms to entire fleets, and from short lead times to horizons spanning several days.

Simple evaluation metrics such as MAE or RMSE can be used to get an overview of general forecast performance and to provide an indication of forecast performance for decisions with (symmetric) linear or quadratic loss functions, respectively. However, in most cases, the true cost of wind power forecast errors will be more complex and depend on externalities.

Systematic evaluation of operational forecasts is however an important business function for forecast users. Whether this is monitoring the quality of the forecasts produced in-house or procured from vendors, regular evaluation supports continuous improvement in forecast performance and end-use. This section provides a guide to the best practices in evaluation of operational forecasts. It begins by reviewing common motivations for continuous and periodic evaluation of operational forecasts, and then discusses different evaluation paradigms for specific use-cases.

5.2.1 Performance Monitoring

Continuous monitoring of forecast performance is best practice in order to develop an understanding of forecast capability and to identify and respond to issues with raw forecast data or its processing. While failure of forecasting systems is extremely rare, weather models, IT systems, and the forecast target (e.g. individual wind farm, portfolio of wind farms, national wind output) are constantly evolving. This has the potential to introduce new and unforeseen sources of error.

5.2.1.1 Importance of Performance Monitoring for Different Time Periods

Short Periods (monthly): While error metrics or contingency tables calculated over short periods do not provide reliable measures of overall performance they can provide an indication of problems with a forecasting system and large errors should be logged and investigated. Abrupt changes in forecast performance can result from errors in data processing, such as incorrect availability information during maintenance.

Long Periods (> 6 months): Changes in performance over longer time scales may be a result of changes to a supplier's numerical weather model(s) or changes in the behaviour of wind power plant as they age. Slow changes may be more difficult to detect, but over time can accumulate significant biases which should also be investigated.

For both cases, it is necessary to dis-aggregate forecast metrics to identify some sources of error. Important factors to consider when dis-aggregating errors are to include lead-time, time of day, power level and weather type.

Regular reporting and tracking of forecast performance over relevant periods can help foster understanding of forecast capability across business functions and support staff and process development.

Recommendation:

- Forecasts performance should be monitored continuously to quickly identify technical problems
- Large errors should be investigated and recorded for future analysis
- Error metrics should be dis-aggregated by appropriate factors, e.g. lead-time, power level
- Regular reporting for error metrics supports forecast users' interpretation of forecast information

5.2.2 Continuous improvement

Forecast evaluation is the first stage in identifying areas for potential improvement in forecasting systems. Periodically evaluating operational forecast performance and its impact

on wider business functions can be a valuable exercise. For example, changes in the way forecasts are used, or the importance of different lead-times or variables may be a cause to change the way forecasts are produced or communicated internally.

In situations where multiple operational forecasts are produced or supplied, regular benchmarking can add value as different services are upgraded over time or exhibit different performance characteristics.

Recommendation:

- Evaluation underpins forecast improvement and insights should be shared with both forecasters and end-users
- Evaluation and improvement should be driven by end-use and business value

5.2.3 Maximization of Forecast Value

Forecast value can be maximized by continuously monitoring and evaluating operational processes of both forecasts and measurement quality. Additionally, the use of forecasts and the interaction with other business processes need to be taken into consideration as well, if they can impact the quality of the forecasts or the correctness and trustworthiness of the evaluation.

The use of a single metric such as a mean absolute or root mean squared error for forecast evaluation may be a way to start a process and can be helpful in identifying errors in the system that can cause unwanted costs. This is a valid and useful approach. It is however recommended to use such simplified methods only for monitoring purposes and not as the primary verification tool (see also chapter 2, especially sections 2.2, 2.3 and 5.1).

Recommendation: The following aspects should be taken into consideration when identifying a "loss function" or "cost function" in the selection process of performance metrics for operational forecasts. Details on some metrics can be found in the Appendix A, a comprehensive database for metrics can be accessed online [9] together with the concepts of the metrics and valuable combinations of metrics, which have also been described in more detail in section 5.1.

- Evaluation should contain a selection of metrics:
  - One metric alone is not indicative of overall forecast performance
  - Use de-compositions of errors to identify the origin of errors. e.g. look at bias and variance alongside MAE or RMSE.
  - Selected metrics should reflect the costs of errors or security constraints to the greatest extent possible based on the user's knowledge of the application's characteristics

  - Box plots, histograms and scatter plots reveal additional important information compared to a "typical error" metric

- Evaluation metric combinations can provide a representative approximation of a “cost function”:
  - subjective evaluation through visual inspection
  - quantitative, dichotomous (yes/no) verification of critical events such as high-speed shut-down or ramps with e.g. contingency tables
  - error ranges per important forecast horizon
  - error ranges per hour of day or forecast hour
  - error frequency distributions in ranges that have different costs levels
  - separation of phase errors and amplitude errors according to their impact
  - parametric tests, bootstrapping can be used to look on individual error measures before averaging

5.2.4 Maintaining State-of-the-Art Performance

If expensive long-term solutions have been established it can be challenging for an end-user to ensure that state-of-the-art performance is maintained. This can be due to the stiffness of the established IT solution (see also Part 1 of this recommended practice), but also due to the fact that there is no monitoring of the performance.

