Intelligent
Embedded Systems

12

André Gensler

Wind Power Ensemble Forecasting

Performance Measures and Ensemble Architectures for Deterministic and Probabilistic Forecasts

kassel university press

Intelligent
Embedded Systems

Band 12

Herausgegeben von
Prof. Dr. Bernhard Sick, Universität Kassel

André Gensler

Wind Power Ensemble Forecasting

Performance Measures and Ensemble Architectures for Deterministic and Probabilistic Forecasts

kassel university press

This work has been accepted by the Faculty of Electrical Engineering / Computer Science of the University of Kassel as a thesis for acquiring the academic degree of Doktor der Naturwissenschaften (Dr. rer. nat.).

Supervisor: Prof. Dr. Bernhard Sick, University of Kassel
Co-Supervisor: Prof. Dr.-Ing. Kurt Rohrig, University of Kassel

Defense day 21st September 2018

Bibliographic information published by Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at http://dnb.dnb.de.

Zugl.: Kassel, Univ., Diss. 2018
ISBN 978-3-7376-0636-3 (print)
ISBN 978-3-7376-0637-0 (e-book)
DOI: http://dx.medra.org/10.19211/KUP9783737606370
URN: https://nbn-resolving.org/urn:nbn:de:0002-406378

www.upress.uni-kassel.de

Printed in Germany

Abstract

This thesis describes performance measures and ensemble architectures for deterministic and probabilistic forecasts using the application example of wind power forecasting and proposes a novel scheme for the situation-dependent aggregation of forecasting models. For performance measures, error scores for deterministic as well as probabilistic forecasts are compared, and their characteristics are shown in detail. For the evaluation of deterministic forecasts, a categorization by basic error measure and normalization technique is introduced that simplifies the process of choosing an appropriate error measure for certain forecasting tasks. Furthermore, a scheme for the common evaluation of different forms of probabilistic forecasts is proposed. Based on the analysis of the error scores, a novel hierarchical aggregation technique for both deterministic and probabilistic forecasting models is proposed that dynamically weights individual forecasts using multiple weighting factors such as weather situation and lead time dependent weighting. In the experimental evaluation it is shown that the forecasting quality of the proposed technique is able to outperform other state of the art forecasting models and ensembles.

Zusammenfassung

Diese Thesis beschäftigt sich mit Qualitätsmetriken und Ensemblearchitekturen für deterministische und probabilistische Vorhersagen anhand des Anwendungsbeispiel der Leistungsvorhersage von Windkraftanlagen und stellt eine neuartige situationsabhängige Kombinationstechnik für Vorhersagen einzelner Vorhersagemodelle vor. Für deterministische Fehlermaße wird eine neue Kategorisierung anhand von Basisfehlermaß und Normalisierungstechnik eingeführt, die die Wahl einer geeigneten Evaluationsmetrik erleichtert. Weiterhin wird ein Schema für eine gemeinsame Evaluation von probabilistischen Vorhersagetechniken mit unterschiedlichen Repräsentationsformen vorgestellt, das die Kombination verschiedenster probabilistischer Vorhersagen ermöglicht. Auf Basis der analysierten Fehlermaße wird eine neue Aggregationstechnik für deterministische wie probabilistische Vorhersagen vorgestellt, die eine dynamische Gewichtung der Einzelvorhersagen anhand multipler Gewichtungsfaktoren wie einer wettersituationsabhängigen und vorhersagezeitabhängigen Gewichtung durchführt. In der experimentellen Evaluation wird gezeigt, dass die Qualität der vorgestellten Technik die von aktuellen “State of the Art” Vergleichsverfahren übertrifft.

Danksagung

Ich möchte mich ganz besonders bedanken bei Prof. Dr. rer. nat. Bernhard Sick, der mir langjährig mit Geduld und Rat zur Seite stand und mich in allen Phasen der Promotion unterstützt und gefördert hat. Weiterhin möchte ich Prof. Dr. Kurt Rohrig danken für die Zweitbegutachtung und die wertvollen Hinweise für die Anfertigung meiner Dissertation.

