EXPERT GROUP REPORT

ON

RECOMMENDED PRACTICES FOR SELECTING RENEWABLE POWER FORECASTING SOLUTIONS

Part 1: FORECAST SOLUTION SELECTION PROCESS

1. EDITION 2019

Submitted to the
Executive Committee of the
International Energy Agency Implementing Agreement
on 13th August 2019

Prepared as part of the IEA Wind Task 36, WP 2.1.

Version:2.6

Date: 13. August 2019

Edited by:

Corinna Möhrlen (WEPROG, DK, DE)

John Zack (UL AWS Truepower, USA)

With Contributions from:

Jeffrey Lerner (Vaisala, USA)

Mikkel Westenholz (ENFOR, DK)

Supported by:

Operatiing Agent Gregor Giebel (Danish Technical University, DK)

Table of Contents

1 BACKGROUND AND OBJECTIVES

1.1 BEFORE YOU START READING

This is the first part of a series of three recommended practices that deal with the selection and design of renewable energy forecasting solutions in the power industry.

The first part "Forecast Solution Selection Process", which is the current document, deals with the selection and background information to be collected and evaluated when designing or renewing a forecasting solution for the power market.

The second part of the series "Benchmarks and Trials", of the series offers recommendation on how to best conduct benchmarks and trials in order to test different forecasting solutions against each other and the fit-for-purpose.

The third part "Forecast Evaluation", provides information and guidelines for the effective evaluation of forecasts and forecast solutions within benchmarks and trials as well as in other applications.
If you already have experience in setting up a forecast solution and you have an up-to-date IT infrastructure, then it is recommended to go straight to part 2 or 3.
The information in these recommended practices documents provides guidance for ongoing or future forecast users who are selecting an external forecast solution, building an internal forecasting capability or evaluating the effectiveness of an existing solution. This includes those who are starting a process to address

- a renewal of their IT infrastructure
- a need for new forecasting products
- an extension or reduction in the number of forecast vendors in their solution
- the building of a forecast solution from scratch

An overview of the decision support tool to help develop structured processes in the design and planning for a new, or renewal of a, forecasting solution can be found in chapter 3, while chapters 1 and 2 provide background information and initial considerations. It is recommended to use the table of contents actively to find the topics that are most relevant for you.

1.2 BACKGROUND

The forecast's effectiveness in reducing the costs for the variability management of power generation from wind and solar farms is dependent upon both the accuracy of the forecasts and the ability to effectively use the forecast information in the grid management decision-making process. Therefore, there is considerable motivation for stakeholders acting in the power market to try to obtain high quality forecasts and effectively use this information as input to other operational processes or trading.

This document is intended to provide guidance to stakeholders who are seeking a forecasting solution that fits their purpose and enables them to work efficiently and economically responsible.

In recent years, carrying out trials or benchmarks seemed to be an industry practice in the power market with an easy and uncomplicated decision process for many. In reality, trials are often expensive for both the end-user and the vendor, are quite complicated, and not entirely conclusive. Benchmarks have little value for commercial vendors, except in their start-up phase, and end-users can often not count on results that reflect state of the art. Further, if trials and benchmark studies lead to a dissatisfying result, forecasting solutions become increasingly criticized for their value. And, providers that may have had the most technically qualified solution at hand, but did not score best at a specific (maybe simplified) test, may be deselected.

This recommended practices document will therefore focus on the key elements to consider when seeking to establish or renew a forecasting solution that fits one's purpose.

In summary, this document provides recommendations and a decision support tool to establish procedures for an effective selection process.

1.3 OBJECTIVES

This document is intended to serve as guidance and best practice for private industry, academics and government for the process of obtaining an optimal wind or solar power forecast solution for their applications and, in particular, it provides guidance to the design and requirements for effective renewable energy forecasting solutions.

These guidelines and best practices are based on years of industry experience and intended to achieve maximum benefit and efficiency for all parties involved.

DEFINITIONS

In the discussion of the process of obtaining the best possible forecasting solution, there are a number of terms and concepts that are used. Several of the key terms and concepts are defined in the following.

Note, these definitions are kept as general as possible with a focus on forecasting processes in the power industry and may not have such a completely general character to be applied to other areas of business.

Request for Information (RFI)*: a RFI allows the client to get information about the state-of-the-art business practices and available commercial products in the preparation or design of a forecast application or solution for a specific target process. By providing information about the target application, a client can ask vendors for their recommendations and experience to solve specific tasks. Such information is useful in the preparation and design of a new system, but also for systems that need to be rebuilt due to changing requirements.*

Request for Proposal (RFP): a RFP is a tender process, where the client prepares a document laying out the requirements of a forecasting solution and asking vendors to propose a solution and price quote. Usually, a set of minimum requirements are provided that become part of a contractual agreement for the awarded vendor.

Renewable Energy Forecast Benchmark: an exercise conducted to test features and quality of a renewable energy forecast such as wind or solar power. The exercise is normally conducted by an institution or their agent and usually includes multiple participants from private industry forecast providers or applied research academics.

Renewable Energy Forecast Trial: an exercise conducted to test the features and quality of a renewable energy forecast such as wind or solar power. This may include one or more participants and is normally conducted by a private company for commercial purposes. A trial is a subset of a Renewable Energy Forecast Benchmark.

Renewable Energy Forecast Product: a specified set of content, format and delivery protocols of forecast information supplied by a forecast system

Renewable Energy Forecast Solution: a set of forecast products and supporting information that address the specific needs of a user's application; it may be based on an external (e.g. supplied by a vendor) or internal (e.g. formulated and managed by the user) forecast system

Renewable Energy Forecast System: an integrated set of IT hardware and software that ingests external data, uses physics-based and/or statistical models to process it and generates a set of forecast products

Renewable Energy Forecast Application: a user's process that has non-negligible sensitivity to the future weather-dependent behavior of renewable energy generators

High-level Overview of a typical state of the art forecasting solution: the components and data flow of a typical state-of-the-art forecasting solution is schematically depicted in Figure 1. This schematic indicates that solutions are a combination of physics-based (such as Numerical Weather Prediction (NWP)) models and statistical methods. In Figure 1, the physics-based methods are denoted by blue objects, the statistical components are depicted by green objects. Components that can be either statistical or physics-based depending on the configuration of the solution are denoted by a combination of green and blue colors.

Almost all current solutions have a structure that is a specific configuration of this general framework. The variations among potential solutions are typically related to the type (i.e. specific method formulations) and number of instances of each component that is included in the solution. For example, a particular solution may use output from many government-center NWP models while another solution may employ the output from only one government-center NWP model.

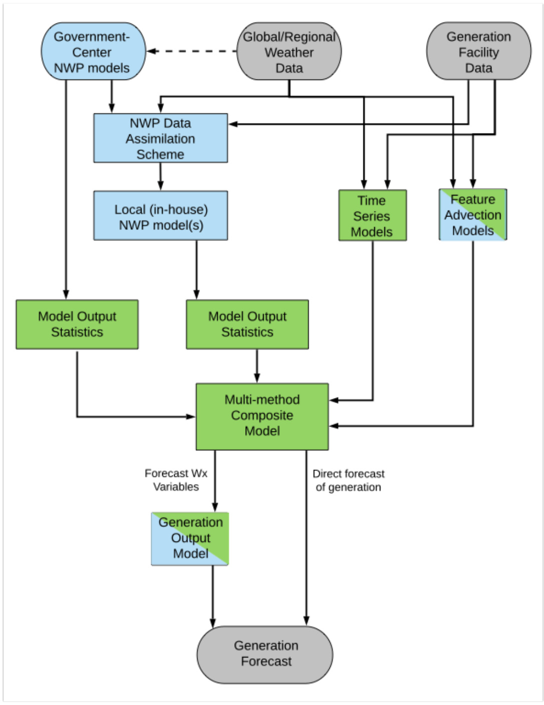

Figure 1: High-level overview of the components and data flow of a typical state-of-the-art forecasting solution.

Another example is the type of statistical models used for the MOS component (which is intended to reduce systematic errors in the NWP output). One solution may use a traditional multiple linear regression approach for this purpose while a different provider might utilize a sophisticated machine-learning model such as an Artificial Neural Network (ANN) or a combination of statistical methods. These system design decisions play a major role in the determination of how well a particular solution is able to meet the requirements of a specific application. Therefore, it is valuable for the user to attempt to gather information that provides an understanding of design differences among alternative solutions.

2 INITIAL CONSIDERATIONS

This part of the IEA Wind Task 36 recommended practice series provides guidelines for those whose task is to provide a plan and justification for a forecasting solution selection process. It intends to assist in finding the necessary information when navigating through the vast jungle of information, opinions and possibilities and ensures that crucial details are being considered.

2.1 TACKLING THE TASK OF ENGAGING A FORECASTER FOR THE FIRST TIME

The most important considerations and first question to answer, when starting out to plan the selection of a forecasting solution is to be clear about the desired outcome. A lot of time and resources can get wasted for all involved parties on trials and benchmarks that are not aligned with requirements, also when planned and conducted by personnel with little or no experience in the subject.

To avoid this, the recommended practice is to carry out a market analysis in the form of a "request for information" (RFI) and to establish a requirement list (see also APPENDIX B).

In some cases, it can be beneficial to test vendors or solutions prior to implementation. The difficulty with this method lies in the evaluation of trials, especially, when they are of short duration. In many cases they do not answer the questions an end-user needs answered, because such tests mostly are simplified in comparison to the real-time application and, but still require significant resources. For such cases, this guideline provides other methods for an evaluation of different forecast solutions/vendors.

The pitfalls and challenges with trials and/or benchmarks are the topic of part 2 of this series of recommended practices. Table 1 summarizes some of the aspects and help the decision process as to where and when trials or benchmarks may not be the best choice when selecting a forecast solution. The column "recommendation" in Table 1 provides other methodologies that may be used to evaluate a forecast solution. Additionally, a typical set of questions to be asked to service providers will be provided in APPENDIX A.

Table 1: Recommendations for initial considerations prior to forecast solution selection for typical end-user scenarios

Scenario	**Limitation**	**Recommendation**
Finding best service provider for a large portfolio (> 1000MW) distributed over a large area	Test of entire portfolio is expensive for client and service provider in terms of time and resources. Simplifying test limits reliability of result for entire portfolio.	RFI and RFP, where service provider's methods are evaluated and incentive scheme on the contract terms provides more security on performance.
Finding best service provider for medium sized Portfolio (500MW< X < 1000MW) over limited area	Test of entire portfolio is expensive for client and service provider in terms of time and resources. Simplifying tests limits reliability of result for entire portfolio.	RFP, where service provider's methods are evaluated. Building of a system that enables change of service provider and incentive scheme may be more efficient than a test in the long run. (More detail on incentive schemes are found in section 3.9.3.2 and Part 3 of this guideline).
Finding best service provider for small-sized portfolio (< 500MW)	Test of portfolio requires significant staff resources, a budget and a minimum of 6 months. Difficult to achieve significance on target variable in comparison to required costs and expenses – trial costs makes solution more expensive.	Test is possible, but expensive. Cheaper to setup an incentive scheme and a system in which the suppliers may be exchanged relatively easily.
Micro portfolio (< 100MW) or single plants	Cost of a trial with many parties can easily be higher than the cost of 1 year of forecasting. Time for a trial can delay real-time experience by up to 1 year.	Evaluation of methodologies and setting up the internal system with an incentive scheme and ease of service provider exchange is more beneficial. (More detail on incentive schemes are found in section 3.9.3.2 and Part 3 of this guideline)

Scenario	Limitation	Recommendation
Forecasts will be used to optimize revenue from the sale of generation in power markets	Best evaluation score is difficult to define, as sale is dependent on market conditions and a statistical score like RMSE or MAE cannot reflect the best marketing strategy, considering the uncertainty of a forecast and the associated costs	Strategic choice of forecast provider and incentive scheme better than real-time test. The best choice may be a solution provider that uses different and less correlated input weather forecasts and weather-to-power models, a unique forecast methodology, and/or has greater flexibility and expandable. Employ incentive scheme to motivate performance optimization and continuous performance improvements (see section 3.9.3.2, Part 3).
Market share of potential service provider is high	Monopolies by forecast providers in the power market mean that forecast errors are correlated among generators. This could lead to higher balancing costs. The forecast error might be low, but the costs for errors may be disproportionately high.	Ask about the market share of a provider and do not choose one with a share > 30% as the only provider!
No measurement data available for park or portfolio (“blind forecasting”)	Only useful for portfolios, where small errors are canceled out and indicative regarding performance. Without measurements, forecast accuracy will be non-representative of what accuracy can be achieved by training forecasts with historical data. Evaluation can only be carried out for day-ahead or long-term forecasts, if measurements are collected throughout the trial.	If you have a portfolio > 500MW, a blind test against a running contract can provide an inexpensive way to test the potential of a new provider. For single sites, the benefits of training are so large (>50% of error reduction at times) that blind forecasting is not recommended. It wastes resources for everybody without providing useful results.

Forecast solution components		
Scenario	**Limitation**	**Recommendation**
Prediction of extreme or rare events is important for the application	Today, extreme (or rare) events are better forecasted, when considering weather uncertainty. Statistical approaches relying solely on historic information may not be sufficient. A PoE50 (probability of exceedance of 50%) needs to have equally high probability in every time step above and below. Another critical issue is that a general forecast solution with a single forecast product will not be able to optimally meet the requirements of a extreme event forecast and vice versa.	The IEA Task 36 WP 3 has been dealing with uncertainty forecasting and provides recommendations for such situations. See "Uncertainty Forecast Information" in_Reference Material. Forecasting solution needs to be weather and time dependent, i.e. only physical methodologies (ensemble forecast systems) fulfill such tasks Extreme event forecasting is a component of a full forecasting solution. If extreme events are an important issue a separate forecast that has different optimization and performance attributes is needed.
Prediction of "critical ramp" events is important for the application	Critical ramp forecasts are part of an extreme event analysis and require probabilistic methods with time dependency. A general forecast, especially with a single or a small number of forecasts cannot be used to define critical ramp forecasts, as their optimization strategy usually dampens extremes and will not adequately be able to warn about critical ramps.	Consider difference between a ramp forecast and a critical ramp as extreme event analysis that requires time + space dependent probabilistic methods such as ensemble forecasts. See references for uncertainty forecasts. In general, critical ramp forecasting is a component of a full forecasting solution. If critical ramp forecasts are an important issue, a separate forecast that has different optimization and performance attributes is needed.
Dynamic reserve	Deterministic forecasts cannot solve reserve requirements.	It is necessary to apply probabilistic methods for reserve calculation for intermittent resources such as wind and solar. More information on this topic has been collected by IEA Task 36 WP 3 that has been dealing with uncertainty forecasting. See "Un-certainty Forecast Information" in the Reference Material.

2.1.1 Purpose and Requirements of a Forecasting Solution

Once the limitations are defined, the next step is to define what objectives the project has. As outlined in Table 1, it poses very different forecasting strategies to the project, if the objective is e.g. system balance of renewables or selling generated electricity at the power market.

When designing a forecast solution the first task is to consider extremes and estimate risks; mean error scores are not that important. Large errors are most significant, as they could potentially lead to lack of available balancing power. The second consideration is to look at the uncertainty of the forecast and make sure to choose a forecast that is uncorrelated to others. The mean error of a forecast is important, but not a priority target, if the objective e.g. is to use a forecast that generates low balancing costs. This is not always the same, because errors that lie within the forecast uncertainty are random.

Such errors can only be reduced by strategic evaluations and decisions, not by methodology. If the objective is to calculate dynamic reserve requirements, probabilistic forecasts are required and should be part of the requirement list. When choosing a forecast solution, understanding the underlying requirements is key to the selection the most suitable solution.

It is not enough to ask the vendors for a specific forecast type without specifying the target objective. For this reason, defining the objective is most important. And, if there is no knowledge in the buyer's organization regarding the techniques required to reach the objective, it is recommended to start with a RFI (see section) from different forecast providers and thereby gain an understanding and overview of the various existing solution and their capabilities.

2.2 INFORMATION TABLE FOR SPECIFIC TASKS AND TARGETS

Table 2 lists a number of targets and points to the chapter or part of this guideline series, where the topic is described in detail. The table provides some typical targets and where to find information on how to achieve the best solution for that target.