Recommendation: It is recommended that performance monitoring takes place, where those forecasts that are relevant for the business processes are compared against a suitable and objective measure.
The most common measures are climatology values, persistence values or comparison to previous periods, such as the previous calendar year. Such techniques can provide motivation and can be set up with a reward scheme for the forecast provider to improve forecasts with time and improved knowledge of the specific challenges and needs of the end-user’s forecast problem. (see Table 5.2)

Table 5.2: List of possible performance monitoring types for evaluation of operational forecasts, incentive scheme benchmarks, tests and trials. The types are not meant to be stand-alone and may also be combined.

Performance Measure	**Comment/Recommendation**
Improvement over persistence	Comparison against persistence is the same as comparing "not having a forecast" to having one. Useful measure for short-term forecasts as a mean of evaluating the improvement of applying forecast information to measurements. Note: be aware of data quality issues when evaluating, especially in the case of constant values that benefit persistence, while the forecast provides a realistic view.
Improvement over past evaluation period / forecast	If improvement is important, the comparison to a past evaluation can be useful, especially in long-term contracts. In this way, the forecaster is forced to continue to improve and the target is moved with the improvements. The payment structure however needs to incorporate the fact that improvements reduce over time and have an upper limit.
Comparison against set targets	If the required performance of a forecasting system can be defined, clear targets should be set and the payment directed according to a percentage from 0-100% of the achieved target.
Categorised error evaluation	An effective evaluation format is categorise errors (e.g. large, medium and small errors) instead of setting a single error target. If large errors pose a critical issue, then improvement on these may be incentivized higher and vice versa. The end-user can in that way steer the development and focus of improvements.

5.2.5 Incentivization

Operational forecasts may be tied to an incentive scheme by which monies are exchanged based on forecast performance. Examples of such arrangements exist in both commercial forecast services and regulation of monopoly businesses. As the terms of the incentive scheme typically include details of how forecasts are evaluated, performing this evaluation poses few risks. However, the evaluation methodology should be carefully considered when negotiating or subscribing to such incentive schemes.

Incentives may take the form of a linear relationship between reward/penalty and a forecast metric such as Mean Absolute Error, which may be normalized to installed capacity, and capped at some minimum/maximum reward/penalty. Similarly, incentives may be based on an event-based metric, accuracy or hit-rate for example, for specific events such as ramps or within-day minimum/maximum generation. The time period over which such an incentive is

calculated and settled will have a large impact on it's volatility as evaluation metrics may vary greatly on short time scales. Longer timescales are conducive to a stable incentive reflective of actual forecast performance rather than variations in weather conditions. The basic evaluation rules developed in section 2 and 4 are equalyy valid here and are recommended to be applied.

In summary, the recommendation is that the formulation of an incentive schemes should consider four factors:

- selection of relevant target parameters (see section 2.3)
- selection of relevant metrics (see sections 5.2,5.1, 5.1.5, 5.4.1)
- selection of relevant verification horizons (see section 2.2)
- exclusion principles (see chapter 3 and section 3.2 and **??**)

The selection process of relevant target parameters is highly dependent on the forecasting solution. The objective and proper setup of verification as well as evaluation metrics and frameworks can be found in 2, 4 and sections 5.1, 5.1.2, 5.3.1.

Recommendation: A set of relevant target parameters needs to be defined to provide a focus area for the forecaster. Comparison to a previous period, to a persistence forecast or a set target that is realistic can circumvent a number of constraints that are difficult to exclude in an evaluation. The most important consideration for any performance incentive scheme is that the scheme should put emphasis on the development and advancement of forecast methods for exactly those targets that are important for the end-user's applications.

Table 5.2 provides a list of possible benchmark types for an incentive scheme.

5.3 Evaluation of Benchmarks and Trials

Key Points
In order to maximise the probability of selecting an optimal forecast solution for an application the performance evaluation uncertainty process should be minimised and non-performance attributes of a forecast solution should be effectively considered.
Evaluation uncertainty can be minimised by a well-designed and implemented performance benchmark or trial protocol.
A benchmark should have three well-designed phases: (1) preparation, (2) execution and (3) performance analysis that each address the key issues associated of three primary attributes of an evaluation process.

As a general guideline, the evaluation of benchmarks and trials needs to follow the three principles of being:

1. **representative**
2. **significant and repeatable**
3. **relevant, fair and transparent**

The principles have been explained in detail in Chapter 2. In this section specific considerations and the application of these principles in benchmarks and trials are provided.

5.3.1 Applying the 3 principles: representative, significant, relevant

The three key attributes of a forecast solution evaluation associated with a trial or benchmark (T/B) are (1) representativeness (2) significance and (3) relevance. If any one of these are not satisfactorily achieved the evaluation will not provide meaningful information to the forecast solution decision process and the resources employed in the trial or benchmark will effectively have been wasted. Unfortunately, it many not be obvious to the conductor of a T/B or the user of the information produced by the T/B whether or not these three attributes have not been achieved in the evaluation. This section will present the issues associated with each attribute and provide guidance on how to maximize the likelihood that each will be achieved.

The conductors of a T/B should consider all the factors noted in the three key areas for a T/B. Part of these are described in detail in section 2 in sections 2.1, 2.2 and 2.3. The following is a reminder with specifics for the T/B case:

1. **Representativeness**
   Representativeness in this context refers to the relationship between the results of a trial or benchmark evaluation and the performance that is ultimately obtained in the operational use of a forecast solution. It essentially addresses the question of whether or not the results of the evaluation are likely to be a good predictor of the actual forecast performance that will be achieved for an operational application. There are many factors that influence the ability of the T/B evaluation results to be a good predictor of future operational performance. Four of the most crucial factors here are:
   (a) size and composition of the evaluation sample,
   (b) quality of the data from the forecast target sites,
   (c) the formulation and enforcement of rules governing the submission of T/B forecasts (sometimes referred to as “fairness”),
   (d) availability of a complete and consistent set of T/B information to all T/B participants (sometimes referred to as “transparency”)
2. **Significance** (see section 2.2) For benchmarks and trials it is specifically important that a result obtained now, should also be obtainable when doing a second test. Or, if a test runs over 1 month, the same result should be obtainable over another randomly selected month.