Ich möchte Stephan Vogt für die gute Zusammenarbeit und die gemeinsame Anfertigung von zwei wissenschaftlichen Artikeln danken, die mir maßgeblich bei der Anfertigung dieser Dissertation halfen. Vielen Dank an Adrian Calma, der sich stets die Zeit genommen hat, bei Problemen und zur Ideenfindung tief in die Thematik einzusteigen und damit zu vielen guten Einfällen beigetragen hat. Weiterer Dank geht an Christian Gruhl und Edgar Kalkowski, die mit vielen guten Ideen und auch bei technischen Fragen stets zur Seite standen. Vielen Dank auch an meine langjährigen Raumkollegen Benjamin Herwig und Janosch Henze, die mich ertragen durften und die ich stets mit zu vielen Fragen gelöchert habe. Natürlich auch vielen Dank an Martin, Tobi, Sven, Jens, Maarten, Daniel, Claudia und alle anderen vom Fachgebiet IES für die Hilfe, die gemeinsamen Ideen und viele gute Stunden.

Vielen Dank auch an Marcel Hahn, Ruben Jubeh und der ganzen enercast GmbH für die enge und erfolgreiche Kooperation im Projekt BigEnergy.

Zuletzt geht mein Dank an meine Familie und ganz besonders Amélie, die mir persönlich immer Mut und Unterstützung gegeben haben in allen Phasen der Promotion.

– Bernhard hat bei mir einen Stein im Brett. –

– Stephan weiß, wer der Schönste im ganzen Land ist. –

– Wenn Christian parallelisieren will, kein Problem. –

– Benjamin hält die Löffelchen-Stellung. –

– Sven trumpft auf. –

– Daniel hört zu hochohmig. –

– Biometrische Systeme sind gegen Adrian machtlos. –

– Martin serviert trendige Longdrinks in Bio-Gemüse. –

– Janosch hat eine Vorliebe für sibirische Staatsmänner. –

– Tobias ist Meister mittelalterlicher Klingen. –

– Edgar hat vom Tuten und Blasen Ahnung. –

– Maartens Größter wird jetzt eingeschult. –

– Jens. It's true. –

Contents

1 Introduction **1**
- 1.1 Renewable Energies . . . 2
- 1.2 Forecasting the Power Generation and Electric Load . . . 5
- 1.3 Overview of Power Forecasting Algorithms . . . 6
- 1.4 Challenges and Goals of this Thesis . . . 7
- 1.5 Overall Concept and Innovation . . . 9
- 1.6 Structure of this Thesis . . . 11
- 1.7 List of Relevant Publications . . . 12

2 Theoretical and Methodical Fundamentals **13**
- 2.1 Power Forecasting Nomenclature and Overall Forecasting Process . . . 13
- 2.2 Numerical Weather Predictions . . . 15
- 2.3 Wind Turbines . . . 18
- 2.4 Wind Power Forecasting . . . 20
  - 2.4.1 Power Forecasting in Operational Practice . . . 21
  - 2.4.2 Deterministic Point Forecasting and Probabilistic Distribution Forecasting 22
  - 2.4.3 Time Series Forecasting and Predictive Regression . . . 24
- 2.5 State of the Art in Deterministic Power Forecasting Models . . . 26
  - 2.5.1 Physical Models . . . 27
  - 2.5.2 Conventional Regressive Models . . . 27
  - 2.5.3 Machine Learning Models . . . 28
  - 2.5.4 Baseline and Hybrid Methods . . . 29
- 2.6 Ensemble Principles and Architectures . . . 30
  - 2.6.1 Ensemble Fundamentals . . . 30
  - 2.6.2 Basic Ensemble Techniques and Construction Principles . . . 33
    - Data Diversity . . . 34
    - Parameter Diversity . . . 34
    - Structure Diversity / Heterogeneous Ensembles . . . 35
    - Other Diversity Types . . . 37
  - 2.6.3 Ensembles for Power Forecasting . . . 37
    - Ensemble Prediction System (EPS) . . . 38
    - Multi-Model Ensembles . . . 39
    - Time-Lagged Ensembles . . . 39
  - 2.6.4 Analog Ensembles . . . 40
- 2.7 State of the Art in Probabilistic Forecasting . . . 41
  - 2.7.1 Probabilistic Forecasting Techniques from Single Predictor Models . . . 41
  - 2.7.2 Probabilistic Forecasts from Ensemble Techniques . . . 42
- 2.8 Quality Assessment . . . 43

2.8.1 Deterministic Error Measures . . . 43
2.8.2 Probabilistic Scoring Rules . . . 44
2.8.3 Statistical Forecast Validation . . . 45
2.9 Application Examples . . . 46
2.9.1 Unit Commitment, Economic Dispatch and Reserve Capacity Planning . 47
2.9.2 Economic Bidding Strategy for Wind Power Producers . . . 48
2.10 Need for Research . . . 49
2.11 Summary of this Section . . . 50