Table 2: Information table of specific targets

Target	Information
How to find the best forecast solution	Section 3
Creating a requirements list	Section 3.3.1, 2.1.1, 3.2.1, and 3.2.2
Deterministic versus Probabilistic	Section 3.2.1 and 3.9.1
Decision support tool and practical guide to forecasting	Figure 1
Evaluation of vendors: interviewing or conducting trial?	Section 3.9 and References in section
Do I need to test reliability and consistency?	Section 3.2.1 and 3.9.3.1
How do I know which forecast solution fits my purpose best?	Section 2.2 and 3.9.4, APPENDIX A
How do I build up sufficient IT infrastructure for a trial?	Part 2: Trial Execution
Which metrics for what purpose?	Part 3: Evaluation of forecasts
Step-by-step guide for trials and benchmarks	Part 2: Trial Execution

3 DECISION SUPPORT TOOL

From a forecast end-user perspective, it is a non-trivial task to decide which path to follow when implementing a forecasting solution for a specific application. Whether this is at a system operator, energy management company, a power producer or power trader, there are always multiple stakeholders involved in the decision-making process. A relatively straightforward way to decide for one path or another is to use a decision support tool. Figure 1 shows a decision support tool aimed at high- level decisions by managers and non-technical staff when establishing a business case for a forecasting solution. The high-level thought construct shown in Figure 1 is targeted to assist in considering the required resources and involvement of departments and staff for the decision process. The decision tool is constructed to begin with initial considerations to establish a "Forecast System Plan". The tool aims to assist in taking a decision on the major dependencies to the planned item. There are cross references in the decision tool and referrals to different decision streams, dependent on the answer at each step of the decision flow.

The starting point at the top reflects the close and intertwined relationship between a potential forecast solution and the IT infrastructure that is intended to support it. Indeed it may not be possible to implement some aspects of a potential forecast solution (e.g. flow of near real-time data from the generation facilities to the forecast system) if the existing or planned IT infrastructure will not be available to effectively enable it. Therefore, the recommended approach is split based on the status of the IT infrastructure. This is intended to emphasize that there should be a parallel and iterative interaction between the assessment/enhancement of IT infrastructure and the development of specifications for a forecasting solution at the very beginning of the forecast solution selection process. The decision support tool in Figure 2 provides a high-level overview of the process for finding the most suitable forecast solution and vendor. The following sections provide guidance in how to use the decision support tool. There are detailed descriptions and explanation for the more detailed planning and design of the decision process.

Notice for the practical usage: To find the detailed recommendations, the numbering of the boxes in Figure 2 correspond to the headlines in the following sections.

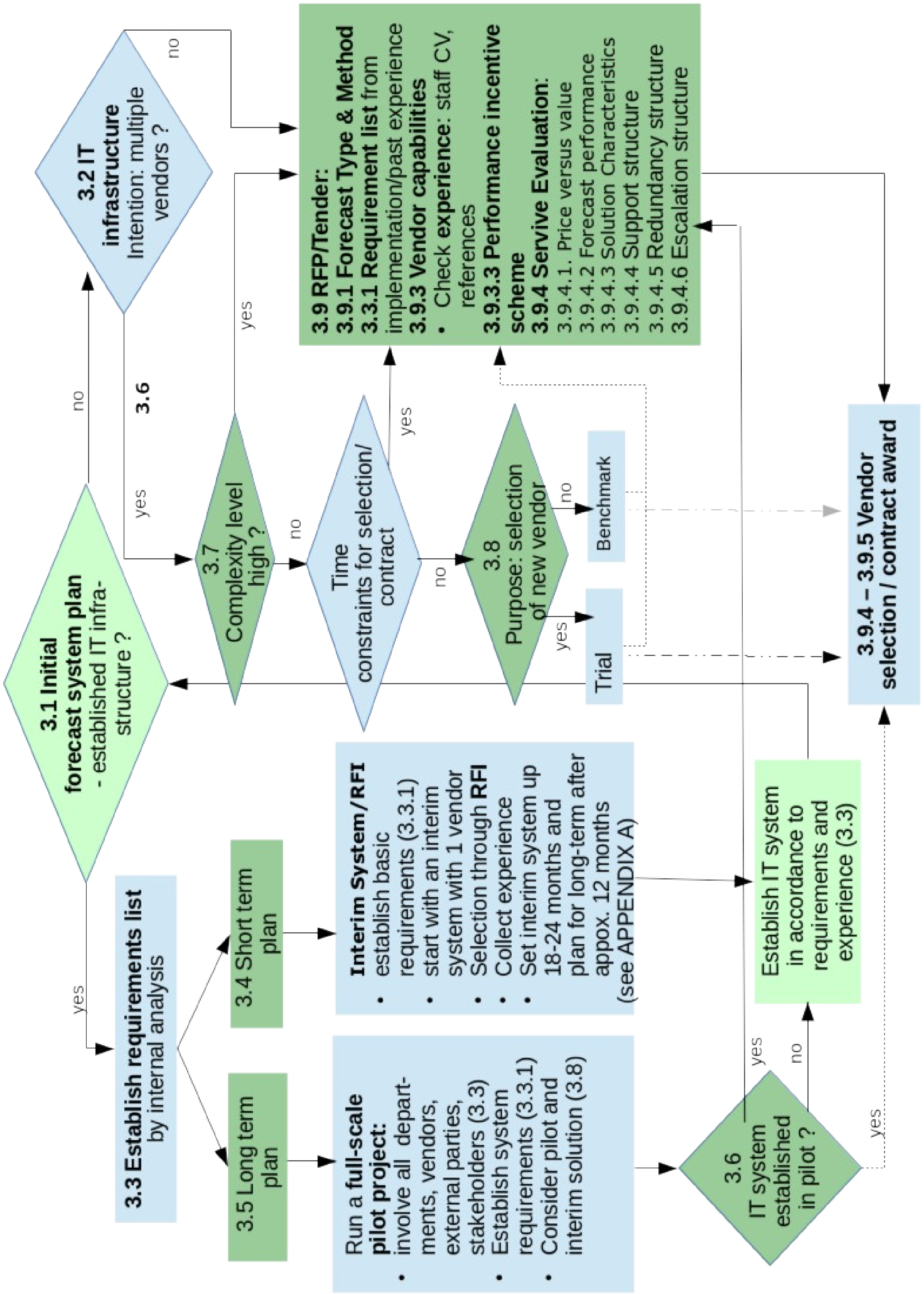
3.1 Initial forecast system plan - established IT infra-structure ?
yes
no
3.2 IT infrastructure Intention: multiple vendors ?
no
yes
3.6
3.3 Establish requirements list by internal analysis
3.4 Short term plan
3.5 Long term plan
3.7 Complexity level high ?
yes
no
Time constraints for selection/ contract
yes
no
3.8 Purpose: selection of new vendor
yes
no
Trial
Benchmark
3.9 RFP/Tender:
3.9.1 Forecast Type & Method
3.3.1 Requirement list from implementation/past experience
3.9.3 Vendor capabilities
• Check experience: staff CV, references
3.9.3.3 Performance incentive scheme
3.9.4 Servive Evaluation:
3.9.4.1. Price versus value
3.9.4.2 Forecast performance
3.9.4.3 Solution Characteristics
3.9.4.4 Support structure
3.9.4.5 Redundancy structure
3.9.4.6 Escalation structure
3.9.4 – 3.9.5 Vendor selection / contract award
Interim System/RFI
• establish basic requirements (3.3.1)
• start with an interim system with 1 vendor Selection through RFI
• Collect experience
• Set interim system up 18-24 months and plan for long-term after appox. 12 months (see APPENDIX A)
Run a full-scale pilot project:
• involve all depart-ments, vendors, external parties, stakeholders (3.3)
• Establish system requirements (3.3.1)
• Consider pilot and interim solution (3.8)
3.6 IT system established in pilot ?
yes
no
yes
Establish IT system in accordance to requirements and experience (3.3)

Figure 2: Decision support tool.

3.1 INITIAL FORECAST SYSTEM PLANNING

The planning of a forecasting system for wind and solar power is a complex task and highly individual. This guideline therefore focuses solely on aspects that are of general planning and management tasks specific to the implementation of wind power or solar power production forecasts into an operational environment.

Note that any information and considerations about forecast technologies or methodologies here has the sole objective to provide guidelines on the impacts of commonly implemented technologies for decision processes, not a recommendation for or against any technology.

There is strong focus on the IT infrastructure as one of the most crucial tasks in the implementation and integration of forecast solutions that are prone to become limiting factors for changes at later stages. For that reason, it is recommended that the IT infrastructure is established or, if already available, evaluated together with the planning of the forecast solution and methodology. Especially the IT solution's ability to develop along with changes in forecast practices, possible statutory changes among others are important aspects to consider. Databases are another aspect to consider, as they are prone to have limitations that prevent changes to incorporate more information or store information in a different way. Such consideration need to take place and should be part of the decision process and the requirement list (see section 3.3).

3.2 IT INFRASTRUCTURE CONSIDERATIONS

The starting point of the tool is the IT infrastructure. If a company has already built an appropriate infrastructure, finding a forecasting solution or a vendor for a specific forecasting solution is less complicated. The reason for this is that in this case, the forecast provider will need to conform to file formats, communication protocols or security constraints, for example. If an IT infrastructure for the forecasting solution is to be established or renewed it should be formulated to efficiently accommodate the technical requirements of the solution. If no IT infrastructure has been built yet, an internal analysis of the needs are required. In this analysis, it is important to know, whether there is a short-term goal with an objective to be reached with time constraints, or whether it is a long-term plan that needs to be satisfied.

The important aspects in the IT infrastructure to be considered are:

- database structure
- communication layer
- monitoring and error handling
- data storage and historic data accessibility

In general a forecast system interface, whether in-house or outsourced requires multiple data streams, starting from measured power and weather variables. Usually, there is a connection to the power unit’s SCADA (Supervisory control and data acquisition) system. However, the measurement data needs storage and a data flow of measurements and other production data from the power plants to the forecaster needs to be added as one more of the various internal data flow processes.

It needs to be decided whether there is a need to access other external data sources, such as NWP data, or the forecast data itself.

Dependent on the setup of the forecasting solution, it is also necessary to evaluate how fast accessible historic data has to be, for example to carry out internal analysis, external data delivery to vendors, etc.

3.2.1 IT impacts for single versus multiple forecast vendors

Impacts on multiple vendor solution:

- infrastructure more complex
- database requirements are higher due to higher data volumes
- strategy required for forecast: mixing versus primary/secondary forecast

IT infrastructure impacts for single vendor solution:

- reliability requirement of solution high
- monitoring requirement higher for up-time
- higher requirements for quality control of forecasts
- less data volume than for multiple-vendor solutions
- database structure less complex than for multiple-vendor solutions

3.2.2 IT requirements for deterministic versus probabilistic forecasts

From an IT infrastructure and architectural perspective, deterministic and probabilistic forecasting solutions are quite different. The database requirements are by a factor of 10 to 100 higher for the latter. Dependent on the way the probabilistic forecasts are used, they add significant amounts to the storage requirements.

Nevertheless, storage and computational resources are changing with changing requirements in industry and hence should not per se be considered a barrier or limitation for the integration or implementation of new technologies. But, they need consideration and careful planning.

The advantages and disadvantages of the deterministic versus the probabilistic solution from a IT perspective are similar to single versus multiple providers in section 3.2.1.

3.3 ESTABLISHMENT OF REQUIREMENT LIST

Establishing a requirement list for a forecasting solution is highly individual and depends on many factors, such as internal requirements and external offerings. Every end-user will have very specific needs to fulfill. There are however common areas that require consideration. This is how the recommendation list in 3.3.1 has to be interpreted.

Two of the fundamental aspects when establishing a requirements list are:

<u>1. Description of the current situation</u>

> In this process, it is imperative to describe exactly all processes, where forecasting is required and how these processes are interlinked. Here it is essential to get the different departments involved, also the IT department. The more accurate you can describe the situation at hand, (e.g. integration plans, use of forecasts, market situation, statutory aspects, IT restrictions, limitations and methods for data exchange exist, current or future challenges, etc.), the more straight forward it will be to (1) ask questions to the forecasting vendors regarding forecasting methodology, but also (2) get clarity of the involved processes enabling forecasting, (3) provision of liabilities and guarantees.

2. Engage forecast vendors, stakeholders and independent consultants

Questions to vendors should be of technical character regarding forecast methodology, but also on available data exchange methodologies, required input data for the models and system support.

If you already have a forecast vendor, it is recommended to engage with the forecaster to discuss the current situation and where the forecaster sees limitations and potential for improvements. Often, forecast providers need to adopt their forecasts to a specific need and even though a new technology may be available, it is not used due to current limitations.

Other vendors, stakeholders and independent consultants may at any stage be engaged, not only when it comes to establishing a new, or renewal of, a forecasting system. For new systems, it is recommended to engage different forecast vendors and stakeholders to provide insight from a variety of experiences.

In all cases, it is essential to describe the planned objective and name limitations, if they are already known. The more information that can be shared the better a vendor, stakeholder or consultant can evaluate what is considered the most appropriate solution.

APPENDIX A contains an additional listing of recommended considerations that are applicable also for RFI's.

3. Description of the envisaged Situation

The description of the envisaged situation is most important for the implementation of a solution. Analysis of the current situation, the forecast vendor(s) input and other organizational and statutory requirements should lay the basis for an envisaged new system. It is recommended to put as much detail into this part as possible. The following requirement list assists in defining all aspects for the planning phase of a forecasting system.

Recommendation in short: Describe (1) the current situation, (2) engage vendors and stakeholders and (3) describe the envisaged situation in great detail. Ask specific questions that are required to get the highest possible level of detail for the decision process.

3.3.1 Requirement List

The following areas are recommended to be considered in the list:

IT infrastructure

- communication/data exchange with the forecast vendor(s)
- communication/data exchange with the asset operation (wind/solar parks)
- database and storage implications
- accessibility of data/information of internal users
- application interfaces to internal tools (e.g. graphics, models, verification, metering)
- information security policies

Forecast Methodology and Attributes

- Specification of weather inputs used by solution provider
- Specification of methods used in weather to power model
- Specification of data/methods used to produce each forecast product
- Forecast time horizons
- Forecast frequency
- Forecast uncertainty

Support and Service

- service level for each product (e.g. 24/7, business hours etc.)
- system recovery
- failure notifications and reporting
- escalation procedures
- service documentation
- contact list for different services
- staff training

Contracting

- contract length
- amendment possibilities
- additional work outside contract
- licenses
- confidentiality (NDA)
- insurances
- sub-contracting
- Price table for each product category

Performance and Incentivization

- Verification methods
- Verification parameter
- Definition of incentive payment structure (e.g. payment/no payment or partial payment)
- Expected accuracy for each forecast horizon

3.4 SHORT-TERM SOLUTION

In the case of a short-term solution, current requirements should be listed and analyzed in accordance with possible time limitations. It is recommended that a short-term solution is sought, if the country's current policy does not seem to be stable to make long-term investments, or a here-and-now issue needs to be solved and experience gained. In such cases, a relatively simple methodology that can be implemented fast and easy is the best way forward.

Today, this can be found by carrying out a RFI, where vendors can suggest how to best and easiest fulfill very specific needs. Due to IT constraints in many organizations, such solutions sometimes are set up with delivery by Email. This is not a recommended practice for security and reliability reasons, but can help to fill a gap between a long-term solution and an urgent need.

Despite the shortcomings, interim solutions are recommended as they are valuable in respect to experience with forecasting data and it's handling inside the organization. If such solutions are employed while a long-term plan is being developed, it can be of great benefit for the long-term solution. Such solutions should last approx. 18-24 months. Planning for a long-term solution should ideally start after 12 months.

Staying with an interim solution can bare disadvantages for the forecast user, if it has real limitations on security (e.g. email delivery) and reliability, as such limitations may not be problematic for a long time, but reliance on non-redundant systems can cause sudden uncontrollable situations arising from missing forecasts of wind and solar power generation.

For this reason, we posted the question about the IT system (see also Figure 1) at the end of the short-term solution, as this is a crucial part in the next step. We recommend that this is taken as a priority topic, once practical experience with forecasting has been gained.

3.5 LONG-TERM SOLUTION

Developing a long-term solution can be cumbersome and difficult, as many aspects have to be considered, from policies to governmental plans or corporate strategies.

A practical way forward is to conduct a full-scale pilot project, where different solutions are tested and verified over a period of at least 1 year.