Often, especially in short intervals, this is not possible due to the different climatic and specific weather conditions that characterize specific periods of a year. In this case, it is necessary to establish mitigating measures in order to generate results that provide a correct basis for the respective decision making.

Such a mitigating measure could be to consume potentially new forecasts in real-time and

(a) compare or blend them with a running system in order to test the value of such a new forecast

(b) evaluate the error structure of a potential new forecast to the error structure of your running system

The both tests can be relatively easy incorporated and tested against the main forecast product, such as a day-ahead total portfolio forecast. It will not reflect the potential or performance and quality of a new forecast in it's entirety, but comparing error structures in form of for example error frequency distributions, ensures that a bias due to a lack of training or knowledge about operational specifics does not provide a misleading impression on quality. Chapter 4 details principles and section 5.1 provides details on suitable metrics.

3. **Relevance** (see section 2.3) Results obtained must reflect relevance in respect to the associated operational task and forecasts for energy applications should follow physical principles and be evaluated accordingly. That means in fact that the b/t task must in some way reflect the future function of the forecasts. If this is not so, the results from a b/t should not be used to select a solution of vendor. Instead it may be used to evaluate other performance measures, such as service, support, delivery etc. Fairness in the evaluation, specific for benchmarks and trials then means that the forecast providers are informed about this different objective. Forecasts also need to be evaluated on the same input and output. If assumptions are made, these assumptions must also be provided in a transparent way to alll participants.

A useful approach is to create a evaluation plan matrix that lists all of the factors noted in the discussion in this section and how the user's evaluation plan addresses them.

5.3.2 Evaluation Preparation in the Execution Phase

The evaluation of a T/B should start in the execution phase in order to prevent errors along the way from making results unusable.
Since there is usually a time constraint associated with T/B's there are a number of aspects that should be considered to ensure meaningful results.

.

Recommendations for the execution phase:

Data monitoring:

Measurement data and forecast delivery should be monitored and logged in order to prevent data losses and to ensure that all relevant data is available for the evaluation. It is recommended that the data monitoring should contain the following tasks:

- test accuracy and delivery performance for fairness and transparency
- monitor forecast receipt to test reliability
- exclude times, where forecasts are missing to prevent manipulation on performance

Consistent Information

The fourth key factor is the availability of a complete and consistent set of T/B information to all participants in the T/B. Incomplete or inconsistent information distribution can occur in many ways. For example, one participant may ask a question and the reply is only provided to the participant who submitted the inquiry.

Develop and refine your own evaluation scripts:

Independent whether it is a first time b/t or a repeated exercise, the execution phase is the time, where the following evaluation has to be planned and prepared. It is recommended to verify metrics scripts or software tool and input/output structures as well as exclusion principles.

5.3.3 Performance Analysis in the Evaluation Phase

The performance analysis has a number of key points that need consideration. These are:

1. Application-relevant accuracy measures of the forecasts
   The key point here is that the metrics that are used in the verification must have relevance for the application. For example, if a ramp forecast is tested, a mean average error only provides a overall performance measure, but is not relevant for the target application. If a vendor knows that performance is measured with an average, the incentive would be to dampen forecasts to reduce the overall average error, which is the opposite of what is required for the application to work. Such an application would have to use a scoring system for hits, misses and false alarms of pre-defined ramping events.

2. Performance in the timely delivery of forecasts
   The key pitfalls in an T/B are often associated with the failure to closely monitor the following aspects:

   (a) Lack of check or enforcement of forecast delivery time
   If forecast delivery is not logged or checked, it is possible for a forecast provider to

deliver forecasts at a later time (perhaps overwriting a forecast that was delivered at the required time) and use fresher information to add skill to their forecast or even wait until the outcome for the forecast period is known. Although one might think that such explicit cheating is not likely to occur in this type of technical evaluation, experience has indicated that it is not that uncommon if the situation enables its occurrence.

(b) Selective delivery of forecasts
This example illustrates how the results might be manipulated with explicit cheating by taking advantage of loopholes in the rules. In this example the issue is that the B/T protocol does specify any penalty for missing a forecast delivery and the evaluation metrics are simply computed on whatever forecasts are submitted by each provider. As a forecast provider it is easy to estimate the "difficulty" of each forecast period and to simply not deliver any forecasts during periods that are likely to be difficult and therefore prone to large errors.
This is an excellent way to improve forecast performance scores. Of course, it makes the results unrepresentative of what is actually needed by the user. Often it is good performance during the difficult forecast periods that are most valuable to a user.

3. Ease of working with the forecast provider
In a T/B support in understanding forecast results and error structures may be a good time to test and evaluate for the future. It should however be considered to communicate to the vendors, if it is a decision criteria, especially in non-refunded situations, where resources are used differently than in contractual relationships.

5.3.4 Evaluation examples from a benchmark

Figure 5.4 shows an example of a forecast evaluation using a box-and-whiskers-plot to visualize the spread in MAPE (mean absolute error as percentage of nominal power) of 5 forecasts of different day-ahead time periods (each column) at two different sites. The distribution within each time period is shown for the 5 forecasts errors. In that way, the spread of forecast performance in each hour of the day-ahead horizon can be visualized. It also shows how some forecasts in some hours show very low errors compared to the average error in that hour, as well as occasionally very high errors.