3 Metrics for Model Comparison of Deterministic Forecasts 51
3.1 Desired Error Score Properties . . . 51
3.2 Basic Error Measures . . . 52
3.3 Score Normalization Techniques . . . 54
3.4 Overview of Composed Error Scores . . . 55
3.5 Deviation Assessment, Correlation, and Model Comparison . . . 55
3.6 Case Study: Error Distribution Effects . . . 57
3.7 Case Study: Correlation of Error Scores . . . 60
3.8 Case Study: Discrimination and Abstraction Ability . . . 62
3.9 Discussion of Deterministic Error Scores . . . 63
3.9.1 General Error Score Properties . . . 63
3.9.2 Use of Normalizations . . . 64
3.9.3 Multiple Forecasting Time Steps and Deviation Assessment . . . 65
3.10 Summary of this Section . . . 66

4 Coopetitive Soft Gating Ensemble 68
4.1 The Coopetitive Soft Gating Weighting Function . . . 69
4.2 Evaluation of the Coopetitive Soft Gating Weighting Function . . . 71
4.3 Conclusion for the Coopetitive Soft Gating Weighting Function . . . 75
4.4 The Coopetitive Soft Gating Ensemble (CSGE) Algorithm . . . 78
4.4.1 Global Soft Gating . . . 80
4.4.2 Local Soft Gating . . . 81
4.4.3 Lead Time-Dependent Soft Gating . . . 84
4.4.4 Model Fusion and Ensemble Training . . . 84
4.4.5 Application Examples of the CSGE Technique . . . 85
4.5 Experimental Results . . . 87
4.5.1 Data Set Used for Evaluation . . . 89
4.5.2 Experimental Setup . . . 90
4.5.3 Case Study: Day-Ahead Performance on Single and Multiple Weather Forecasting Models . . . 91
4.5.4 Case Study: Intraday Performance on Single and Multiple Weather Forecasting Models . . . 94
4.5.5 Case Study: Performance Development Using a Varying Number of Weather Forecasting Models . . . 95
4.6 Properties of the CSGE Model . . . 98
4.7 Conclusion of this Section . . . 100

5 Probabilistic Forecasting Techniques 102
5.1 Prediction Spaces . . . 103
5.2 Representations of Predictive Distributions . . . 103
5.3 Desired Model Properties of Predictive Distributions . . . 106
5.4 Visual Verification of Predictive Distributions . . . 109
5.4.1 Visual Verification of Reliability . . . 109
5.4.2 Visual Verification of Sharpness . . . 111
5.5 Overview of Forms of Predictive Distribution Construction . . . 114
5.6 Predictive Distribution Construction from Single NWP Predictor Models . . . 114
5.6.1 Parametric Density Functions . . . 114
5.6.2 Kernel Density Estimation . . . 115
5.6.3 Analog Ensemble . . . 116
5.6.4 Quantile Regression . . . 117
5.6.5 Prediction Interval Forecasting . . . 118
5.7 Predictive Distribution Construction from Ensemble Predictors . . . 121
5.7.1 Density Function Sampling (EPS Ensemble) . . . 121
5.7.2 Distribution Fitting / Model Output Statistics . . . 122
5.7.3 Ensemble / Kernel Dressing . . . 124
5.7.4 Forecasting the Skill Category from Risk Indices . . . 125
5.8 Summary of this Section . . . 127

6 Evaluation Metrics for Probabilistic Forecasts 128
6.1 Scoring Rules and Score Decomposition . . . 128
6.1.1 Continuous Ranked Probability Score (CRPS) . . . 132
6.1.2 Ignorance Score (IGN / CRIGN) . . . 133
6.1.3 Quantile Score (QS) . . . 134
6.2 Experimental Evaluation . . . 136
6.2.1 Score Discrimination Ability . . . 136
6.2.2 Evaluation of Distributions Using the Quantile Score and Decomposition 139
6.2.3 Influence of Bias on Decomposed Scores . . . 140
6.2.4 Influence of Dispersion on Decomposed Scores . . . 142
6.2.5 Influence of Number of Quantiles . . . 142
6.2.6 Varying Parameter Characteristics of Probabilistic Forecasting Techniques 144
6.3 Discussion and Conclusion of Probabilistic Error Scores . . . 145
6.3.1 Score Applicability . . . 145
6.3.2 Decomposition Characteristics . . . 146