The advantage of such a pilot project is that there is the possibility to verify and evaluate different solutions and their fit for purpose over a longer time span.

Moreover, a pilot project is characterized by:

- Participation of all relevant internal and external stakeholders
- Iterative establishment and validation of solution requirements
- Possible use as an interim solution

The disadvantage is that it takes a long time and hence is costly and it is not given that there is a very clear winning solution to a specific area or task. On the other hand, to find the most appropriate long-term solution needs many considerations, not only technically, but also economically and whether a solution is future compatible, i.e. capable of solving growing capacities and requirements expected to become part of the solution at a later stage. So, the experience of the vendor in adjusting, maintaining and developing a solution with changing needs may be a challenge for some and the business philosophy for others. Such vendor policies can be identified and clarified when carrying out long-term tests. The box therefore feeds into the question about an appropriate IT system. If this has not been established, it is recommended to prioritize the IT before going further.

The end of a pilot project has therefore 3 further paths:

(1) vendor selection
(2) redefining requirements to start a solution bottom up
(3) carrying out a RFP with the identified requirements.

3.6 GOING FORWARD WITH AN ESTABLISHED IT SYSTEM

In the case an IT system has been established and new vendors or a renewal of the system is the objective for the project, there are various possibilities to move forward.

Crucial in this phase is again to set target and objectives. If the target is to find out, whether there exist forecast vendors on the market that may provide forecasts with other methods or for a lower price, it may be a good way forward to carry out a trial or benchmark.

Dependent on the structure of the system, or complexity of the system and time constraints, a benchmark/trial or a RFP as alternative are recommended. One crucial criterion when deciding on the two alternatives RFP or trial/benchmark in existing IT environments is whether the IT structure can handle multiple suppliers.

If this is not the case, any evaluation against an existing supplier can be cumbersome and at times impossible. The recommended practices guideline part 2 is going into detail with the topic of evaluations being:

- **representative** (including consistency)
- **significant** (including repeatable)
- **relevant** (including fair and transparent)

These are the key points when carrying out a comparison.

3.7 COMPLEXITY LEVEL OF THE EXISTING IT SOLUTION

Apart from accuracy or statistical skills of forecasts, there are also other aspects to be considered when choosing a forecast supplier. It has been observed that such evaluations based on non-technical skills or skills leading to forecast performance for a specific purpose have been underestimated in their importance. One of these aspects is the ability to improve, which is fully excluded with a trial/benchmark as sole decision-making criterion (besides price) as capability of vendors. It is often forgotten that long-term experience in a specific area can provide significant advantages. And, verifying only a small part of a complex system for practical reasons may result in misleading results (see 3.63.6 "representative", "significant" and "relevant").

Complex systems are seldom easy to simulate in trials and will always disqualify some participants, when it comes to the selection process. To conclude, the complexity of a system and the purpose of a forecast within a complex corporate structure are significant aspects to consider in a forecast solution selection.

Recommendation: The path to follow in case of complex structures and requirements are best performed by a RFP process, where core capabilities should be evaluated, when choosing a forecasting solution.

3.8 SELECTION OF A NEW VENDOR VERSUS BENCHMARKING EXISTING VENDOR

If there are no time constraints and the complexity level of the running system is not too high, or a new system is in the process of being built, a trial or a benchmark exercise can be very useful.

Recommendation: Conduct a trial in case a new vendor has to be selected and a trial can be carried out in such a way that the results are fair, transparent, representative and significant. Carry out a benchmark, if the purpose is not from the outset to engage a new vendor, but also to compare the capabilities of a vendor with other vendors or against newer technology. In both cases the invited vendors need to be notified of the purpose of the exercise.

3.9 RFP EVALUATION CRITERIA FOR A FORECAST SOLUTION

If complexity levels are high and if time constraints do not allow for a lengthy trial or benchmark, the RFP should be compiled with care in order to fulfill all requirements and yet not ask for more than needed.

The most important evaluation criteria for a forecast solution to be defined in a RFP is:

- the type of forecast that is required (e.g., hours-, day-, or week-ahead)
- suitability of available methods for optimally satisfying the forecast requirements
- compliance to requirements

It is recommended that this first step should be vendor independent. And, if this cannot be defined, it is recommended to first conduct an RFI to scan the industry on their capabilities and their recommendation which type and methodology should be applied for the specific needs. APPENDIX B contains typical questions for an RFI.

Only when the forecast type and methodology is defined, the vendor comes into play. The additionally important factors to consider here are:

- capabilities (experience)
- support and maintenance services

The sections below describe these considerations in detail.

3.9.1 Forecast Type and Methodology

Most users will agree that they want to obtain forecasts with the best possible forecast accuracy for their application. A benchmark or a trial has in the past often been viewed as a way to determine which provider is most likely to deliver the best possible forecast performance. In theory, this is a reasonable objective. In practice, it is not recommended to rely solely on a test.

The following subsections will address a number of key issues associated with the dilemma of finding the best forecasting solution with a simple and non-costly exercise for both the end-user and the forecast provider.

3.9.1.1 Forecast solution Type

Single versus multiple forecast providers

It has been widely documented (e.g. Nielsen et al., 2007, Sanchez, 2008) that a composite of two or more independent state-of-the-art forecasts will often achieve better performance (accuracy) than any of the individual members of the composite over a statistically meaningful period of time. Indeed, many of the FSPs internally develop their approach and services on that basis. And, there are well founded reasons for an end-user to consider the use of multiple FSPs to achieve better forecast accuracy. However, in a practical sense, there are several advantages and disadvantages that should be considered. When building up a solution, it is recommended to consider the following aspects:

Benefits of using multiple vendors

(1) There are a number of FSPs in today's forecast market that exhibit performance that is close to the state-of-the-art. It may be advantageous for reliability to assemble a set of state-of-the-art forecasts, unless they are highly correlated.

(2) Higher forecast accuracy can often be achieved by blending forecasts from multiple uncorrelated[1] FSPs.

[1] Uncorrelated forecasts here means ideally that both the underlying weather information and weather to power conversion model is not the same. At least one part must be different, where the weather input has more weight.

Drawbacks of using multiple vendors

The benefits of having multiple vendors also contain inherent challenges for the end-user:

(1) Increased internal costs, even if two "cheap" vendors may be less costly than one high-end forecast vendor, employing multiple vendors increases internal costs significantly due to increased amounts of data and IT processes.

(2) Blending algorithms need to be intelligent. Multiple forecasts can be beneficial, but only if the algorithm is intelligent to only blend/mix in case of all forecasts being available and easy to retrain when forecast statistics change. With two forecast vendors this is relatively easy. If there are more than two, it becomes more complex.

(3) Forecast improvements are difficult to achieve with a multi-forecast provider solution. When improvements are achieved on the vendor side, the blending algorithm is becoming inconsistent and can result in worse scores than before, unless long-term historic data can be delivered. In other words, the handling and the improvement of forecasts are complex and difficult with multiple forecasts.

(4) Multi-vendor Solutions cannot be incentivized as easily to achieve continuous performance increase over time. Although incentive schemes can be a good way to provide resources to the FSP for continuous improvements, in a multi-vendor environment, this can be counter productive, as changing statistical characteristics of forecasts can have a bad influence on the resulting blended forecast. Any end-user needs to be aware of this pitfall, when choosing a solution and take mitigating measures.

(5) Multiple points of failure - with multiple forecast providers, the IT infrastructure needs to contain more logic to deal with one or more data streams when there are, for example, delivery disruptions, timeliness, or quality issues.

3.9.1.2 Deterministic versus Probabilistic

Many forecasting tasks need a discrete answer. For that reason forecasting solutions have been mostly fed with deterministic forecasts in the past. Although weather forecasts and hence also power forecasts of intermittent resources such as wind and solar power, contain inherent uncertainties, probabilistic forecast products have been associated with forecasts not being discrete. The probability of an generic power generation at time x cannot be used in a trading application with the purpose to bid into the market.

As penetration of variable generation resources increase and digitialization increases, the uncertainty information for decision taking can and is being processed by algorithms, also those whose output needs a discrete answer. Deterministic forecasts by default suppress the underlying uncertainty in the forecasts. By using probabilistic forecasts this uncertainty can be taken into consideration in the decision processes.

The most common products of uncertainty or probabilistic forecasts are the probability of exceedance (PoE) values, typically given as PoE05, PoE50 and PoE95, quantiles, or percentiles or confidence bands (see Glossary for definitions).

The advantage of probabilistic/uncertainty forecasts in comparison to the deterministic "best guesses" is the possibility to act upon the probability of an event to occur, rather than being surprised, when the deterministic forecast is wrong. In power markets, for example, a probability of exceedance of 50% (PoE50) is an important parameter for a system operator, as such forecasts prevent the market to be able to speculate against system imbalance. Extreme ramping, high-speed shut-down risk, unit commitment and dynamic reserve allocation are other examples, where probabilistic forecasts are beneficial or required. In other words, wherever there are some kind of uncertainty and extreme to be considered that may have impact on a decision or the costs of a process, probabilistic forecasts provide the necessary information to an end-user to take a decision upon some objective uncertainty criteria.

Recommendation: When establishing or renewing a forecasting system, the question should not be posed on advantages and disadvantages for deterministic or probabilistic forecast solution, but rather whether a deterministic solution can fulfill the objective of the application.

Information about probabilistic methodologies can be found in the References Material under "Uncertainty Forecast Information", especially in a review on probabilistic methods for the power industry (Bessa et al. (2017)) .

3.9.2 Forecast horizons

The forecast horizons play a major role in the ability to plan using forecasts. Today, there are 5 types of forecast horizons applied in the power industry:

1. Minute-ahead forecasts or nowcasts (0-120min)
2. Hours-ahead forecasts (0-12 hours)
3. Day-ahead forecasts (0-48 hours)
4. Week-ahead forecasts (0-168 hours)
5. Seasonal forecasts (monthly or yearly)

The **Minute-ahead forecasts** are in literature also sometimes referred to as *ultra-short term forecasts or nowcasts* and are mainly used in areas with high penetration and high complexity in system operation or significant risk for high-speed shut down and extreme events. These forecasts are either based on a statistical extrapolation of measurements or weather input together with measurements generated on minute basis.

The recommended practice depends on the severity and costs of the target value. For situational awareness, a simple extrapolation of measurements may be sufficient. For extreme events (e.g. ramps, high-speed shut down) the involvement of weather related forecasts in high time resolution is recommended.

Hours-ahead forecasts, or sometimes referred to as s*hort-term forecasts,* correct a day-ahead forecast by using real-time measurements and extrapolate from local real-time observations an improved view of the current state and the next few hours.
There are different methods available from simple extrapolation of measurements to advanced weather and distance- dependent algorithms. It's recommended to get details of a short-term forecast methodology described by the vendors, as quality and usability can differ strongly with availability of data, quality of measurement data etc.

If the target is e.g. ramp forecasting, system control, a very large fleet or quality issues with measurement data not dealt with by the end-user, simple algorithms are often not capable of providing a sustainable picture of the next few hours.

The **Day-ahead forecasts** are widely-used forecasts for general system operation, trading and short-term planning. Traditionally, they are based on a combination of weather models and statistical models.

The **Week-ahead forecasts,** *sometimes referred to as long-term forecasts,* are usually applied in cases where the focus is not on forecast accuracy, but on forecast skill, e.g. in situations, where trends prevail over granularity. These forecasts are most valuable as a blending of a number of different forecasts or from an ensemble predication system, where the small-scale variability is reduced. If this is done, such forecasts can serve to reduce reserve costs and generate more dynamic reserve allocation as well as auctions.

The **Seasonal forecasts** *sometimes referred to as ultra-long-term forecasts,* predict variations due to seasonal and or climate variability. They may be derived based on climatology, correlation to various climate indices and oscillatory phenomena, climate models, or a combination of these methods. Ensemble methodologies are the most preferable method due to the inherent uncertainty on such time frames. The most simple method is to analyze past measurements.

Recommendation: Key when choosing a methodology is to carefully analyze the accuracy requirements of the task to solve. For trading of futures in a trading environment a simple methodology may be sufficient. Tasks such as grid balancing, grid infrastructure planning or long-term capacity planning however require more advanced methodologies. It is recommended to choose the method according to the need to capture quantities only (simple method) or capture also climatic extremes (advanced method).

3.9.3 Vendor Capabilities

3.9.3.1 Experience and Reliability

Experience is a key element of a successful vendor and implementation of the forecasting solution. It can usually be evaluated by the selected references that are provided and measured by conducting interviews with customers of similar type or by asking for information about the vendor’s background and experience with similar customers. If a vendor is new to the market that may not be possible. In this case, staff resources and experience of the key staff is usually indicating, whether the experience level for the minimum requirements is given.

Reliability is also connected to experience, as it implies the reliable implementation and real-time operation of a forecasting service. It is an important aspect and may be derived by requiring examples of similar projects and interviewing references. It can also save a lot of work and resources in comparison to carrying out a trial, if reliability and experience with respect to e.g. complex IT infrastructure, security aspects, reliable delivery and provision of support etc. are a more crucial aspect than specific statistical performance scores.

Recommendation: Ask vendors to describe their experience and provide references and CV of key staff members.Ability to maintain state-of-the-art performance

The previous section provided an overview of all of the considerations for the technical aspects of forecast type and methodology.

In order to assure that the forecast vendor can maintain state-of-the-art performance it is recommended to verify, whether the provider engages in ongoing method refinement/development and forecast improvement activities.

Recommendation: Evaluate by asking the vendor to provide information about:
 * research areas and engagement
 * references to staff publications of e.g. their methodology, project reports
 * references of participation in conferences/workshops
 * percent of revenue reinvested into research and development

3.9.3.2 Performance incentive Schemes

A performance incentive scheme is the most effective way to ensure that a forecaster has an incentive to improve forecasts over time and also allocates resources to it. By setting up a performance incentive scheme, the client acknowledges that development requires resources and vendors have not only an economic incentive to allocate resources for further developments, but can also influence their reputation. Incentive schemes do not have to be enormously high, but usually range between 10-30% of the yearly contract sum.

Establishing a performance scheme

What must be key to a performance incentive scheme is that it reflects the importance of the forecast parameters that are incentivized for the client!

The evaluation of such forecast parameters should be selected according to:

1. the objective of the forecasting solution
2. the use/application of the forecasts
3. the available input at forecast generation time

The **objective (1)** in this context is defined as the purpose of the forecast. For example, if a forecast is used for system balance, an evaluation should contain a number of statistical metrics and ensure that there is an understanding of the error sources that the forecaster can improve on. A typical pitfall is to measure performance only with one standard metric, rather than a framework of metrics reflecting the cost or loss of a forecast solution. For example, if a mean absolute error (MAE) is chosen to evaluate the performance in system balance, an asymmetry in price for forecast errors will not be taken into account. Also, if e.g. large errors pose exponentially increasing costs, an average metric is unsuitable.

The use or **application of forecasts (2)** is defined in the context of where forecasts are used in the organization and where these have impact and influence on internal performance metrics or economic measures. For example, a wind power forecast that a trader uses for trading the generation of a wind farm on a power market has two components: revenue and imbalance costs.

The revenue is defined by the market price for each time interval, whereas the cost is defined by the error of the forecast, the individual decision that may have been added to the forecast and the system balance price. When evaluating a forecast in its application context, it is important to choose an evaluation that incentivizes the vendor to tune the forecast to the application. A forecast that is optimized to avoid large errors may create lower revenue. However, if income is evaluated rather than revenue, such a forecast may be superior due to lower imbalance costs. On the other hand, if the end-user makes changes to the forecast along the process chain, the forecast evaluation must stop, where it is outside the forecast vendor’s influence.

The available input at **forecast generation time (3)** is most important when evaluating short-term forecasts that use real-time measurements. For example, if the forecast is evaluated against a persistence forecast with corrected measurements rather than with the measurements that were available at the time of forecast generation, the evaluation is to the disadvantage of the forecaster. The same applies, if aspects that affect the forecast such as curtailments, dispatch instructions, turbine availability, are not taken out of the evaluation or are corrected.

Recommendation: When incentivizing a forecast solution with a performance incentive, the evaluation need to consider the non-technical constraints in the forecast and the parts that a forecaster does not have influence upon. A fair performance incentive scheme needs to measures the performance of a forecast by blacklisting any measurement data that is incorrect or corrupt, that contains curtailments, dispatch instructions, reduced availability or other reductions outside of the forecasters influence. Evaluation against persistence forecasts also need to be done with the available data at the time of forecast generation to not give advantage to persistence. Additionally, single standard statistical metric (e.g. MAE or RMSE) alone cannot be recommended.