Figure 5.5 shows an example of an evaluation of errors by time of day for a fixed lead time of 3 hours. It illustrates a very large spread in errors during certain times of the day, as would be expected.
Nevertheless, if such evaluations are compared between different forecast providers an evaluation of the "most costly errors" may reveal a very different result than, if only an average metric per forecaster would be used.

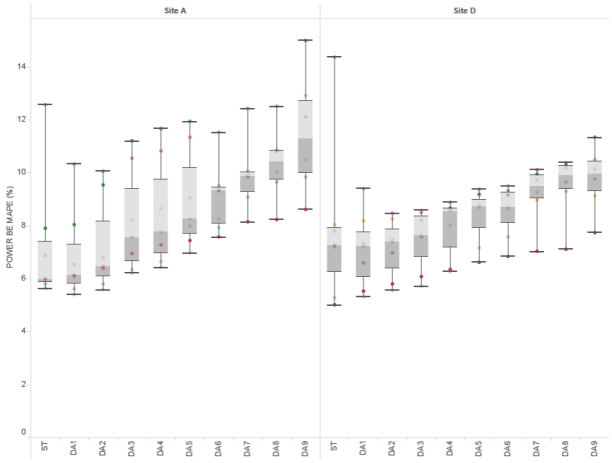

Figure 5.4: Example of a box-and-whisker-plot verification at two different sites (left and right panel) for different look ahead times (x-axis; DAx is x^t *h* hour of day-ahead forecast) and mean absolute percentage error (MAPE; y-axis).

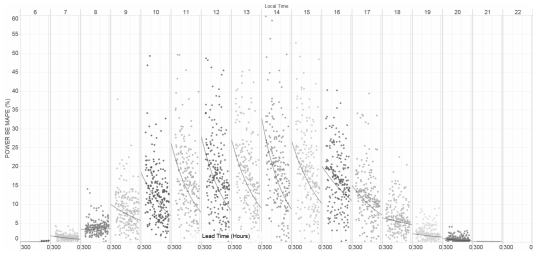

Figure 5.5: Example of a forecast error scatter plot by time of the day (top x-axis) for 3-hours lead times and forecast error (y-axis)

5.4 Evaluation of Development Techniques

Key Points

Maintaining forecast performance at a state-of-the-art level is an important objective for any end-user, but especially for those with complex IT infrastructure systems or multiple suppliers of forecasts that are bound to statistically consistent forecasts over a period of time for highest performance.

This Section outlines how analysis, diagnostics and evaluation of improvements need to be structured in order to ensure sustained improvement over time without radical changes in existing infrastructures and the typical pitfalls associated with such evaluations.

5.4.1 Forecast Diagnostics and Improvement

The improvement of a forecast over time is especially important in an operational environment, where the IT infrastructure is complex and the amount of resources required to change a forecast service provider is high relative to the likely gain in forecast performance that would be produced by such a change. The following recommendations may therefore be applied for any of such cases, where an end-user is bound to a forecast solution.

Improvements over time and the importance of a forecast solution being able to develop over time in a real-time environment is difficult to measure. Also, the improvement of forecasts may have a steep curve in the first years, or when constant changes in the system become less frequent.

However, over time, forecast performance has a limit and the rate of improvement will be reduced. This needs to be taken into account as much as the ability of a forecast solution to develop over time to maintain a state of the art character.

Table 5.2 is a guideline for the evaluation of forecasts and diagnostics for such improvement monitoring (see also 5.2.5).

5.4.2 Significance Test for new developments

Forecast vendors and researchers are always seeking for improvements and new developments, testing and investigating new technology or techniques to add value to specific tasks in the forecasting arenas. Whenever a new development is ready for testing, the researchers or technical staff are confronted with the question, whether the new technique outperforms the older or current state of the art. Due to time constraints, data limitations or lack of historical available forecasts or measurements, this is often a difficult question to answer.

The following example demonstrates such a typical situation and presents and outlines the overall considerations that need to be taken, followed by the choice of metrics and test on significance on the results.

Initial Considerations

A forecasting model that can take various inputs, such as online measurements in an autoregressive manner, weather forecasts or other predictive features, generates power forecasts, which estimate the future electricity production. In order to decide which model is most suitable, it is necessary to evaluate its quality by comparing the forecast against power measurements. Typically, the errors of a separate test data are compared against each other in order to then decide in favour of one of the models. Which error measure is chosen should be individually adjusted to the corresponding application.

The evaluation should be performed strictly on test data that were not used to calibrate the respective model. Otherwise, it can easily happen that models are favoored, which have adapted too much to the training data without being able to generalize for future unknown situations. If several models are compared, they should also have been jointly trained on data that does not originate from the test set.

In the case of wind power forecasts, it is furthermore essential to select the test data from a continuous period. The data cannot be considered temporally independent. If one were to randomly assign individual samples to a training and a test set, one would assign both sets of random samples that share a large part of the information. As a result, preference would also be given to models that are over-adapted to the training data.

In addition to the error measure, other aspects can also play a role. For example, one is faced with the question whether an established model should be replaced. For several reasons it may seem attractive not to replace it even though another one shows a smaller error. For instance, because confidence in the model functionality has been built up, or because a change in the model requires additional effort. Such or similar cases make it necessary to examine the significance of the estimated error values. The critical question behind this is whether the extent of the test data considered is sufficient to form the basis for a decision.