7 Probabilistic Coopetitive Soft Gating Ensemble 147
7.1 Combination of Probabilistic Forecasts . . . 147
7.2 The Probabilistic Coopetitive Soft Gating Ensemble Technique (PCSGE) . . . 148
7.3 Experimental Results . . . 150
7.3.1 Experimental Setup . . . 150
7.3.2 Case Study: Day-Ahead Performance on Single and Multiple Weather Forecasting Models . . . 153
7.3.3 Case Study: Intraday Performance on Single and Multiple Weather Forecasting Models . . . 155
7.3.4 Case Study: Reliability and Sharpness Properties . . . 157
7.3.5 Case Study: Overall Reliability and Sharpness Analysis . . . 159
7.4 Discussion and Conclusion of this Section . . . 162

8 Conclusion **163**
8.1 Summary of Contents of this Thesis 163
8.2 Directions of Future Research 164
8.2.1 Deterministic and Probabilistic Error Measures 164
8.2.2 Ensemble Techniques 164

A Acronym Definitions **170**

B Definition of Mathematical Symbols **172**

C Full Results of Error Score Distribution Comparison **173**

D Full Results of Analysis of Deterministic Error Scores **174**

E Proof: Quantile Score **177**

F Proof: Relationship of IS and QS **179**

G Proof: Minimum Value of Interval Score **182**

Chapter 1

Introduction

Overcoming the climate change arguably is one of the greatest challenges in human history. Global warming is a world-wide phenomenon that describes a steady increase in atmospheric temperature since the beginning of the industrial revolution in the second half of the 18th century. While some researchers believe global warming to be a natural phenomenon, the majority of the climate research community presumes climate change to have human causes due to the massive emittance of carbon dioxide and other greenhouse gases. In [222], the Intergovernmental Panel on Climate Change (IPCC) concluded that the higher concentration of greenhouse gases in the atmosphere can be attributed to human emittance activities. Furthermore, it is concluded that greenhouse gases are the cause for the rise of atmospheric temperatures.

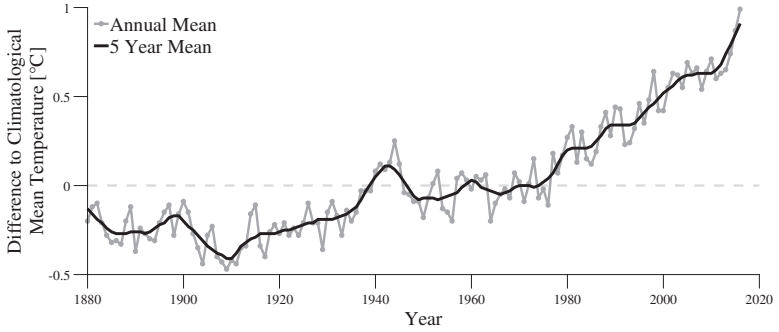


Figure 1.1: Development of global surface temperature since 1880. [170]

According to analyses of National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA) in [170, 171], the temperature in relation to the early 19th century has risen by about 1° C in 2017, as is being shown in Fig. 1.1. As the amount of energy in the atmosphere rises with increasing atmospheric temperatures, more extreme weather events can be observed frequently. Long-term effects of an increasing atmospheric temperature are, besides others, the rise of the sea levels, the loss of biodiversity, and the destruction of ecosystems. Furthermore, the frequency and strength of natural disasters, such as floods, droughts, or hurricanes, is increased. As is being described by the World Economic Forum in [153], climate-change induced factors, such as natural disasters, extreme weather events, ecosystem collapses, and the failure of climate-change mitigation of

the global community are seen as the most critical dangers for worldwide prosperity.

Following the Kyoto Protocol from 1997, in the 2015 United Nations Climate Change Conference (COP21) representatives of 195 states agreed to limit global warming to 1.5° C – 2° C compared to pre-industrial levels until the end of the century [233]. According to a recent report [51] of the German Meteorological Service (DWD), the COP21 goals can only be achieved by a timely and ambitious reaction of the international community. To accomplish this goal, the emission of carbon dioxide has to be reduced drastically. In order to achieve the COP21 targets, the technological development ideally is accompanied by market-oriented instruments which encourage environmentally conscious behavior, e.g., instruments such as the EU emission trading system [62].

Table 1.1: World-Wide CO_2 emissions in 2015 [154].