More details on the purpose and interconnection of statistical metrics for evaluation of incentive schemes can be found in part 3 of this recommended practice and in the references under "Evaluation and Metrics".

Structure of a performance incentive payment

The structure of performance incentive scheme is an individual process and contractual matter between parties.

When establishing the structure of a performance incentive it is recommended to consider that by choosing a maximum and minimum, the maximum value provides budget security to the end-user, also when e.g. changing from a very simple solution to an advanced one with much higher performance. The latter provides security to the forecaster to ensure that the basic costs for generation of forecasts are covered.

Adding a sliding structure in between ensures the forecaster always has an incentive to improve, also when it is foreseeable that the maximum may not be achievable.

Recommendation: it is recommended to apply a maximum incentive payment and a maximum penalty or minimum incentive. A sliding change is preferable over for a boolean (yes|no) decision for incentive payments, as it always encourages forecast improvement efforts.

3.9.4 Evaluation of services

The recommended practice in any evaluation is to consider a number of factors that contribute to the value that a user will obtain from a forecast service. It is not possible to provide a complete list of factors to consider.

However, the most important factors that should be addressed are the following elements:

- Price versus value and quality
- Forecast Performance
- Solution Characteristics
- Speed of delivery
- Support structure
- Redundancy structure

The issues associated with each of these aspects will be addressed in the following subsections in more detail.

3.9.4.1 Price versus Value and Quality

The value of a forecast may or may not be directly measurable. In most cases however, the value can be defined for example in terms of cost savings or obligations and in that way provide an indication of the expected value from a certain solution.

Prices are difficult to evaluate. A low price often indicates that not all requirements may be fulfilled in operation or not all contractual items are accepted and left to the negotiations. For these reasons, care has to be taken in the evaluation process.

Some services and methods are more expensive than others on e.g. computational efforts, required licenses, database requirements, reliability, etc. Unless prices are driven by competition in a overheated market, a service price is normally coupled to the requirements and acceptance of contractual items. Some items such as reliability, customer support or system recovery can have high prices, but can always be negotiated to a different level. In an RFP end-users need to be aware of the relation between cost, value and associated service level to prevent vendors from speculation on negotiable item in the requirement list.

Recommendation: Following a decade of experience in the forecasting industry, the recommended practice on price evaluation is to connect technical and contractual aspects to the price and consider to let vendors detail contractual aspects that may be associate with high service costs separately, especially, if a fixed cost price is requested.

An example could be the requirement of full system recovery within 2 hours in a 24/7/365 environment. If there is no penalty associated, a vendor may ignore this requirement, which may result in a much lower price.

Requesting transparent pricing eases evaluation and makes sure that speculations regarding negotiable aspects of a service can be clearly compared.

3.9.4.2 Forecast Performance

Forecast performance evaluation should contain a number of metrics that are representative for the need to the forecast user. It is recommended to establish an evaluation framework for the performance evaluation. How to establish such a framework is dealt with in Part 3 of this recommended practice guideline.

3.9.4.3 Solution Characteristics

The solution characteristics of a forecast service also contains much value for an end-user and should get attention in the evaluation. It can be defined in terms of the available graphical tools, ease of IT services for retrieving data or exchanging data in real-time as well as historical data, customer support setup and staff resources connected to the forecasting solution.

This can be key for the operational staff to accept and be comfortable with a forecast service as well as having confidence in the service. Additional work that may be connected, but outside the scope of the operational service can also be key elements for a well functioning service.

Recommendation: Ask the vendor to describe how the system will be built up, how communication and support is envisaged and let them provide examples of graphics (if applicable).

3.9.4.4 Support Structure

Customer service is often under-estimated and in most cases second to an accuracy metric when selecting a vendor. Support can be a costly oversight if, for example, costs are related to a continuously running system or extreme events, where the user needs an effective warning system and related customer service. Support can have a relatively large cost in a service contract and may provide a false impression on service prices, if, for example support is only offered at business hours.

Recommendation: Definition of the required support structure should be part of the requirement list for any forecasting solution. For real-time forecasting solutions end-user need to ensure that there is an appropriate support structure in place. Considerations of the real-time environment, own resources and which of the forecasting business practices are of significance to the user should be carried out. Especially, where processes are supposed to run every day in the year.

Key elements for the customer support is:

- the responsiveness of the provider, when issues arise
- live support in critical situations

A support structure and its management for operational processes additionally need to bind the following strategic areas together:

(a) Customer Support
(b) Operations Software and Service
(c) IT Infrastructure

The customer support (a) should be handled by a support platform, ideally with different forms for contact, e.g. telephone hotline and email ticket system.

Any end-user needs to ensure that third-party software used in the operational environment (b) is licensed and renewed and maintained according to the licensing party’s recommendations.

The IT infrastructure (c) should ideally be ISO 9001 and ISO 27001 certified in cases, where real-time operation and security is of paramount importance.

3.9.4.5 Redundancy Structure

Redundancy depends very much on the end-users needs to maintain a frictionless and continuous operation. Forecasting is mostly carried out in real-time, which has an inherit requirement of being functional all the time. While there are many processes and targets for forecasting that may not require large redundancy and permanent up-time, the following recommendation is targeted to those end-users where forecasting is to some extend mission critical.

There are a number of different redundancy levels that need consideration and that can be achieved in various ways:

(1) Physical delivery of the service via IT infrastructure
(2) Content of the delivery via Forecasting methods

The delivery of the service (1) is connected to the IT infrastructure. Redundancy measures may be a combination of any of these:

- ➔ Delivery from multiple locations to mitigate connectivity failures
- ➔ Delivery from multiple hardware/servers to mitigate individual server failure
- ➔ Delivery with redundant firewalls to mitigate hardware failure
- ➔ Delivery through a ISP using Email, etc.

The redundancy of the forecast content is equally important as the physical delivery of the data, but often neglected.

It is recommended to consider any combination of the following redundancy measures for correct forecast content:

- ➔ redundant providers of weather input
- ➔ redundant/multiple providers of forecast service
- ➔ redundant input and mitigation strategy for weather models
- ➔ redundant input and mitigation strategy to power conversion models

Recommendation: Define the required redundancy level according to the importance of a permanent functioning service and the impact of delivery failure to other internal critical processes.

3.9.4.6 Escalation Structure

It is recommended for high-level contracts, where forecasting is critical to the end-users processes to get information about escalation structures in case of failure. This is especially important when employing only one forecast provider.

Recommendation: An end-user needs to have a description about structure and corresponding responsibilities for their operations staff in order to incorporate such information into own escalation structures in case of emergencies.

Table 4: Recommendation of a three tier escalation structure.

Escalation Level	Forecast service providers coordination	End-user side coordination
Level 1: failure to deliver service	Technical Staff	Operations Staff Project manager
Level 2: failure to recover or implement service	Project manager	Project manager Department manager
Level 3: failure to solve failure/recovery	General management	General management

Each level of escalation ideally contains the following structured process:

- Formulation of the problem/failure
- Root cause analysis
- Coordination of action plan for troubleshooting inclusive responsibilities
- Coordinated action plan progression
- Escalation to the next level or closure of escalation procedure

4 FINAL AND CONCLUDING REMARKS

While every forecasting solution for wind and/or solar power generation contains very individual processes and practices, there are a number of areas that all forecasting solutions have in common. For any industry it is important to establish standards and standardized practices in order to streamline processes, but also ensure security of supply with a healthy competition structure.

This document is providing state of the art practices that have been carefully collected by experts in the area and reviewed by professionals and experts in an appropriate number of countries with significant experience in wind energy forecasting. The recommendations are to encourage both end-users and forecast service providers to bring focus to areas of practice that are common to all solutions. The document will be updated as the industry moves towards new technologies and processes.

The key element of this recommended practice is to provide basic elements of decision support and thereby encourage end-users to analyze their own situation and use this analysis to design and request a forecasting solution for wind and/or solar power generation that fits their own purpose rather than applying a “doing what everybody else is doing”-strategy.

This document is also intended to serve forecast service providers new to the market or those wanting to evolve to a new level of service and support as a guideline to state of the art practices that should be incorporated into business practices.

References Material

NOTE: Access to references at IEA Wind Task 36 webpage: http://www.ieawindforecasting.dk

Forecast solutions, Trials and Benchmarks

IEA Wind Task 36: *Recommended Practices Guideline for the Implementation of Wind Power Forecasting Solutions Part 2: Designing and executing forecasting bench-marks and trials.* Online access: http://www.ieawindforecasting.dk

Corinna Möhrlen, John Zack, Jeff Lerner, Aidan Tuohy, Jethro Browell, Jakob W. Messner, Craig Collier, Gregor Giebel, *Recommended Practices for the Implementation of Wind Power Forecasting Solutions Part 1: Forecast Solution Selection Process*, Proc. 17th Int. Workshop on Large-Scale Integration of Wind Power into Power Systems, Stockholm, Sweden, October 2018.
Online Access: http://download.weprog.com/wiw18-133_recommended-practice_selection-process.pdf

C. Möhrlen, C. Collier , J. Zack , J. Lerner , *Can Benchmarks and Trials Help Develop new Operational Tools for Balancing Wind Power?,* Proc. of 16th International Workshop on the Large-Scale Integration of Wind Power into Power Systems, Paper WIW-292, Berlin, Germany, 2017. Online access: http://download.weprog.com/WIW2017-292_moehrlen_et-al_v1.pdf

Evaluation and Metrics

IEA Wind Task 36: *Recommended Practices Guideline for the Implementation of Wind Power Forecasting Solutions Part 3: Evaluation of forecast solutions.*
Online access: http://www.ieawindforecasting.dk

C. Möhrlen, C. Collier , J. Zack , J. Lerner , *Recommended Practices for the Implementation of Wind Power Forecasting Solutions Part 2&3: Designing and executing forecasting benchmarks and trials and evaluation of forecast solutions,* Proc. of 16th International Workshop on the Large-Scale Integration of Wind Power into Power Systems, Paper WIW-160, Berlin, Germany, 2017. Online access:
http://download.weprog.com/wiw18-160_recommended-practice_benchmark-evaluation.pdf

Anemos.Plus Project DELIVERABLE D-1.3 (), *Towards the definition of a standardised evaluation protocol for probabilistic wind power forecasts*. Online available: http://www.anemos-plus.eu/images/pubs/deliverables/aplus.deliverable_d1.3-protocol_v1.5.pdf

Gensler, André & Sick, Bernhard & Vogt, Stephan. (2016). *A Review of Deterministic Error Scores and Normalization Techniques for Power Forecasting Algorithms.* 10.1109/SSCI.2016.7849848. Online access: https://ieeexplore.ieee.org/document/7849848

Jensen, T., Fowler, T., Brown, B. Lazo, J., Haupt S.E. (2016), *Metrics for evaluation of solar energy forecasts*, NCAR Technical Note NCAR/TN-527+STR. Online available: http://opensky.ucar.edu/islandora/object/technotes:538

Nielsen, H.A., Nielsen, T.S., Madsen, H., San Isidro Pindado, M.J., Marti, I.: *Optimal combination of wind power forecasts*, Wind Energy **10**(5), pp. 471-482, 2007. Online: https://onlinelibrary.wiley.com/doi/abs/10.1002/we.237

Frías Paredes, L., Stoffels, N., Statistical analysis of wind power and prediction errors for selected test areas, EU 7th Framework project Safewind, Deliverable Dp-7.1. Online available: http://www.safewind.eu/images/Articles/Deliverables/swind.deliverable_dp-7.1_statistical_analysis_v1.6.pdf

Sánchez, I.: *Adaptive combination of forecasts with application to wind energy*. International Journal of Forecasting 24(4), pp. 679–693, 2008. Online: https://doi.org/10.1016/j.ijforecast.2008.08.008

Uncertainty Forecast Information

Bessa, R.J.; Möhrlen, C.; Fundel, V.; Siefert, M.; Browell, J.; Haglund El Gaidi, S.; Hodge, B.-M.; Cali, U.; Kariniotakis, G. *Towards Improved Understanding of the Applicability of Uncertainty Forecasts in the Electric Power Industry*.
Energies 2017, 10, 1402. Online access: https://www.mdpi.com/1996-1073/10/9/1400
http://www.mdpi.com/19961073/10/9/1402
Dobschinski, J., Bessa, R., Du, P., Geisler, K., Haupt, S.-E., Lange, M., Möhrlen, C., Nakafuji, D., Rodriguez, M. d.l.T., *Uncertainty Forecasting in a Nutshell: Prediction Models Designed to Prevent Significant Errors,* IEEE Power and Energy Magazine, vol. 15, no. 6, pp. 40-49, Nov.-Dec. 2017. doi: 10.1109/ MPE.2017.2729100

C. Möhrlen and J.U. Jørgensen, *Chapter 3: The Role of Ensemble Forecasting in Integrating Renewables into Power Systems: From Theory to Real-Time Applications*, Integration of Large-Scale Renewable Energy into Bulk Power Systems - From Planning to Operation, Editors: Du, Pengwei, Baldick, Ross, Tuohy, Aidan (Eds.), pp 79-134.

Möhrlen, C., Bessa, R., Giebel, G., Jørgensen, J.U., *Uncertainty Forecasting Practices for the Next Generation Power System*, Proc. 16th International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plants Germany, of 16th International Workshop on the Integration of Solar Power into Power Systems, 2017. Online available: www.ieawindforecasting.dk/publications

Presentations:

Möhrlen, C., Collier, C. ,Zack , J., Lerner, J.A., *Can Benchmarks and Trials Help Develop new Operational Tools for Balancing Wind Power?*, Proc. of 16th International Workshop on the Integration of Solar Power into Power Systems, Paper WIW-126, Berlin, Germany, 2017. Online access: www.ieawindforecasting.dk/publications

Möhrlen, C., Zack, J., Lerner, J.A., Tuohy, A., Browell, J., Messner, J.W., Collier, C., Giebel, G. RECOMMENDED PRACTICES FOR THE IMPLEMENTATION OF WIND POWER FORECASTING SOLUTIONS - Part 1: FORECAST SOLUTION SELECTION PROCESS and Part 2&3: DESIGNING AND EXECUTING FORECASTING BENCHMARKS AND TRIALS AND EVALUATION OF FORECAST SOLUTIONS, Proc. of 17th International Workshop on the Integration of Solar Power into Power Systems, Paper SIW-126, Berlin, Germany, 2018. Online access: www.ieawindforecasting.dk/publications

Glossary and Abbreviations

Ensemble Forecasting	Ensemble forecasts are sets of different forecast scenarios, which provide an objective way of evaluating the range of possibilities and probabilities in a (weather or weather related) forecast
Probabilistic Forecast	General description of defining the uncertainty of a forecast with objective methods. These can be ensemble forecasts, probability of exceedance forecasts, or other forms of measures of uncertainty derived by statistical models.
Quantile	A quantile is the value below which the observations/forecasts fall with a certain probability when divided into equal-sized, adjacent, subgroups.
Quartile	quantiles that divide the distribution into four equal parts.
Percentile	Percentiles are quantiles where this probability is given as a percentage (0-100) rather than a number between 0 and 1
Decile	quantiles that divide a distribution into 10 equal parts.
Median	the 2nd quantile, 50th percentile or 5th decile, i.e. the value, where the distribution has equally many values above and below that value.

Abbreviations

FSP	*Forecast service provider*
NWP	*Numerical Weather Prediction*
EPS	*Ensemble Prediction System*
NDA	*Non-disclosure Agreement*
RFI	*Request for Information*
RFP	*Request for Proposals*
TSO	*Transmission system operators*
ISO	*Independent system operator*

APPENDIX A: Clarification questions for forecast solution

In order to define the objectives and possible solutions for a forecasting system, it is recommended to follow an overall structure:

1. Describe your situation

In this process, it is imperative to describe exactly those processes, where you need forecasting in the future. Here it is essential to get the different departments involved, especially the IT department. The more accurate you can describe the situation you need to solve with forecasting (e.g. which IT restrictions, limitations and methods for data exchange exist, current or future challenges, etc.), the more straight forward it will be to (1) ask questions to the vendors regarding forecasting methodology, but also (2) get clarity of the involved processes enabling forecasting.

2. Ask Questions to the vendors

The questions to the vendors should be of technical character regarding forecast methodology, but also on available data exchange methodologies, required input data for the models and system support.