Evaluation of Significance

One way to evaluate the significance of the error values is to evaluate the distribution of the error measures of a model across different locations. In the following, the relevant aspects of the results of the study in [24] are summarized. It compared different machine learning models to weather forecasting and real-time measurement based forecasting. The box plot shown in Figure 5.6 shows the distribution of the error measures of 29 wind farms in northern Germany. The error measure used here is the root mean square error (RMSE) which is applied to nominal power normalized time series. The individual boxes represent the error distribution of one of the six models used. The triangular markers indicate the confidence range of the median. If these ranges do not overlap for two models, the medians are different under normal distribution assumption to a 5% significance level. This corresponds to a visual representation of a t-test.

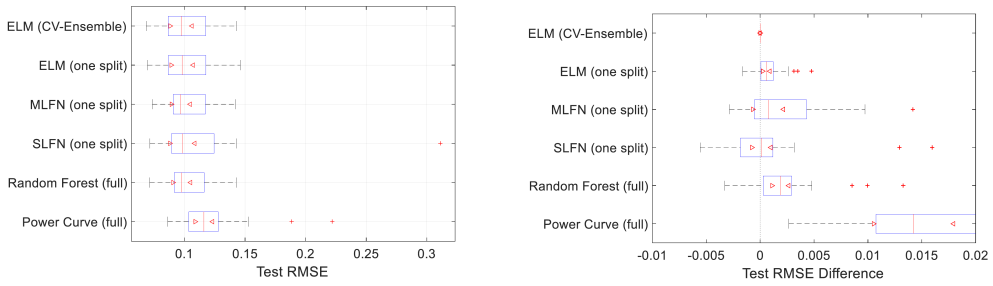

Figure 5.6: RMSE distribution for six different forecasting models forecasting for 29 wind farms in the North of Germany (left figure). Pairwise differences RMSE for each single model in comparison to the wind farm RMSE of the reference model ELM (CV-Ensemble) [24] (right figure).

Figure 5.6 (left) shows, that only the power curve model has a significantly higher RMSE. All others cannot be clearly distinguished. The reason for this can be found in the broad

distribution. This can be explained to a greater extent by the different local properties, such as the number of turbines per wind farm or the local orography. When considering the paired differences, local influences can be partially eliminated.

Figure 5.6 (right) shows the distribution of the difference between a model and a reference model (ELM (CV-Ensemble)) across all 29 wind farms. If the distribution of a model is significantly in the positive range, it can be assumed that the reference model is significantly better. Thanks to these pairwise differences, it can now be established that two other models have a significantly worse result.

5.5 Use cases

Key Points
The section presents a number of use cases that illustrate how an evaluation in a specific part of the power and energy sector should ideally be designed and executed. In the ***Energy Trading and Balancing, ramping forecast in general and for reserve allocation****, forecasts are a crucial part of the processes for balance-responsible parties as well as system operators. And yet, many mistakes are made in the evaluation and incentivization of forecasts that effectively often lead to results that are unsatisfactory and create mistrust in the ability of forecast service providers to have skills to provide useful forecasts.*

5.5.1 Energy Trading and Balancing

In energy trading, forecasts of multiple variables are used in order to provide situational awareness and support quantitative decision-making. Costs accrue on the basis of forecasts and energy prices at multiple look-ahead times. An example is forecasts used at the day-ahead stage and then again at an intra-day look-ahead time frame for the same trading period, and the relative price of buying and selling energy at different times.

Furthermore, prices, particularly imbalance prices, may be influenced by the cumulative forecasts and forecast errors of all market participants creating dependency between wind power forecast errors and the price at which resulting imbalances are settled. Similarly, unrelated events may cause large price movements that result in an otherwise unremarkable forecast error having a large financial impact. Therefore, care must be taken when designing an evaluation scheme that is reflective of forecast performance and not externalities.

5.5.1.1 Forecast error cost functions

If trading decisions are based on a deterministic power production forecast, it is tempting to try and evaluate the ‘cost’ of forecast errors based on energy prices.

For example by taking the cost of under forecasting to be equal to the difference between the day-ahead price and the system sell price (the opportunity cost of having to sell at the system sell price rather than day-ahead price), and taking the cost of under forecasting to be equal to the difference between the system buy price and the day-ahead price (the cost of having to buy back the energy not produced at a higher price than it was sold for).

This approach has several problems:

1. price asymmetry:
   Traders are aware of the asymmetry in imbalance prices and have a view of whether the market is likely to be long or short, as such they do not naively trade the forecast production and will hedge against penalizing prices. It is therefore not representative to assume the day-ahead forecast is contracted.

2. adjustment opportunities:
   The intra-day market and flexibility within the traders portfolio provide opportunities for adjustment between the day-ahead market and imbalance settlement which may influence both the value and volume of traded energy, and potentially the imbalance price.

3. forecast error correlation:
   Renewable power forecast errors are often highly correlated across the entire market and therefore to the market length and total imbalance. As a result, evaluating forecast errors based on imbalance cost will not discriminate between forecast performance and correlation with imbalance prices and one may incorrectly interpret reduced 'cost' as improved forecast skill.

For these reasons it is recommended that (normalized) mean absolute error be used as part of an evaluation matrix of other relevant metrics when evaluating deterministic renewable power forecast performance for trading applications (see 4, 5.1). Additionally, a real-example of a market analysis and evaluation of how different trading strategies influence tne costs in comparison to the revenue can be studied at [25], and [26].