World	**Total**	**Coal**	**Oil**	**Gas**	**Other**	**Total**	**Coal**	**Oil**	**Gas**	**Other**
Total sectors (GT CO_2)	32.2	14.8	10.8	6.4	0.2	100 %	46 %	34 %	20 %	1 %
Power and heat generation	13.7	9.9	0.8	2.8	0.1	42 %	31 %	3 %	9 %	0 %
Other energy industry own use	1.7	0.4	0.5	0.7	0.001	5 %	1 %	2 %	2 %	0 %
Manufacturing industry[1]	6.1	3.9	1.0	1.2	0.04	19 %	12 %	3 %	4 %	0 %
Road transport	5.5		5.5	0.08		17 %		17 %	0 %	
Other transport[2]	1.8		1.7	0.1		6 %		5 %	0 %	
Residential sector	1.9	0.3	0.6	1.0	0	6 %	1 %	2 %	3 %	0 %
Other buildings[3]	1.5	0.3	0.7	0.5	0.005	5 %	1 %	2 %	1 %	0 %

[1] emissions from non-energy use excluded and feedstock use of fuel are excluded [2] includes aviation and international marine [3] includes agriculture and forestry

Table 1.1 gives an overview of worldwide carbon dioxide emissions. As can be seen from the table, over 32 Gigatons (GT) of CO_2 are being released from fossil fuel combustion in the observed year. Conventional fossil fuel combustion power plants contribute substantially to climate change, with 42 % of global CO_2 emissions being released in the power and heat generation sector. Therein, coal-fired power plants are responsible for 31 % of global CO_2 emissions, or over 13.7GT, in 2015 [62]. Also other categories, such as transportation, contribute significantly to worldwide emission of climate-relevant gases.

In order to reduce global warming, alternative forms of energy production which do not or only barely emit greenhouse gases have to be found. Nuclear power plants are a popular "clean" and efficient form of energy production without CO_2 emission, but this form of power generation has the drawback of problematic degradation products, and possible safety difficulties, as has been exhibited prominently by the disasters in Tschernobyl and Fukushima. Therefore, *renewable energies* (RE) have become a popular alternative to conventional power plants during the past decades.

1.1 Renewable Energies

Renewable energies (RE) generate electric power from renewable resources. Typical power plants of this type are photovoltaic panels, water mills, biogas-fired power plants, or wind turbines. Each form of RE has its justification, the optimal power plant choice in many cases depends on the particular location of the power plant and other external conditions. All of these forms of RE have been researched and installed globally with varying support from the local governments, and already contribute substantially to the energy mix of certain power grids. For instance, within the European Union, over 142 GW of installed wind power capacity has been constructed until the end of 2015 [63]. In particular in Germany, RE have achieved a relatively large contribution to the overall power generation due to strategic subsidies in

the RE sector. According to the Fraunhofer Institute for Wind Energy and Energy System Technology (IWES), RE were the most important source of electricity for the first time in 2014 [68]. The overall power generation in Germany is shown in Fig. 1.2.

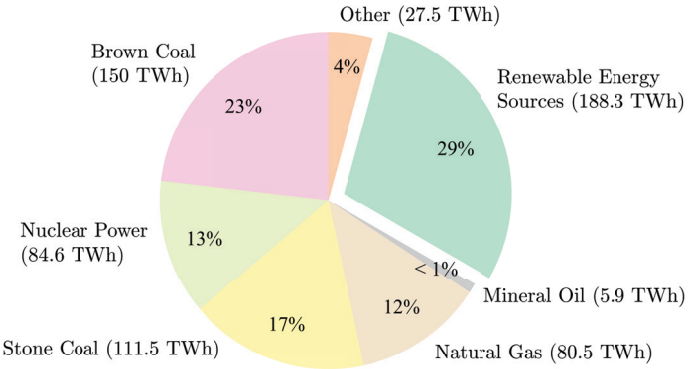


Figure 1.2: Gross Power Generation of Germany in 2016. [1]

As is illustrated in the figure, while around 188 TWh of produced energy is generated from RE power plants, still 40 % of the power is generated using coal-fired power plants. The German ministry of economics (BMWi) aims to increase the percentage of RE to at least 40 % in 2025, and 80 % in 2050 [29]. Furthermore, the goal is to shut down the last nuclear power plants until the year 2022. Thereby, it is planned to emit 40 % less greenhouse gases in 2030 in comparison to 1990. Within RE, wind power has by far the largest electrical infeed into the power grid with 41 % of total RE power generation, as illustrated in Fig. 1.3. Photovoltaic and biomass power also contribute substantially to the power mix with 24 % and 20 %, respectively.