TYPICAL QUESTIONS FOR PART 1

Processes: Which processes require forecasting

Data:

- How will the data flow internally be solved: data storage, data exchange, data availability ?
- Which data do we collect that may assist the forecaster to improve accuracy

Data Formats:

- Which formats are required for applications, data exchange and storage ?

Applications:

- Who/which department will use the forecasts, are new applications required to make use of the forecasts ?

Education:

- Is it required to train staff in how to use forecasts ?

Policies:

- Are there policies, political or legal restrictions to be aware of when exchanging data with a forecaster ?

TYPICAL QUESTIONS FOR PART 2

The following are typical questions to get some overview of what is state-of-the-art in forecasting for renewables and what products are available on the market for a specific purpose.

- Describe the methodology you will use when generating forecast for (wind| solar|…)
- How many years of experience do you have in this specific area or related areas
- Required data fields for the forecasting model for the trial
- Time scales and IT requirements for the data for the forecasting model
- Required data for vendor's model, if adopted and used “live”
- Applicable charges for a trial with vendor
- Vendor’s forecast model forecast horizons

APPENDIX B: TYPICAL RFI QUESTIONS PRIOR TO OR IN AN RFP

Methodology

- What unique services can you provide that may address our needs ?
- What input weather data is used
- What methodology is used for power generation for the long-term (>1 days ahead) and short-term forecasting (0...24h).
- Can uncertainty forecasts or probability bands be provided ?[2] If yes, which methodology is being used.
- What are the minimum requirements for wind farm site data?
- Can a Graphical User Interface be provided to visualise forecasts ? If yes, please describe it in detail (e.g. platform dependence, user management, in-house installation or web-based).

Service Level

- What kind of service level does the provider offer (ticket system, personal support, call center, online support, etc.)
- What kind of service level is recommended for the specific service.
- Does the provider have outage recovery guarantee

Contract and Pricing

- What are restrictions and preferences on the pricing structure of your service (e.g. price per park, per MW, per parameter, per time increment)?
- What restrictions/preferences does the provider have in responding to RFPs ?

Experience

- Can the vendor provide minimum of 3 examples of your work that is applicable to our needs (e.g. forecast accuracy, references, methodology)?
- Does the company have significant market shares in the market/area of business
- Additionally, can your company supply products or information that you consider relevant for us when setting out an RFP ?

[2] For a review on methodologies see reference material in section

EXPERT GROUP REPORT

ON

RECOMMENDED PRACTICES FOR SELECTING RENEWABLE POWER FORECASTING SOLUTIONS

Part 2: DESIGNING AND EXECUTING FORECASTING BENCHMARKS AND TRIALS

1. EDITION 2018

Submitted to the Executive Committee of the International Energy Agency Implementing Agreement
on 13th August 2019

Prepared as part of the IEA Wind Task 36, WP 2.1.
Version:3.5
Date: 13. August 2019

Edited by:
Corinna Möhrlen (WEPROG, DK)
John Zack (UL AWS Truepower, USA)
Jeffrey Lerner   - Vaisala, USA

With Contributions from:
Jakob Messner, Anemos Analytics, DK
Jethro Browell, University of Strathclyde, UK
Craig Collier, DNVGL, USA
Aidan Tuohy, EPRI, USA
Justin Sharp, Sharply Focused, USA
Mikkel Westenholz, ENFOR, Denmark

Supported by: Operating Agent Gregor Giebel (DTU, DK)

Table of Contents

1 INTRODUCTION TO BENCHMARKS AND TRIALS

1.1 BEFORE YOU START

This is the second part of a series of three "recommended practices" documents that deal with the development and operation of forecasting solutions. This document "Execution of Benchmarks and Trials" deals with the configuration and steps for carrying out a benchmark or trial of different forecasting solutions prior to selection.

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. The third part "Forecast Evaluation" provides information and guidelines regarding effective evaluation of forecasts, forecast solutions and benchmarks and trials.

If your main interest is in selecting a forecasting solution or verifying the quality of your forecast solution, please move on to part 1 or part 3 of this recommended practices guideline, respectively.

1.2 BACKGROUND

The effectiveness of forecasts in reducing the variability management costs of power generation from wind and solar plants is dependent upon both the accuracy of the forecasts and the ability to effectively use the forecast information in the user's decision-making process. Therefore, there is considerable motivation for stakeholders to try to obtain the most effective forecast information as input to their respective decision tools.

This document is intended to provide guidance to stakeholders on a primary mechanism that has been used extensively in the past years to assess the accuracy of potential forecasting solutions: benchmarks and trials.

This guideline focuses on the key elements to carry out a successful trial or benchmark and on typical pitfalls. It will also provide recommendations as to when it is beneficial or too risky or expensive in terms of resources to carry out a trial or benchmark.

1.3 DEFINITIONS

The two main terms and concepts “trial and benchmark” that are used in this recommended practice shall be defined in the following. Note, the focus has been on forecasting processes in the power industry and the definition may not have a completely general character to be applied to other areas of business.

Renewable Energy Forecast Trial: an exercise conducted to test the features and quality of a renewable energy forecast such as wind or solar power. This may include one or more participants and is normally conducted by a private company for commercial purposes. A trial is a subset of a Renewable Energy Forecast Benchmark.

Renewable Energy Forecast Benchmark: an exercise conducted to determine the features and quality of a renewable energy forecast such as wind or solar power. The exercise is normally conducted by an institution or their agent and multiple participants including private industry forecast providers or applied research academics.
It should be noted that “forecasting trials and benchmarks” will be abbreviated with “t/b” throughout this document for simplicity.

1.4 OBJECTIVES

The guidelines and best practices recommendations are based on years of industry experience and intended to achieve maximum benefit and efficiency for all parties involved in such benchmark or trial exercises. The entity conducting a trial or benchmark taking the recommendations provided in this guideline into consideration will have the following benefits:

1. *Being able to evaluate, which of a set of forecast solutions and forecast service providers (FSP) fits best the need, specific situation and operational setup*
2. *Short term internal cost savings, by running an efficient t/b*
3. *Long term cost savings of forecast services, by following the trial standards and thereby help reduce the costs for all involved parties*

2 INITIAL CONSIDERATIONS

This section is targeted to the task of engaging a forecast service provider (FSP) and how to navigate through the vast amount of information.

2.1 DECIDING WHETHER TO CONDUCT A TRIAL OR BENCHMARK

The most important initial consideration when planning a forecasting trial or benchmark (t/b) is to be clear about the desired outcome.

The following tables provide information about the benefits and drawbacks of conducting a t/b as a key part of the selection process. Before a decision is made to conduct a t/b it is recommended to go through these tables and determine if the effort is warranted.

A possibly attractive alternative approach for a forecast user that wishes to evaluate a set of forecast solutions for their ability to meet the user's needs is to engage an independent trial administrator. An experienced and knowledgeable administrator can act as a neutral third party and advocate for both the vendors and the end-users in the design and execution of a t/b and the evaluation and interpretation of the results. Such an arrangement builds trust in the process among all parties.

An effective administrator can take the requirements from the user and ensure they are realistically incorporated into the trial design. There obviously is a cost to engage such an administrator but it may actually be more cost effective for the user and generate more reliable information for the user's decision-making process.

2.1.1 Benefits of Trials and Benchmarks

Table 1: Decision support table for situations in which trials/benchmarks are determined to be beneficial

Situation	Benefit
Real-time trial for an entire portfolio	Information gain is greater and more representative, but costs are higher; provides the best estimate of the error level and which solution/FSP is best for the target applications
Real-time trial for a selected number of sites	Lower cost but still a substantial information gain if sites are well selected; provides a reasonable idea about the error level and a good indication of which solution/FSP fits is best for the target applications
Retrospective benchmark with historic data for a specific time period separate from a supplied training data set	Low cost In multi-FSP systems, the error level of an additional FSP is secondary, while the correlation with other FSPs determines whether the additional FSP improves the overall error of a multi-FSP composite forecast
Blind forecast without historic measurements	Test to get an indication of the accuracy of forecasts from an FSP in the upstart phase of a project, where no historical data are available. Excludes statistical methods, which need historical data. An inexpensive way to get an indication of forecast accuracy for larger portfolios (> 500MW), where measurement data handling is complex. NOTE: There is an inherent risk that the result may be random and FSP use different methods for blind forecasting and forecasting with measurement data. See also Table 2 for limitations of this approach.

2.1.2 Limitations with Trials and Benchmarks

Table 2: Decision support table for situations in which trials/benchmarks are determined to contain limitations and a t/b is not recommended.

Situation	Limitation	Recommendation
Finding best service provider for large portfolio (> 1000MW) distributed over a large area	Trial for entire portfolio is expensive for client and FSP in terms of time and resources. Limiting scope of trial limits representativeness of results for entire portfolio.	RFI and RFP in which FSP's methods are evaluated and the use of an incentive scheme in the contract terms provides more security of performance than a limited trial.
Finding best service provider for a medium sized portfolio (500MW< X < 1000MW) over a limited area	Trial for entire portfolio is expensive for client and service provider in terms of time and resources. Limiting scope of trial limits representativeness of results for entire portfolio.	RFP in which FSP's methods are evaluated. Design of a system that enables an easy change of FSP and use if an incentive scheme is more a more cost effective approach than a trial.
Finding best service provider for small sized portfolio (< 500MW)	Trial for entire portfolio usually requires significant staff resources for about 6 months	Trial is feasible, but expensive. Difficult to achieve significance on target variable in comparison to required costs and expenses – trial costs makes solution more expensive. Less expensive to setup an incentive scheme and a system where the FSPs can be changed relatively easily.
Finding best service provider for micro portfolio (< 100MW) or single plants	Cost of a trial with many parties can easily be higher than the cost of a 1-year forecasting contract. Time for a trial can delay operational forecast utilization by up to 1 year!	Select FSP based on an evaluation of methods and experience. Design a system that enables an easy change of FSP and use an incentive scheme for FSP performance
Power marketing	Best score difficult to define, as sale of energy is also dependent on market conditions and a statistical forecast performance	More efficient and timely to perform back test of historical forecasts combined with historical prices, or make a strategic

	score such as RMSE or MAE does not reflect the best marketing strategy	choice with an performance incentive.
Market share of FSP in a specific power market is high	FSP monopolies in a specific power market mean that forecast errors are correlated and hence increase balancing costs.	Ask about the market share of a provider and do not choose one with a share > 30% as the only provider!
Blind forecasting, i.e. no historic measurements data available	Without measurements the value of a trial is very limited due to the significant improvement from statistically training forecasts and the importance of recent data for intra-day forecasts Evaluation can only be meaningfully done for day- ahead or longer forecasts. Some FSP may us different methods for forecasting with and without historic data (statistical methods need historical data to function!)	Results are limited to testing quality on upstart phase of new projects, where no historical data exist (see also Table 1). For single sites, the benefits of training are so large (>50% of error reduction at times) that blind forecasting is not recommended. For larger portfolios it can provide an indication of quality - for physical conversion methods only!

2.2 TIME LINES AND FORECAST PERIODS IN A TRIAL OR BENCHMARK

Time lines and forecast periods need to be set strictly in a trial or benchmark in order to achieve a fair, transparent and representative exercise.

The following time lines should be considered:

(1)Start and stop dates of the t/b must be fixed

(2)Start and stop dates must be the same for all FSPs

(3)Pre-trial setup and test dates for IT infrastructure (including any required security protocols) for trial must be specified and enforced

(4)Delivery times of forecasts must be set and enforced

(5)Forecasts for periods with missing forecasts from one FSP must be excluded for all FSPs

2.3 1-PAGE “CHEAT SHEET” CHECKLIST

The following checklist is provided to help trial organizers save time, apply best practices, and avoid common pitfalls when designing and executing forecast trials. It has been compiled by leading forecast vendors and researchers with many years experience.

Forecast Trial Checklist

--Preparation--

☐ Determine outcomes / objectives

☐ Consult expert with experience

☐ Establish timeline and winning criteria

☐ Decide on live or retrospective trial

☐ If live trial with datafeed, begin datafeed setup

☐ Gather metadata (use IEA checklist spreadsheet)

☐ Determine if adequately resourced to carry out

☐ Obtain historical data

☐ Invite forecast service providers

☐ Distribute historical and meta-data

☐ Finalize datafeed configuration (if applicable)

☐ Allow two weeks Q&A prior to start

☐ Begin

--During Trial--

☐ Develop validation report

☐ Check interim results

☐ Provide interim results (if no live data being provided)

☐ End

--Post Trial--

☐ Provide final results

☐ Notify winner(s)

☐ Contract with winner(s)

☐ Start Service

3 PHASES OF A BENCHMARK OR TRIAL

There are three main phases of a trial or benchmark exercise: preparation ahead of the trial, actions during the trial, and post-trial follow up.

3.1 PHASE 1: PREPARATION

The time required for the pre-trial preparation is significant and should not be underestimated to insure a successful outcome. If the operator of the trial has no experience in renewable energy forecasting or running a t/b, it would be prudent to contact an experienced individual, organization or forecast provider to obtain feedback on what can reasonably be accomplished given the target time line and objectives. Part 1 of this recommended practice contains a decision support path that may be useful for determining the proper course of action.

3.1.1 Key Considerations in the Preparation Phase

Once the objectives of the t/b are known (see Section 1.1 Background and 1.2 Objectives), there are some key decisions to be made that will play a major role in determining the complexity of the trial.
They are:

(1) **Choice of forecast horizon**

Are forecast horizons less than 6 hours operationally important? If the answer is "no", establishing a live data feed may not be necessary. Although there are advantages of running a trial with a live data feed, it is one of the most time consuming aspects of trial preparation.

Are forecast lead times greater than “day-ahead” operationally important? If the answer is no, this will reduce the volumes of data that need to be processed saving time and resources.

If many lead times are of operational importance, consider that the performance of different providers will likely vary across lead times, therefore, different lead times, e.g. hour-ahead, day-ahead and week-ahead, should be evaluated separately.

(2) **Weather conditions for the exercise:**

Will the benchmark take place during periods of more difficult to predict weather conditions that reflect the organization’s difficulties in handling renewable generation, e.g. windy or cloudy periods? The answer here should be "Yes" to insure the sample size of harder-to-forecast events is sufficient. If the answer is "No", the trial operator should strongly consider doing a retrospective forecast (also known as "hindcast") that includes the types of conditions that are critical for the user’s application.

(3) **Historical data/observations for the exercise:**

For locations in which there are significant seasonal differences in weather conditions and the associated renewable generation levels and variability, it is best to provide 12 months or more of historical data from the target generation facilities to the FSPs for the purpose of training their forecast models. However, if it is not feasible to make this amount of data available or if the target location does not exhibit much seasonal variation, most FSPs can typically train their forecast models reasonably well with 3-6 months of on-site historical observations.

It should be noted that advanced machine learning methods often exhibit significantly greater performance improvement over less sophisticated methods as the training sample size increases. Thus, FSPs that employ the latest and most advanced machine learning prediction tools may not be able to demonstrate the ultimate value of their approaches, if only short historical data sets are provided. If 6-12 months of data are not available, the trial operator might consider another location or conduct a longer trial on the order of 4-6 months to monitor forecast improvements over time as more data becomes available to the FSPs to improve the quality of the training of their prediction models.

In general it is recommended that the t/b operator should provide a data set of the typical length that is available data for the application that is the target of the t/b. If more historical data is available for a t/b than in the typical application, care should be taken in the evaluation of methods, as e.g. machine learning methods might outperform e.g. physical methods in the trial, but perform worse in the real application due to the benefits associated with the longer data sets.

(4) **Representativeness:**

Is the benchmark location representative from a wind-climatology perspective of the scope of locations for which the operator will ultimately require operational forecast services? That is, the trial operator should select a location that is needed for subsequent forecasting or a location with a similar climatology. Operators should also be aware of the randomness of forecast performance on single locations, if a large area with many sites is the target.

It should be noted that forecast performance exhibits a significant “aggregation effect”. That is the magnitude and patterns of forecast errors vary substantially depending on the size and composition of the forecast target entity. Thus, the characteristics of forecast errors for an individual turbine, a single wind park and a portfolio of wind parks will typically be quite different and the forecast evaluator should be very careful when inferring forecast performance characteristics from one scale of aggregation (e.g. a single wind park) to a different scale (e.g. a geographically diverse portfolio of wind parks) (see also part 3 of this recommended practice for more details on evaluation methods).