If trading decisions are based on probabilistic power production forecasts those forecasts should be evaluated as described in section 4.2. If probabilistic forecasts of both power production and prices are used, it is important that the dependency structure between power and price forecast errors is correct. Various metrics exists to measure this, such as the multivariate energy score [5] and p-variogram score [11, 27]. Details are beyond the scope of this document.

5.5.2 General Ramping Forecasts

Power ramps can have significant impact on power system and electricity market operation and are of interest to decision-makers in both domains. However, as ramps comprise a

sequence of two or more forecasts, metrics that only compare predictions and observations at single time points are not suitable for evaluating ramp forecasts. Event-based evaluation in the form of contingency tables and associated metrics provide a tool-set for evaluating these forecasts.

Once an event is defined, such as ramp defined as a particular change in wind energy production over a particular time period, occurrences in forecasts and observations can be labeled and a table of true-positive, false-positive, true-negative and false-negative forecasts can be produced. From this, the skill of the forecast at predicting such events can be evaluated.

The definition of a ramp will influence the forecast tuning and evaluation results. It is recommended that the definition reflects the decision(s) being influenced by the forecast. For example, this could be related to a commercial ramp product definition, or the ramp rates of thermal power plant used in balancing. Furthermore, if an economic cost can be assigned to each outcome, then the forecasting system can be tuned to minimize costs, and the relative value of different forecasting systems can be compared.

In general terms, the following methods and metrics are recommended as basis for the evaluation of ramp forecasts:

- Contingency tables and statistics derived from the tables provide an evaluation framework
- Ramp definitions should reflect operational decision-making
- The cost implications of different types of errors should be considered when comparing different forecasting systems

In the next sections, a number of examples are described to demonstrate how evaluation should be planned and that illustrates the pitfalls in the metric selection process.

5.5.2.1 Amplitude versus Phase

Ramping events cause shortage or overproduction and risk for congestion in the power system for relatively short time frames. For this reason, many system operators have different levels of reserve time frames and also forecasting time frames that provide the possibility to allocate different types of reserve to counteract ramps that have been forecasted insufficiently strong (amplitude) and/or are wrong in phase. On system operator level it is often described that the amplitude is more important than the exact timing (phase).

In this case, it is necessary that the evaluation method does not punish the forecaster stronger for a phase error than an amplitude error. This means for example that using a root mean square error to evaluate ramps is incentivizing a forecaster to dampen amplitudes and optimize on phase. Sometimes it is referred to the **"forecaster's dilemma"** when the end-user defines a metric for evaluation such that the target is opposite of what the end-user asks for and needs. The forecast provider then either tunes forecasts to the metric or to what

the end-user likes to see and risks to be punished (e.g. loose a contract), when evaluated. See also [28].

Recommendation: When a forecaster should be incentivized for amplitude in a ramp forecast, the evaluation metric cannot be an average error measure such as mean absolute error or root mean square error. If these average error metrics are used, the data to be evaluated has to be prepared to:

- reflect only cases that contain ramps of a certain strength
- widen ramp events with a forward/backward window of $1-2$ hours to allow for phase errors

Additionally, either a contingency test with hit rate, misses and false alarms have to be used in the evaluation of the forecasts to reflect the focus on amplitude.

5.5.2.2 Costs of false alarms

Ramps can have different costs in a power system. In some systems, too fast up-ramping causes congestion or in some way over-production that needs to be dealt with (case 1). The opposite case, the down-ramping can cause that there is power missing on the grid that is not available and the fast primary reserve causes high costs (case 2). In case 1, the system operator has to be able to reduce ramping capacity of the wind farms or have other highly flexible resources on the grid to level out the overproduction. In case 2, lacking energy can cause high costs for fast ramping resources on primary reserve level or outages, which are unwanted.

The consequence is that the cost profile for up-ramping and down-ramping is usually different. Also, the cost of not forecasting a ramp that occurs (false-negative) can be significantly higher than the cost of preparing for a ramp, which does not occur (false-positive). The only way to verify, whether a forecast is sufficiently good in predicting a specific type of ramping event is to use contingency tables, where the forecast skill can be computed and visualised.

5.5.3 Evaluation of probabilistic Ramp forecasts for Reserve Allocation

The primary scope of reserve predictions is to reduce balancing costs via dynamic allocation of reserve and if possible with the help of non-fossil fuel capacity.
If a system operator (SO) or balance responsible party (BRP) can more dynamically schedule reserve, the costs for imbalances become lower and the energy system more efficient.

This was the scope of a study that will be presented as an example of the evaluation of a real-time environment application that needed a practical solution in order to reduce costs for reserve allocation for the end-user [29]. The evaluation strategy and results of the study can be considered a kind of guideline on how to best manage renewable energy imbalances in a market system.

In this sample control area there are approximately 40 wind farms. The permanent allocation of reserves for the control area amounted at the outset to +/-10% and up to +/- 30% of installed capacity of wind, dependent on the time of the year, i.e. there are large seasonal reserve allocation differences. In our example area the wind generation is correlated and strong ramps occur. However, it is seldom to observe that the wind generation ramps down in a dramatic speed. Ramp-ups are faster than down-ramps and it is very unlikely that an instant total wind ramp down to zero can occur in the control area.