(5) **Metrics:**

Are the metrics that will be used to evaluate the forecasts meaningful to the success of my project? There are a wide variety of well-documented error metrics that penalize forecast errors differently. For example, root mean squared error penalizes large errors more than small errors. It is important to choose a metric, or set of metrics, that reflects the value of an improved forecast to the user’s application and can discriminate between different forecast solutions. Please refer to part 3 of this recommended practice for details on metric selection.

3.1.2 Metadata Gathering in the Preparation Phase

Details of the forecast trial, such as location and capacity of the target generator, are required by all FSPs and comprise the trial Metadata. Appendix A “Metadata Checklist” provides the information that is typically needed by FSPs for participation in a trial and is designed to be used as a spreadsheet form that is completed during the preparation phase of a t/b.

This should also include the desired format (filename and content) of the forecasts you’ll be comparing. The best way to communicate the forecast file format to multiple FSPs is to provide an example file.

3.1.3 Historical Data Gathering in the Preparation Phase

On-site observations of power production or the renewable resource (e.g., irradiance or wind speed at hub height) are critical for helping the FSPs statistically “train” their forecast models and thus reduce error and bias in the forecasts. Good quality data is critical. ”Good quality" means that the data does not, for example, contain many gaps or unrepresentative values. Curtailed power data should be accompanied by plant availability or a curtailment flag.

Data time intervals should be regular and there should be a clear documentation of the units, how the observations were averaged, the time zone of the data, and whether there's a shift in time due to daylight savings time. Appendix A of this document has a concise list of the necessary historical data attributes required to efficiently start a t/b.

3.1.4 IT/Data Considerations in the Preparation Phase

Most organizations have constraints on the amount of IT resources available for a t/b. Therefore, it is best to plan ahead or keep the sending and receiving of data very simple. The primary IT issue is typically the selection and setup of data formats and communication protocols that will be used for the t/b operator to send data to the FSPs and for the FSPs to send forecasts to a platform designated by the t/b operator.

There are many possibilities for data formats, which range from a simple text file with comma separated variables (CSV) to more sophisticated XML or openAPI formats. Similarly, there are a wide range of communication protocols that can be used. These range from the relatively simple Secure Shell File Transfer Protocol (SFTP) to more sophisticated web service or API structures. The more sophisticated structures have advantages and there are many IT companies and resources that support these structures but they almost unavoidably increase the complexity of the setup.

Unless adequate IT resources or knowledge are available for all participants (especially the operator) it is recommended that simple data formats and communication resources be employed for a t/b. This typically means the use of the CSV data format and an SFTP data communications protocol.

If a live trial is planned (most common), but real-time data will not be made available to the FSPs, then a place for each FSP to send forecast files will need to be setup. One of the metrics that is often used to evaluate an FSP is the timeliness of forecast delivery. In this case, it is important that a mechanism to verify the time of delivery be established.

If real-time data is provided by the t/b conductor, it is typically easiest to create a common password-protected file server directory from which FSPs can download the data via a protocol such as SFTP. Another approach is to use SFTP to push data files to each FSP. This typically requires more effort, especially for the t/b operator.

Historical data can be provided to FSPs in the same data format via the same communication protocol. However, it often requires a SCADA engineer or expert on third party software to extract the historical data for the SCADA (or other) data archive.

Another often-overlooked data-related issue is the legal agreements required to disseminate data from possibly multiple data provider entities (e.g. the wind facility owners/operators) to multiple data user entities (e.g. the FSPs in the t/b). This may be relatively simple in cases in which the user (such as a generator fleet operator) owns all the data and is willing to make it available for the t/b with few restrictions. However, it be a very complex and time consuming process in cases in which the user (e.g. a system operator) does not own the data and merely serves as a conduit from the multiple data owners with different data dissemination restrictions to the data users.

In such cases, the process of formulating and executing the required legal documents (such as non-disclosure agreements (NDAs)) can cause substantial delays in the initiation of a t/b and perhaps even change its scope.

See Appendix B for example formats in csv and xml.

3.1.5 Communication in the Preparation Phase

Anonymizing the FSPs for all communication is considered a best practice as it ensured transparency of the available information, promotes competition and entry from smaller FSPs trying to become more established in the industry. Communication via email therefore should always be consistent with blind copies to all FSPs.

Consistent in this context means always sending and sharing emails with the same group of FSP users. Common information sharing engenders trust and the perception of fairness in the benchmark or trial process. In the preparation phase, it is not uncommon that the FSPs will have questions that could affect how the trial is conducted.

For this reason, it is recommended to have a 2-week question and answer period *before* the official start date to allow FSP participants to ask questions that then can be answered in a living document that contains all questions and answers up to the present time. All participants should be notified whenever this document is updated.

The importance of frequent and clear communication cannot be overstated when conducting a t/b. Not only will the t/b operator receive the most accurate forecasts, it will make it much easier the next time a t/b is executed to gage the state-of-the-art in forecasting technologies and features.

3.1.6 One-week test run in the Preparation Phase

It is recommended to that a one-week test period be conducted before the official start date of the t/b to identify and remove any technical issues that could invalidate forecast results. This helps to improve the likelihood that all results can be included in the final validation calculations without the need for omitting the first part of the t/b.

3.2 PHASE 2: DURING BENCHMARK/TRIAL

Often the most successful forecast provider is one that can show steady improvement over time. Providing an interim validation report will not only prepare the trial operator for the final validation report but will give important feedback to the FSPs.

3.2.1 Communication during the T/B

In a well-designed t/b, most of the communication between the trial operator and FSPs should be during the pre-trial period. However, issues often arise especially during a live trial with a real-time data feed. It may be helpful to all t/b participants to establish an open forum during the first part of the live t/b period (e.g. the first 2 weeks) to provide a way to effectively and uniformly resolve all issues early in the t/b period However, it is strongly recommended that if any attributes of the t/b are changed at any point during the live part of the t/b, the changes should be communicated to all participants immediately as they might require action on the FSP’s part. Examples might include: changing the forecast validation metric, if there are unreported outages that should be omitted for future model trainings, or if the location of the data feed or forecast file destination has changed. It should be emphasized that all communications related to the t/b should be distributed to all FSPs without exception. Additional communication with individual FSPs (including forecast incumbents) can be interpreted as bias on the part of the operator of the t/b and in some cases may actually bias the t/b result due to information that impacts forecast design, production or delivery not being equally available to all FSPs.

3.2.2 Forecast Validation and Reporting during the T/B

Forecast validation reports are often compiled during the t/b. With forecast data coming in at regular intervals, the t/b operator has real data to feed into the validation report. If the t/b has a duration of several months (i.e., >3 months), it is recommended to provide at least one interim report to FSPs that include anonymized results from all FSPs. This benefits the trial operator as errors in the evaluation process or the report generation can be flagged earlier and ways to make the report generation more efficient can be realized. The interim report benefits the FSPs as course-corrections can be made during the t/b to improve the forecasts.

If there are several FSPs participating, efficiencies can be realized by automating part or most of the validation metrics especially as the forecast file format should be the same from all FSPs.

3.3 PHASE 3: POST TRIAL OR BENCHMARK

The post trial phase is an important aspect of the t/b because FSP selection will likely occur during this phase based on the criteria set out at the start of the t/b. (see recommended practices part 1 on "evaluation of services and decision support").

3.3.1 Communication at the end of the T/B

If the trial operator hasn't already done so, an email should be sent within a week before the end date of the t/b to alert FSPs that the end of the trial is near and to communicate the timeline for sharing results and re-iterate the specifications of the FSP selection process.

3.3.2 Forecast Validation and Reporting at the end of the T/B

If an interim report was provided during the trial, then the final report can either be an updated version of the validation report expressing the bulk metrics or appended month-by-month forecast validation results. For transparency and to promote further forecast improvements, it is recommended that the t/b operator share the anonymized forecast results from each FSP at the time-interval frequency that forecasts were being made at (e.g., hourly). This will help FSPs discover where forecasts are similar or different from the competition which may spawn improved methodologies.

4 BEST PRACTICES

Although there are many different ways that a t/b may be conducted, there are some common elements of a successful t/b that provide the t/b operator with the best forecast solution and the participants with useful knowledge of where their forecast ranks among the competition.

The following are some selected best practice recommendations:

(a) A clear purpose for the t/b exercise
(b) Pre-defined and explicit accuracy metrics and solution selection criteria
(c) A clear time line (start/end dates, selection announcement, contract award)
(d) Anonymized forecast results. Ask FSP's approval to share results. This helps FSPs find ways to improve their forecast accuracy and see their shortcomings.
(e) Question & answer period before benchmark period begins (~ 1-2 weeks)
(f) Sufficient time allocated for testing the transfer of data between participant(s) and operator
(g) Prompt communication to participants regarding any changes or answers to questions that arise
(h) Consistent forecast file format requested of all - example file sent to all
(i) Consistent data formats (both observations and forecast files) ideally as close to (if not identical to) what the trial operator needs, once contract is executed.
(j) Providing the same historical and project metadata to all participants
(k) Allocation of sufficient resources by the t/b conductor to furnish data and perform validation

(l) PITFALLS TO AVOID

The following list describes a few common mistakes and how to avoid them in the design, setup and execution of a forecast t/b.

The consequences of errors and omissions in trials are often underestimated. However, if results are not representative, the efforts that have gone into a t/b can effectively be wasted. Some of these common pitfalls can be expensive to the operator because they result in placing the operator in a position of making a decision without having truly objective and representative information to base it on.

1. **Poor Communication**

   All FSPs should receive the same information. Answers to questions should be shared with all FSPs. Fairness, and perception of fairness, are important when running and evaluating the results of trials.

2. **Unreliable Validation Results**

   Don't compare forecasts from two different power plants or from different time periods. Forecast performance will vary depending on location and specific time periods. Only forecasts for the same period and location/power plant/portfolio should be compared.

3. ***Examples of Bad Design***

   (a) A trial with 1 month length during a low-wind month
   (b) No on-site observations shared with forecast providers
   (c) Hour-ahead forecasts initiated from once a day data update
   (d) Data only processed in batches or at the end of a real-time trial – this is an invitation for cheating to the FSPs. In most cases, there will be some that use the opportunity to do so

4. ***Examples of Missing or Non-communicated Data***

   (a) daylight savings time changes are not specified
   (b) data time stamp represents interval beginning or ending not specified
   (c) plant capacity of historical data differs from present capacity
   (d) data about curtailment and maintenance outages not provided

5. **Possibility of Cheating**

   In any type of competition, cheating is a reality. If there are not taken precautions, results may be biased and decisions are taken upon incorrect results. It is recommended that the possibility of cheating is considered with seriousness and avoided, where possible.

   Typical situations, where cheating is being observed are:

   - Forecast t/b being carried out for a period of time for which FSPs are given data. Recommendation: separate historical data from t/b period.

   - if there is one or more incumbent FSP with a longer history of data, this should be taken into consideration in the evaluation, as such an FSP may not be able or willing to modify forecast models for the purpose of being “comparable” in a t/b. Recommendation: see limitations in Table 2 and part 3 of this recommended practice.

Other observed situations, where cheating is happening is:

- Missing forecasts: FSP leave out “difficult situations” as missing forecasts are often not penalized. However, missing data may bias “average” forecast metrics, potentially resulting in the formulation of incorrect conclusions. Recommendation: remove dates where forecasts are missing for one FSP for all FSPs

- If delivered forecasts from a FSP as part of a live trial are not downloaded, moved or copied in accordance with the operational process being simulated, and certainly before the time period being forecast, FSPs can potentially renew forecasts with high accuracy due to fresher information being available. Recommendation: Such an omission should not be underestimated and care taken for the evaluation.

5 REFERENCE MATERIAL

5.1 REFERENCES

[1] J. Kehler and D. McCrank, Integration of wind power into Albertas electric system and market operation, Proc. of IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, Pittsburgh, PA, pp. 1-6. doi: 10.1109/PES.2008.4596824, 2008.

[2] E. Lannoye, A. Tuohy, J. Sharp, V. Von Schamm, W. Callender, L.Aguirre, Solar Power Forecasting Trials and Trial Design: Experience from Texas, Proc. of 5th International Workshop on the Integration of Solar Power into Power Systems,, Brussels, Belgium, ISBN: 978-3-9816549-2-9, 2016.

[3] E. Lannoye, A Tuohy, J Sharp, and W Hobbs, Anonymous Solar Forecasting Trial Outcomes, Lessons learned and trial recommendations, Proc. of 7th International Workshop on the Integration of Solar Power into Power Systems, Paper SIW-126, Berlin, Germany, 2017.

[4] C. Möhrlen, C. Collier , J. Zack , J. Lerner , Can Benchmarks and Trials Help Develop new Operational Tools for Balancing Wind Power?, Proc. of 7th International Workshop on the Integration of Solar Power into Power Systems, Paper SIW-126, Berlin, Germany, 2017. Online access: http://download.weprog.com/WIW2017-292_moehrlen_et-al_v1.pdf.

[5] Bessa, R.J.; Möhrlen, C.; Fundel, V.; Siefert, M.; Browell, J.; Haglund El Gaidi, S.; Hodge, B.-M.; Cali, U.; Kariniotakis, G. Towards Improved Understanding of the Applicability of Uncertainty Forecasts in the Electric Power Industry. Energies 2017, 10, 1402. Online access: http://www.mdpi.com/1996-1073/10/9/1402

5.2 CONFERENCE PAPERS

Corinna Möhrlen, Recommended Practices for the Implementation of Wind Power Forecasting Solutions Part 1: Forecast Solution Selection Process, Proc. 17th International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant, Stockholm, Sweden, October 17.-19, 2018.
Online Access: http://www.ieawindpowerforecasting.dk/publications

Corinna Möhrlen, John Zack, Jeff Lerner, Aidan Tuohy, Jethro Browell, Jakob W. Messner, Craig Collier, Gregor Giebel
Part 2&3: DESIGNING AND EXECUTING FORECASTING BENCHMARKS AND TRIALS AND EVALUATION OF FORECAST SOLUTIONS, Proc. 17th International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant, Stockholm, Sweden, October 17.-19, 2018
Online Access: http://www.ieawindpowerforecasting.dk/publications

C. Möhrlen, R. Bessa, Understanding Uncertainty: the difficult move from a deterministic to a probabilistic world, Proc. 17th International Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant, Stockholm, Sweden, October 17.-19, 2018

Online Access: http://www.ieawindpowerforecasting.dk/publications

C. Möhrlen, R. Bessa, G. Giebel, J. Jørgensen,G. Giebel, Uncertainty Forecasting Practices for the Next Generation Power System, Proc. 16th Int. Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant**, Berlin (DE), 26-29 June 2017.**
Online Access: http://www.ieawindpowerforecasting.dk/publications

C. Möhrlen (WEPROG, Denmark), C. Collier (DNV GL, USA), J. Zack (AWS Truepower, USA), J. Lerner (Vaisala, USA)
Can Benchmarks and Trials Help Develop new Operational Tools for Balancing Wind Power?, Proc. 16th Workshop on Large-Scale Integration of Wind Power into Power Systems as well as on Transmission Networks for Offshore Wind Power Plant_**, Berlin (DE), 26-29 June 2017.**
Online Access: http://www.ieawindpowerforecasting.dk/publications

5.3 PRESENTATIONS

C. Möhrlen, C. Collier , J. Zack , J. Lerner , Can Benchmarks and Trials Help Develop new Operational Tools for Balancing Wind Power?, Proc. of 7th International Workshop on the Integration of Solar Power into Power Systems, Paper SIW-126, Berlin, Germany, 2017. Online access: http://download.weprog.com/WIW17-292_MOEHRLEN-ET-AL_PRESENTATION_20171028.pdf

J. W. Zack (2017)**,** Wind and solar forecasting trials experience: do's and don'ts, Part 2 UVIG 2017 Forecasting Workshop, Atlanta (US), 21-22 June 2017
Online access: http://www.ieawindforecasting.dk/-/media/Sites/IEA_task_36/Publications/forecast_trials_session_4_uvig2017_jzack.ashx?la=da
C. Collier (2017), Why Do Forecast Trials Often Fail to Answer the Questions for which End-Users Need Answers: A Forecaster's Point of View UVIG Forecasting Workshop, Atlanta (US), 21-22 June 2017. Online access: http://www.ieawindforecasting.dk/-/media/Sites/IEA_task_36/Publications/forecast_trials_session_4_uvig2017_ccollier.ashx?la=da
T. Maupin (2017), Wind and Solar Forecasting Trials: Do's and Don'ts, Part 1 Best practices. UVIG 2017 Forecasting Workshop, Atlanta (US), 21-22 June 2017. Online access:http://www.ieawindforecasting.dk/-/media/Sites/IEA_task_36/Publications/forecast_trials_session_4_uvig2017_ccollier.ashx?la=da

5.4 GLOSSARY

T/B: *Trial and Benchmark*
FSP: *Forecast Service Provider*

Forecast Creation Time: *The time at which a forecast is created. This is useful when determining skill at different lead times though usually deliver time will be used instead.*

Forecast Delivery Time: *Similar to creation time, only this is the time the forecast was actually received by the end user. This is then used to define what lead time should be ascribed.*

Forecast Lead Time: *The time between the delivery (or creation) time and the beginning of the first interval being forecasted. For example, a forecast delivered at 8:30, where the first entry is for 5-minute period ending 9:05 has a 30 minute lead time.*

Forecast Horizon Time: *The time of the last forecast interval relative to the delivery time. For instance, a day head forecast with hourly intervals from midnight to midnight the following day has a horizon time of midnight on date+2*

Forecast Interval: *The length of time between the forecast start time and the forecast end time.*

Forecast Valid Time: *The time interval for which a forecast is valid. The last valid time is the same forecast horizon.*

Appendix A: Metadata Checklist

The following checklist (Table A.1), when filled out, will greatly aid FSPs in configuring forecasts efficiently. Many of the essential questions relevant to benchmark and trial forecast model configuration are provided here.
Note that the following table is an example and may not contain all necessary information required for the FSP to setup a solution for your purpose. The table is meant to serve as a guideline and can be copied, but should be carefully adopted to the specific exercises before sending out to FSP with questions filled in. If this is done with care, it will expedite forecast configuration and save back and forth communication time.