5.5.3.1 Definition of Error Conditions for the Forecast

Fundamental for forecasting is that a criteria for success and error can be defined. Given the fact that certain swings in the data are unrealistic or possibly so extreme that the operational cost of self-balancing would be too high, there was a need to work with probabilities. One way of doing this is to define that, if a forecast value lies within a band, the result is a success and if it lies outside the band, it is a false alarm. A constant very wide reserve band would imply 100% success, but would not be affordable.

The gain lies in finding a balanced criterion considering the following questions:

- How many failures can be tolerated ?
- What is the allowed maximum error ?
- Which frequency of reserve under-prediction is allowed ?
- What is the cost of spilled reserve ?

These questions are related or determined by the SO's operational experience and standards to which the SO must be conforming. Figure 5.7 illustrates the challenges of deciding how many outliers can be accepted to reduce costly spill, a dilemma every balance responsible party has to deal with. The static allocation of reserves is very expensive, especially if all extremes should be covered. Even, if extremes are not covered always, there is a lot of spill (black areas in Figure 5.7) in comparison to a dynamic allocation of reserves.

The difficulty for such a situation is to find objective criteria suitable for evaluation of a model result, which relates to operation and presents incentives for the forecaster to reduce the spill by maximizing coverage of extremes. Standard statistical metrics do not provide answers to this optimization task, because (1) it is not the error of 1 forecast any more and (2) the target is whether the allocation was sufficient and cheaper than allocating with a constant "security band".

With contingency statistics it is possible to ask the right questions:

Hits and Misses Analysis show the percentage of time the band was too small Positive and negative reserve allocation can be split up to reflect use of tertiary reserve allocation (cheaper) instead of primary reserve (high expenses)

The following analysis was carried out to reflect these objectives:

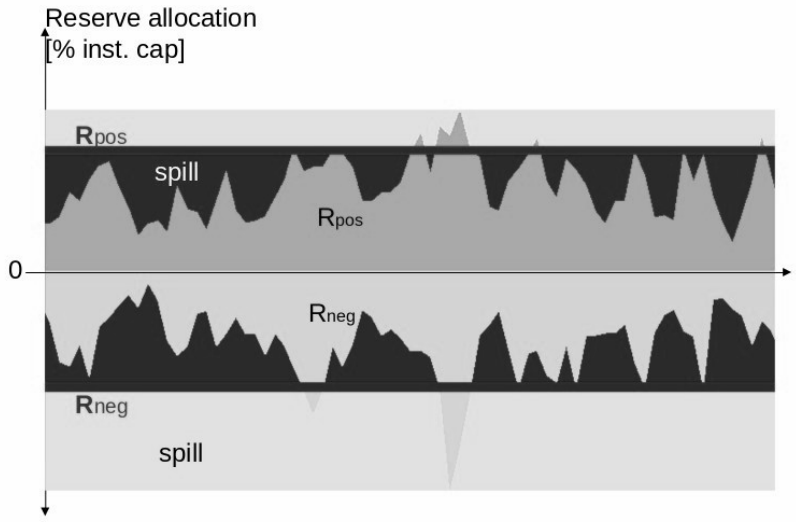

Figure 5.7: Illustration of the "reserve allocation dilemma" of costly spill versus covering all possible ramping events. Here, R_{pos} is the dynamic positive reserve, R_{neg} is the dynamic negative Reserve, the upper linear borders R_{pos} and R_{neg} are the static reserve allocation, the black area and the outer light grey areas are the spill for the dynamic and static allocation of reserves, respectively.

Table 5.3: Applied metrics in the evaluation matrix for the reserve allocation example in [29]. The input forecasts are split up in 9 percentile bands from P10..P90 and a minimum and maximum.

Metrics		Purpose	Input forecasts
BIAS		average to gain overview	MIN
MAE		average to gain overview	P10
RMSE		average to gain overview	P20
Inside Band		consistency forecast-deployment	P30
$R_{coverage}$		forecasted reserve deployment	P40
Hit rate	Total	achievable percent of activated reserve	P50
	R_{pos}	as above for pos reserve	P60
	R_{neg}	as above for neg. reserve	P70
Misses	Total	avg under-predicted reserve	P80
	R_{pos}	as above for pos reserve	P90
	R_{neg}	as above for neg. reserve	MAX
Spill	Total	avg over-predicted reserve	
	R_{pos}	as above for pos reserve	
	R_{neg}	as above for neg reserve	

1. A BIAS, MAE and RMSE provide an overview of the plain statistical capabilities of the various forecasts

2. Contingency tables for hit rate, misses, spill and reserve coverage have been computed to provide metrics for further optimization of the task

Table 5.3 shows the evaluation matrix of metrics and their purpose in the verification and further optimisation. The study [29] concluded that the real reserve deployment will not be able to cover the shortage or overcapacity for about two hours per day in average. Their 5760 hours of evaluation was not considered very robust to draw final conclusions and to set long-term strategies, it was found that the results provided the information necessary to enhance the optimisation task and follow it's progress closely over some time.

Bibliography

[1] Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559 – 570, 2000.

[2] Madsen H., Pinson P., Kariniotakis G., Nielsen HA, and Nielsen TS. Standardizing the performance evaluation of short-term wind power prediction models. *Wind Engineering*, 29(6):475˘489, 2005.

[3] G. Candille and O. Talagrand. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131(609):2131–2150, 2005.

[4] Towards the definition of a standardised evaluation protocol for probabilistic wind power forecasts. Technical report, Anemos.Plus Project, 2012.

[5] Tilmann Gneiting, Larissa I. Stanberry, Eric P. Grimit, Leonhard Held, and Nicholas Alexander Johnson. Assessing probabilistic forecasts of multivariate quantities with an application to ensemble predictions of surface winds. *TEST*, 17:211–235, 2008.

[6] S. J. Mason. Understanding forecast verification statistics. *Meteorological Applications*, 15(1):31–40, 2008.

[7] Wilks. *Statistical Methods in the Atmospheric Sciences*. Third edition, 2011.

[8] T. L. Jensen, T. L. Fowler, B. G. Brown, J. K. Lazo, and S. E. Haupt. Metrics for evaluation of solar energy forecasts. Technical report, 2016.

[9] WWRP/WGNE Joint Working Group on Forecast Verification Research.

[10] Jakob W. Messner, Pierre Pinson, Jethro Browell, Mathias B. Bjerregård, and Irene Schicker. Evaluation of wind power forecasts—an up-to-date view. *Wind Energy*, 23(6):1461–1481, 2020.

[11] Mathias Blicher Bjerregård, Jan Kloppenborg Møller, and Henrik Madsen. An introduction to multivariate probabilistic forecast evaluation. *Energy and AI*, 4:100058, 2021.

[12] GLENN W. BRIER. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950.

[13] E. S. Epstein. A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor*, 8:985–987, 1969.

[14] Ensemble verification metrics. Technical report, European Center for Medium Range Forecasting, 2017.

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

[16] Wwrp/wgne joint working group on forecast verification research. `https://www.cawcr.gov.au/projects/verification/#Methods_for_probabilistic_forecasts`. Accessed: 2021-10-11.

[17] Receiver operating characteristic. `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`. Accessed: 2021-10-11.

[18] TM Hamill and J Juras. Measuring forecast skill: is it real skill or is it the varying climatology? *Q.J.R. Meteorol. Soc.*, (132):2905 – –2923, 2006.

[19] D.S. Wilks. A skill score based on economic value for probability forecasts. *Meteorol. Appl.*, 8:209 – –219, 2001.

[20] D.S. Richardson. Skill and relative economic value of the ecmwf ensemble prediction system. *Quart. J. Royal Met. Soc.*, 126:649 – –667, 2001.

[21] Verification of climate forecasts: How is forecast skill or accuracy measured? – what aspects of forecast quality is measured by various scores? `http://indico.ictp.it/event/a09161/session/29/contribution/17/material/0/0.pdf`. Accessed: 2021-10-11.

[22] T. Jensen, T. Fowler, B. Lazo J. Brown, and Haupt S.E. Metrics for evaluation of solar energy forecasts. Technical report, NCAR, 2016.

[23] Zhang. A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111:157, 2015.

[24] S. Vogt, A. Braun, J. Koch, D. Jost, and R.J. Dobschinski. Benchmark of spatio-temporal shortest-term wind power forecast models. 2018.

[25] Corinna Möhrlen, Markus Pahlow, and Jess U. Jørgensen. Untersuchung verschiedener handelsstrategien für wind- und solarenergie unter berücksichtigung der eeg 2012 novellierung. *Zeitschrift für Energiewirtschaft*, 36(1):9–25, March 2012.

[26] Corinna Möhrlen, Markus Pahlow, and Jess U. Jørgensen. Author's english translation of *(Untersuchung verschiedener Handelsstrategien für Wind- und Solarenergie unter Berücksichtigung der EEG 2012 Novellierung* / investigation of various trading strategies for wind and solar power developed for the new eeg 2012 rules.

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

[28] Ravazzolo F Lerch S, Thorarinsdottir TL and Gneiting T. Mforecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1):106 – –127, 2017.

[29] Jørgensen J.U. Möhrlen, C. Reserve forecasting for enhanced renewable energy management. 2014.

Appendix A

Standard Statistical Metrics

Commonly applied standard statistical metrics

Mean Absolute Error (MAE): *The average of all absolute errors for each forecast interval. Measures the average accuracy of forecasts without considering error direction.*

$$\frac{1}{n}\sum_{i=1}^{n}(f_i - m_i)$$

Mean Absolute Percent Error (MAPE): *This is the same as MAE except it is normalized by the capacity of the facility.*

Root Mean Square Error (RMSE): *Measures the average accuracy of forecasts without considering error direction and gives a relatively high weight to large errors*

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - m_i)^2}$$

Root Mean Square Percent Error (RMSPE): *As above normalize by plant capacity.*

BIAS: *Indicates whether the model is systematically under- or over-forecasting*

$$\frac{1}{n}\sum_{i=1}^{n}(f_i - m_i)$$

Correlation: Correlation is a statistical technique that is used to measure and describe the STRENGTH and DIRECTION of the relationship between two variables.

$$r(x,y) = \frac{COV(x,y)}{STD_x \cdot STD_y} = \frac{\sum (x-\bar{x})\cdot(y-\bar{y})}{N \cdot STD_x \cdot STDy}$$

where f are the forecasted values, m are the measurements, COV is the covariance, STD is the standard deviation.

Standard Deviation: A measure of the spread or dispersion of a set of data. The more widely the values are spread out, the larger the standard deviation. It is calculated by taking the square root of the variance.

$$\text{STD} = \sqrt{\left(\frac{\sum \left((f_i - \bar{f}_i)^2\right)}{n}\right)}$$

Variance: A measure of the average distance between each data point and the data mean value; equal to the sum of the squares of the difference between each point value and the data mean.

$$\sigma^2 = \frac{\sum \left((f_i - \bar{f}_i)^2\right)}{n}$$