Table A.1: Example of a Metadata Checklist

Wind Power Forecast Trial Checklist	
Metadata	
Name of site(s) as it should appear in datafile	
Latitude and longitude coordinates of sites	
Nameplate capacity of each site	
Will a graphical web tool be needed?	
Turbine make/model/rating	
Number of turbines	
Hub height of turbines	
Please attach suitable plant power curve	
Forecast output information	
Forecast output time intervals (e.g., 15-min, 1-hourly)	
Length of forecast required	
Timezone of forecast datafile	
Will local daylight savings time be needed?	
Forecast update frequency (e.g., once a day, every hour)	
Value of Forecast	
Which variables will be forecasted and validated?	
Which forecast horizons are being validated?	
Which metrics are being used to gage forecast performance?	
List criteria for determining winning forecast provider	
Will results be shared as a report? Will results be anonymized?	
On what frequency will results be shared with forecast provider?	
Historical Data Checklist	
Is the data in UTC or local time?	
Is the data interval *beginning* or *ending* or *instantaneous?*	
What are the units of the data?	
If met tower histories being provided, indicate height of measurements.	
Realtime Data Checklist (if applicable)	
Is the data in UTC or local time?	
Is the data interval *beginning* or *ending* or *instantaneous?*	
What are the units of the data?	

Email and Telephone number of technical point of contact (POC)	
Email and Telephone of datafeed POC	
Name and email of users that need website access	
Person name and email that filled out this checklist:	

Appendix B: Sample forecast file sturctures

Back and forth communication can sometimes delay the start of a trial or benchmark. One of these delays is getting the forecast file output format just right for the beginning of the trial. Standardization of the format will make the trial operators life much easier when time comes to validating forecasts. A best practice here is for the trial operator to use a format that is already in use or a format that has already proven to work in operations.

Table B.1 below shows the first few fields of a forecast file template.

Plant Output	Acme Wind Farm	1.11.2017 4:00	1.11.2017 5:00	1.11.2017 6:00	1.11.2017 7:00
Power	MW	41.43	41.43	41.43	40.89
Windspeed	m/s	11	10	10	10
Time zone: Central European Summer Time (CEST)					
Intervals: hour ending					
Date time format: dd.mm.yyyy hh:mm (e.g., 06.08.1969 08:30)					

Table B.2 shows typical XSDs for forecasts and SCADA data in a b/t, usable also with WebServices

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
 <xs:element name="WindForecast">
  <xs:complexType>
   <xs:attribute name="VendorCode" type="xs:string" use="required" />
   <xs:attribute name="ImportTime" type="xs:dateTime" use="required" />
   <xs:sequence>
    <xs:element name="CUSTOMER">
     <xs:complexType>
      <xs:attribute name="name" type="xs:string" use="required" />
      <xs:sequence>
       <xs:element name="Forecast">
        <xs:complexType>
         <xs:attribute name="MWaggregated" type="xs:double" use="required" />
         <xs:attribute name="time" type="xs:dateTime" use="required" />
         <xs:sequence>
          <xs:element name="Probability">
           <xs:complexType>
            <xs:attribute name="P95" type="xs:double" use="required" />
            <xs:attribute name="P50" type="xs:double" use="required" />
            <xs:attribute name="P05" type="xs:double" use="required" />
            <xs:attribute name="max" type="xs:double" use="required" />
            <xs:attribute name="min" type="xs:double" use="required" />
           </xs:complexType>
          </xs:element>
          <xs:element name="WindFarms">
           <xs:complexType>
            <xs:sequence>
             <xs:element name="WindPark1">
              <xs:complexType>
               <xs:attribute name="id" type="xs:string" use="required" />
               <xs:attribute name="mw" type="xs:double" use="required" />
              </xs:complexType>
             </xs:element>
            </xs:sequence>
           </xs:complexType>
          </xs:element>
         </xs:sequence>
        </xs:complexType>
       </xs:element>
      </xs:sequence>
     </xs:complexType>
    </xs:element>
   </xs:sequence>
  </xs:complexType>
 </xs:element>
</xs:schema>
```

- *SCADA XSD for exchange of real-time measurements*

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="WindSCADA">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" name="WindPark">
          <xs:complexType>
            <xs:attribute name="ID" type="xs:string" use="required" />
            <xs:attribute name="Time" type="xs:dateTime" use="required" />
            <xs:attribute name="Mw" type="xs:decimal" use="required" />
            <xs:attribute name="Availabilty" type="xs:decimal" use=" optional" />
            <xs:attribute name="CurrentActivePower" type="xs:decimal" use=" optional"/>
            <xs:attribute name="Curtailment" type="xs:string" use="optional" />
            <xs:attribute name="WindSpeed" type="xs:decimal" use="optional" />
            <xs:attribute name="WindDirection" type="xs:decimal" use="optional" />
            <xs:attribute name="AirTemperature" type="xs:decimal" use="optional" />
            <xs:attribute name="AirPressure" type="xs:decimal" use="optional" />
            <xs:attribute name="Outage" type="xs:decimal" use="optional" />
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

EXPERT GROUP REPORT

ON

RECOMMENDED PRACTICES FOR SELECTING RENEWABLE POWER FORECASTING SOLUTIONS

Part 3: Evaluation of Forecasts and Forecast Solutions

1. EDITION 2018

To be Submitted to the
Executive Committee of the
International Energy Agency Implementing Agreement
on 1st March 2019

Prepared as part of the IEA Wind Task 36, WP 2.1.

Version: 1.3

Date: 28 February 2019

Edited by:

Corinna Möhrlen (WEPROG DK, DE)

John Zack (UL AWS Truepower, USA),

Jakob Messner (Anemos Analytics, DK),

Jethro Browell (University of Strathclyde, UK)

With Contributions from:

Craig Collier (DNVGL, USA), Aidan Tuohy (EPRI, USA)

Supported by: Operating Agent Gregor Giebel (DTU, DK)

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 of the IEA Wind Task 36 (2016-2018) as part of workpackage 2.1.

The editors want to thank everybody that has been part of the meetings, workshops and sessions and conributed in the discussions, provided feedback or other input throughout the past 3 years. Special thanks to Stephan Vogt for the provision of the significance test example in section 5.4.2.

IEA Wind Task 36, February 28, 2019

Editors:
Dr. Corinna Möhrlen (WEPROG) <com@weprog.com>
Dr. John Zack (UL AWS Truepower) <john.zack@awstruepower.com>
Dr. Jakob W. Messner (Anemos Analytics) <jakob.messner@posteo.net>
Dr. Jethro Browell (University of Strathclyde) <jethro.browell@strath.ac.uk>

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 development and operation of forecasting solutions in the power market. 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.

1.2 Introduction

The evaluation of forecasts and forecast solutions is an obligation 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 design and outline forecast evaluations with this importance in mind, give this part the required attention and thereby ensure that 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 first 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], [4], [12], [6]) which will also specifically be 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 first version of the recommended practices guideline focuses on:

1. *Impact of forecast accuracy on application*
   First, it's often difficult to define the forecast accuracy impact to the bottom line as forecasts are just one of many inputs. Second, trials or benchmarks often last longer than anticipated or 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 exists no universal 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 tailored forecast solutions 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 with three principles for an evaluation to be:

1. **representative**

2. **significant**

3. **relevant**

How to setup an evaluation process and achieve these principles is the core of this recommended practices 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 the second base principle of evaluation and verification tasks. If forecasts are evaluated against data that inherit errors, results may still show some significance, but may no longer be considered trustworthy, nor 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, how to evaluate benchmarks and trials and new forecasting techniques or developments. 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 on 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 decision 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 of the evaluation sample is one of the most important representativeness factors. The size 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 use of a small sample increases the probability that the conclusions 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. The predictability of these events is often lower (i.e. higher forecast errors) than that of typical weather conditions. Therefore, a small sample that contains such very unusual events may lead to an overestimation of the typical magnitude of forecast errors. Conversely, a small sample that has no difficult-to-forecast events may lead to an underestimation of the typical forecast error. However, the performance of the forecasts under unusual weather conditions may be very important to the user's application and therefore an assessment of how different forecast systems perform under these conditions may be very valuable information to the solution selection process. Thus,there are two key points that the user should keep in mind when using a small evaluation sample. First, conclusions from a small sample will always be less reliable (i.e. more uncertain) than those from a larger sample. Second, the user should make an effort to understand the composition of the small sample by examining the relationship between the weather conditions in the sample relative to an estimate of the climatological (i.e. long-term) distribution (e.g. was the sample dominated by typical conditions or were there one or more atypical events?) for the site or region and also by examining the forecast error distributions (e.g. were almost all of the forecast error magnitudes clustered around the average magnitude or were there a significant number of outliers?) (see also 5.1.1, 4.1.4,).

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.

The composition of an evaluation sample is another key issue. The composition should be constructed so that all significant modes of variation of the forecast variable (e.g. wind 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. However, if this is not practical than at least there should at least be a representative sample of the most important modes for the application (e.g. high wind season when the speeds are near cutout or periods when the wind speed is frequently in the highly sensitive steeply sloped part of the turbine power curve).

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 two purposes: (1) as training data for the statistical components of each forecast system and (2) evaluation of the forecast performance. If the data has many quality issues then the representativeness of both 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) values that are unrepresentative of meteorological conditions because of undocumented outages or curtailments. If a substantial of data with these issues is used is used in the evaluation process for either of the two 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 to 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 is 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 situation 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 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 are 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 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 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 to base 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.

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.

Method 2: Bootstrap Resampling

A simple alternative method is to simulate different data sets, through the use of 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. [**Diebold1995**] proposed a variation of this t-test to account for temporal correlations in the data and can therefore provide a more accurate significance quantification. [**Messner2018**] 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" 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.

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. 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; standards and methods for collecting and reporting data are available from multiple sources referenced in this section*
- *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, service periods, 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 significance of evaluation results. Therefore, it is crucial to assure 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 of 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 turbines SCADA system maintenance.

Recommendations on minimum technical requirements is going 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 into 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.

- For *relevant* evaluation results, minimum standards for measurement data precision and quality have to be ensured and maintained.

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 verification or evaluation of forecasting tools. This section will discuss the main aspects to be considered for any measurement data collection and archiving. In the following we limit the description for the purpose of verification or 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.

- To receive *relevant* results, remove events from the evaluation data set that are effected by non-weather related production constrains 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, or if no aggregation over time is carried out, for example, if hourly values are provided that are not hourly averages of higher resolution data, but instantaneous values taken at the start of the hour, this should be communicated to the forecast provider to assure optimum forecast performance for the intended application. Furthermore, it is strongly recommended to aggregate the measurement data 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.

Archiving data is dependent on the way the further processing of the data is planned. In most cases, it is useful to archive data in a database. There are many different structures of data bases 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. 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, bad, corrupt or misleading data signals can have an almost unlimited amount of reasons, which means that specific limits, operating ranges and validity checks need to be established when dealing with observational data. While 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.

- 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.

3.5 Filtering processes and Data Preparation

The filtering process and data preparation are crucial whenever dealing with measurements or observational data in the evaluation process. A number of parameter have been identified as being important to consider in the preparation phase of any verification/evaluation. Messner et al. [2018]) recommended the following requirements:

- **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 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. However, if this is not practical than at least there should at least be a representative sample of the most important modes for the application (e.g. high wind season when the speeds are near cutout or periods when the wind

speed is frequently in the highly sensitive steeply sloped part of the turbine power curve).

- **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.

- **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, especially 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, 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.

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 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 my 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, 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 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 not 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 to produce a representative evaluation forecast performance which is relevant to the forecast's end use.

Quantitative evaluation methods are usually the core of the evaluation framework since they allow to objectively rank different forecast models. Typical choices of quantitative metrics are the (root) mean squared error, the mean absolute error or the quantile score (see [**Messner2018**] for details) for continuous forecasts and various quantities derived from contingency tables for forecasts of binary forecasts.

As emphasized in Section **??**, 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

The most common error metrics used in the wind industry summarize 'typical' error 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, 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 means fewer small errors too.

Examples of typical error metrics are discussed in section 5.1 and especially in section 5.1.1.

4.1.2 Outlier/Extreme Error

Another important attribute is the prevalence of large errors. Some applications aim to prepare for large errors, such as managing reserve energy or other risk management. Calculating metrics based on historic 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 positive and that 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 level error and then evaluate the performance of a particular forecast solution at predicting such events. There are four possibilities for each predicted

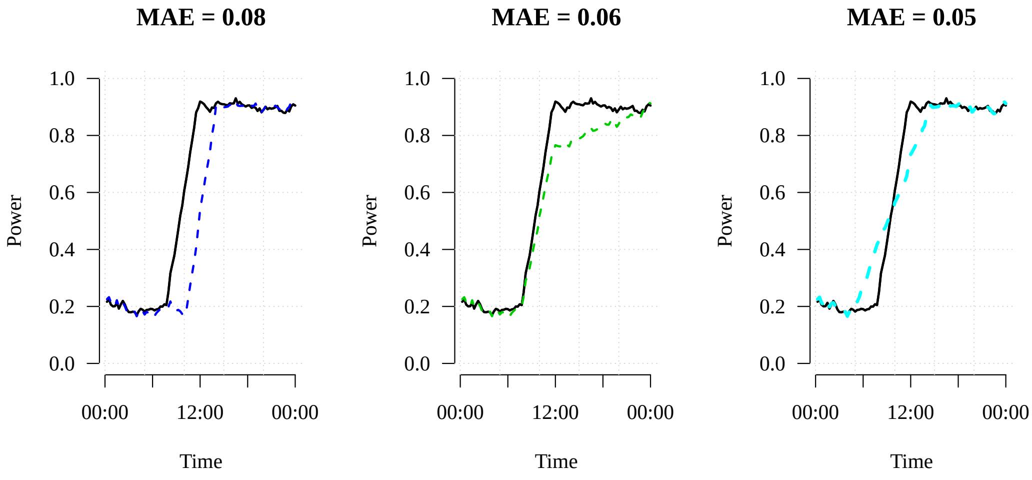

Figure 4.1: Examples of different types of ramp forecast error. Actual power is shown as solid black lines, forecasts are colored dashed lines. From left to right: phase or timing error, level error and ramp rate error. The mean absolute error (MAE) for each forecast is shown above the plots. Despite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

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.

4.1.5 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 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.

4.2 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 are considered from the outset and are known 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 following set of principles:

- *Verification is subjective*
  it is important to understand the limitations of a chosen metric
- *Verification has an inherent uncertainty*
  due to its dependence on the evaluation data set
- *Evaluation should contain a set of metrics*
  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 and in particular wind power in power markets. There are certainly many other tasks and applications of weather and power forecasts in the power 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. 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.

5.1.1 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 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 5.1.2). Errors of a well calibrated forecast model should always be scattered around zero. A frequency distribution that has a center 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.

5.1.2 Choice of Verification methods

When evaluating forecasts one or several evaluation methods or metrics to measure and compare the forecast performance have to be selected. 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.

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 5.1.3.

Dichotomous Event Evaluation

One may quantify desirable qualities 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 analyzing the number of correct positive, false positive, correct negative and false negative predictions of particular events [3], [1]. Table 5.1 provides an example table to carry out such categorical evaluations.

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

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 except of a comprehensive list of categorical statistics tests published by the Joint World Weather Research Program (WWRP) and Working Group Numerical Experimentation on Forecast Verification (WGNE) and provides the most common used metrics and their characteristics, relevant for forecast applications in the power industry. Details, equations and more comprehensive explanation on the use of these as well as references can be found (online) in [1]. It is recommended to apply these categorical statistics in particular for applications, where standard average metrics do not provide a measure of the true skill of a forecast to predict a specific event. In wind energy forecasting applications this is in particular important for extreme event analysis, ramping and high-speed shut-down 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 [14] and Richardson [10] 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.

Analyzing Forecast Error Spread with Box and Wiskers Plots

The box-and-whiskers plot is a visualization tool to analyze 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.

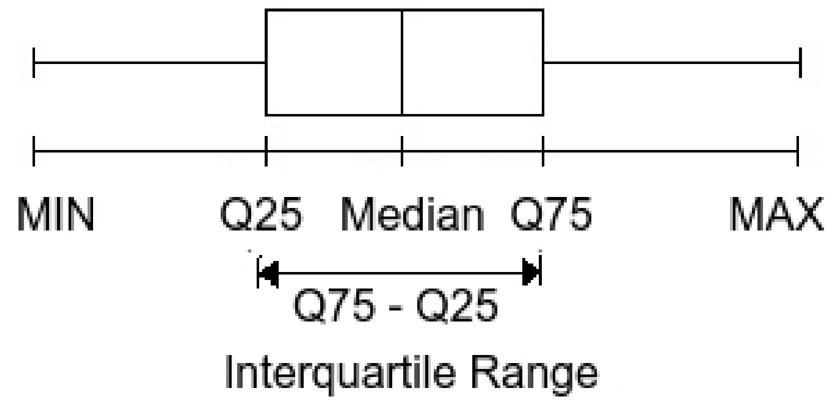

Figure 5.1: Principle of a box-and whiskers plot. The plot displays a five-number summary of a set of data, which is the minimum, first quartile, median, third quartile, and maximum. In a box plot, a box from the first quartile to the third quartile is drawn to indicate the interquartile range. A vertical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

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) the graphical or table 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 given margin [[6]]. In case (2) the error distribution of a forecast can be derived 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 analyze 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. [6] 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.

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

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 [**Madsen2005**].

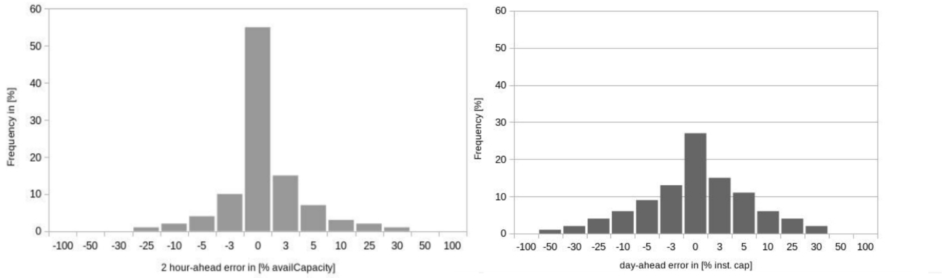

Figure 5.2: Examples of two histograms showing typical frequency distribution of errors for a 2-hour forecast horizon (left) and a day-ahead horizon (right).

5.1.3 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.

Best practice in this context is to following a procedure, where the evaluation/verification reflects the importance of a forecasts in it's role of the business processes and provides incentives for the forecast service provider to generate forecasts that fit the outlined (and verified) 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 application, the key time frames and a ranking of relative importance.

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: SDE, SDBIAS, StDev, VAR, CORR
   - 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 [6] and [**messner**], specific metrics and explanation of metrics can be found in [4], [15] for deterministic forecasts inclusive solar forecasting and for probabilistic forecast metrics in [12]. Significant tests can be found e.g. in [13].

Evaluation Matrix

Establishing an evaluation matrix is complex, but can be straight forward 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 5.1.3 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.

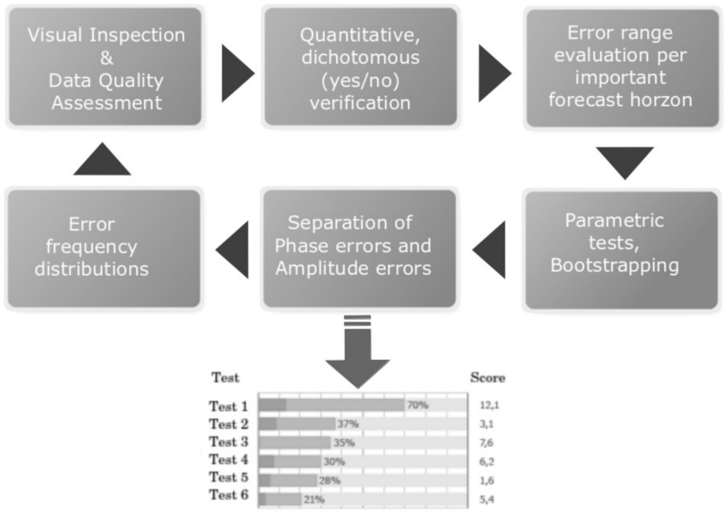
Visual Inspection & Data Quality Assessment
Quantitative, dichotomous (yes/no) verification
Error range evaluation per important forecast horzon
Parametric tests, Bootstrapping
Separation of Phase errors and Amplitude errors
Error frequency distributions
Test
Score
Test 1
70%
12,1
Test 2
37%
3,1
Test 3
35%
7,6
Test 4
30%
6,2
Test 5
28%
1,6
Test 6
21%
5,4

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.

Such a evaluation matrix provides important information in a comprehensive way and can be applied for comparisons as well as 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 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.

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 [1] 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 MAPE 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":
  1. subjective evaluation through visual inspection
  2. quantitative, dichotomous (yes/no) verification of critical events such as high-speed shut-down or ramps with e.g. contingency tables
  3. error ranges per important forecast horizon
  4. error ranges per hour of day or forecast hour
  5. error frequency distributions in ranges that have different costs levels
  6. separation of phase errors and amplitude errors according to their impact
  7. 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 a 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 useful 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 to not set one error target, but categorise errors instead e.g. large, medium and small errors. 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:

1. selection of relevant target parameters (see section 2.3)
2. selection of relevant metrics (see sections 5.2,5.1, 5.1.3, 5.4.1)
3. selection of relevant verification horizons (see section 2.2)
4. exclusion principles (see chapter 3 and section 3.2 and 3.5)

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, 5.1.1, 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 maximize the probability of selecting an optimal forecast solution for an application the performance evaluation uncertainty process should be minimized and non-performance attributes of a forecast solution should be effectively considered. Evaluation uncertainty can be minimized 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 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 of 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 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 is 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.

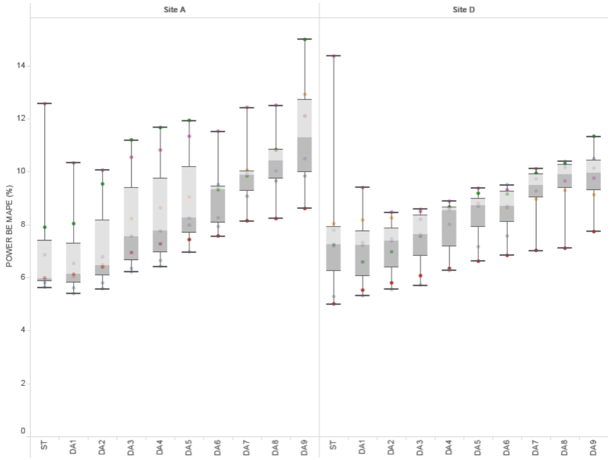

Figure 5.4: Example of a box-and-whisker-plot verification at two different sites (left and right panel) for different look ahead times (x-axis; DAx is $x^t h$ hour of day-ahead forecast) and mean absolute percentage error (MAPE; y-axis).

Figure 5.5 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.

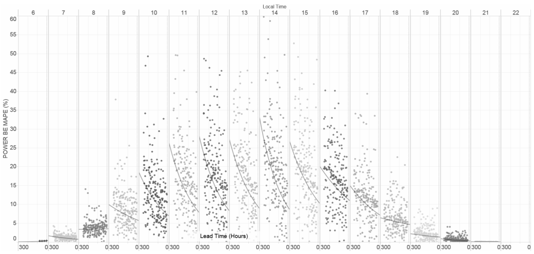

Figure 5.5: Example of a forecast error scatter plot by time of the day (top x-axis) for 3-hours lead times and forecast error (y-axis)

5.4 Evaluation of Development Techniques

Key Points
Keeping State of the Art in forecasting is an important aspect 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 exchange a forecast service provider is in no relation to the gain in forecast performance. Other cases of this type may be a statistical dependence of a or multiple forecasts going into a tool for further processing. 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 any forecast has a limit and the rate of improvement reduces. This needs to be taken into account equally much as the ability of a forecast solution for develop over time to keep 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 auto-regressive 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 favor 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 favored, 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 to 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 of 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 [13] are summarized. It compared different machine learning models for 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.

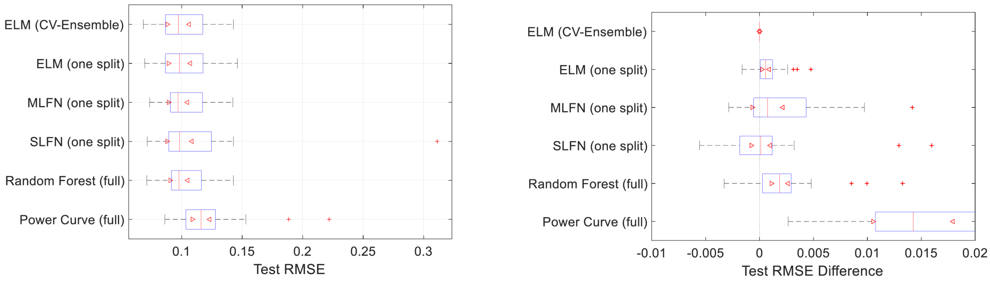

Figure 5.6: RMSE distribution for six different forecasting models forecasting for 29 wind farms in the North of Germany (left figure). Pairwise differences RMSE for each single model in comparison to the wind farm RMSE of the reference model ELM (CV-Ensemble) [13] (right figure).

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 carried out. In the ***Energy Trading and Balancing, ramping forecast in general and for reserve allocation****, forecasts are today a crucial part of the processes at balance responsible parties, but also 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 it is reflective of forecast performance and not externalities.

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:
   Wind power forecast errors are 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 wind 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 [8], and [7].

If trading decisions are based on probabilistic power production forecasts those forecasts should be evaluated as described in section 4.1.5. 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 [2] and p-variogram score [11]. 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.

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 [5].

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.

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 schedule reserve more dynamic, 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 solutions in order to reduce costs for reserve allocation for the end-user [9]. 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.

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 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 criteria 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 conform. 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.

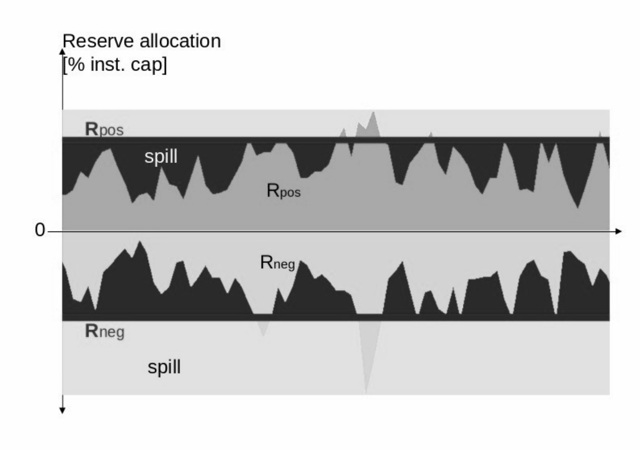

Figure 5.7: Illustration of the "reserve allocation dilemma" of costly spill versus covering all possible ramping events. Here, R_{pos} is the dynamic positive reserve, R_{neg} is the dynamic negative Reserve, the upper linear borders R_{pos} and R_{neg} are the static reserve allocation, the black area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

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:

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 [9] 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.

Table 5.3: Applied metrics in the evaluation matrix for the reserve allocation example in [9]. 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	

Bibliography

[1] WWRP/WGNE Joint Working Group on Forecast Verification Research. In: Berlin, Germany. URL: `http://www.cawcr.gov.au/projects/verification/`.

[2] Tilmann Gneiting et al. "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds". In: *TEST* 17.2 (2008), pp. 211–235. DOI: `10.1007/s11749-008-0114-x`.

[3] TM Hamill and J Juras. "Measuring forecast skill: is it real skill or is it the varying climatology?" In: *Q.J.R. Meteorol. Soc.* 132 (2006), 2905˘2923. DOI: `10.1256/qj.06.25`.

[4] Fowler T. Brown B. Lazo J. Haupt S.E. Jensen T. *Metrics for evaluation of solar energy forecasts*. Tech. rep. NCAR, 2006. URL: `http://opensky.ucar.edu/islandora/object/technotes:538`.

[5] Ravazzolo F Lerch S Thorarinsdottir TL and Gneiting T. "MForecaster's Dilemma: Extreme Events and Forecast Evaluation". In: *Statistical Science* 32.1 (2017), 106˘127. DOI: `10.1256/qj.06.25`.

[6] Kariniotakis G. Nielsen HA Nielsen TS. Madsen H. Pinson P. "Standardizing the Performance Evaluation of Short-Term Wind Power Prediction Models". In: *Wind Engineering* 29.6 (2005), 475˘489. DOI: `10.1260/030952405776234599`.

[7] 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*. URL: `http://http://download.weprog.com/WEPROG_Trading_strategies_EEG2012_ZEFE_71-2012-01_en.pdf`.

[8] Corinna Möhrlen, Markus Pahlow, and Jess U. Jørgensen. "Untersuchung verschiedener Handelsstrategien für Wind- und Solarenergie unter Berücksichtigung der EEG 2012 Novellierung". In: *Zeitschrift für Energiewirtschaft* 36.1 (2012), pp. 9–25. ISSN: 1866-2765. DOI: `10.1007/s12398-011-0071-z`. URL: `https://doi.org/10.1007/s12398-011-0071-z`.

[9] Jørgensen J.U. Möhrlen C. "Reserve forecasting for enhanced Renewable Energy management". In: Berlin, Germany, 2014. URL: http://download.weprog.com/Paper_WIW14-1035_moehrlen_joergensen_online.pdf.

[10] D.S. Richardson. "Skill and relative economic value of the ECMWF ensemble prediction system". In: *Quart. J. Royal Met. Soc.* 126 (2001), 649˘667.

[11] Michael Scheuerer and Thomas M. Hamill. "Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities∗". In: *Monthly Weather Review* 143.4 (2015), pp. 1321–1334. DOI: 10.1175/mwr-d-14-00269.1.

[12] *Towards the definition of a standardised evaluation protocol for probabilistic wind power forecasts*. Tech. rep. Anemos.Plus Project, 2012. URL: http://www.anemos-plus.eu/images/pubs/deliverables/aplus.deliverable_d1.3-protocol_v1.5.pdf.

[13] Braun A. Koch J. Jost D. Dobschinski R.J. Vogt S. "Benchmark of Spatio-temporal Shortest-Term Wind Power Forecast Models". In: Stockhlm, Sweden, 2018.

[14] D.S. Wilks. "A skill score based on economic value for probability forecasts". In: *Meteorol. Appl.* 8 (2001), 209˘219.

[15] Zhang. "A suite of metrics for assessing the performance of solar power forecasting". In: *Solar Energy* 111 (2015), 157.

Appendix A

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.

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

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.

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