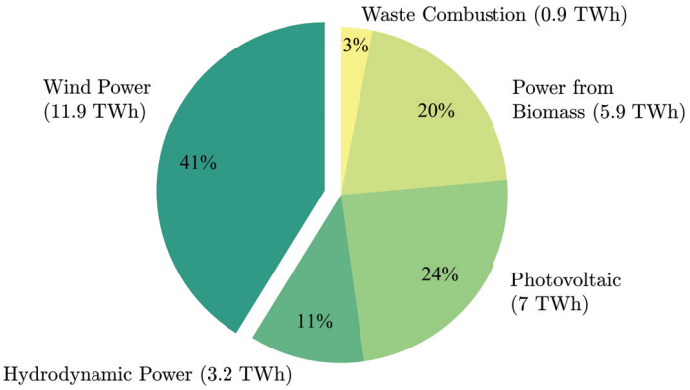


Figure 1.3: Gross Renewable Power Generation of Germany in 2016. [1]

On a theoretical level, renewable energies have a number of advantages over conventional sources of energy:

- Renewable energies are available everywhere on the world in practically unlimited quantity. They therefore can be created decentralized, which reduces the costs of transportation of primary energy sources, such as coal or oil. Furthermore, the political dependency on particular regions of the world is decreased.

- Many forms of renewable energies can easily be made accessible. The principle of creation is straight-forward, and in many cases does not pose any security risks unlike nuclear or combustion based power plants.

- In contrast to conventional centralized power plants, renewable energy power plants typically are distributed, smaller power plants. Therefore electrical energy does not have to be transported over longer distances, but can be utilized near the area of creation, or for self-sustaining facilities and residential areas.

- Renewable energies and conventional power plants complement each other, where conventional power plants close the gap between energy demand and production by renewable energy power plants.

- Job opportunities for specialists for construction and maintenance of renewable energy power plants are created all over the world. For instance, in Germany 371000 jobs were located in the renewable energies sector in 2013 [30].

These advantages partly are the reason for the success of RE over the last decades. However, there are also a number of disadvantages which are characteristic of RE power plants:

- Many forms of RE are intermittent, they depend mainly on the weather and therefore cannot be controlled in the same way it is possible with conventional fossil fuel combustion power plants. Typical intermittent RE are photovoltaic panels and wind turbines.

- The power production of RE typically takes place in rural areas, which historically were not designed for the massive infeed of electrical power. This complicates the design and maintenance of power grids, and possibly requires a dramatic reconstruction of the existing grid structure.

- Power plant and grid maintenance are more complicated due to the more distributed locations of power plants.

- Some forms of RE (e.g., wind turbines) only can be used effectively in particular regions of a country, which may lead to the construction of additional (trans)national power lines.

- Domestic political problems may arise when power infrastructure is created near residential areas.

- The destruction of biological ecosystems may occur if the power infrastructure is created within nature reserves or in bird migration routes.

Having laid out the challenges that arise from the massive electrical infeed from renewable sources of energy, it becomes apparent that the power grid operation becomes more complex. In particular the volatile power generation characteristics of RE power plants create challenges for a safe and reliable power grid operation which is mainly guaranteed by quantifying the

amount of future power generation of RE power plants. The need for forecasting the future state of RE power plants and the power grid therefore is discussed in the following section (Section 1.2).

1.2 Forecasting the Power Generation and Electric Load

In order to transmit the generated power, power plants are connected to a power grid, an interconnected network for the delivery of electrical power from power producers to consumers. Traditional power grids consist of power generating power plants, a transmission grid, and power consumers. In a power grid, the consumed electrical power has to equal the power demand in order to function properly, as the intermediary storage of electrical energy is both inefficient and expensive (though this aspect is becoming increasingly important). Traditionally, power grids were constructed in hierarchical fashion from the power plant to consumers, and power plants were non-intermittent. As all forms of power plants have certain ramp-up characteristics when changing the target power generation, the power grid can not react instantly to changes in the demand of electrical consumers. While industrial consumers have to declare the planned consumption of large quantities of electrical energy, this solution is not feasible for the entirety of residential consumers and small enterprises. Therefore, power plant operators *forecast* the future electrical demand. This process is commonly known as *load forecasting*. The electrical demand is mainly influenced by the time-dependent factors, such as time of day, week-day, season, and whether there are holidays, all of which have particular demand characteristics. In markets with a high percentage of electrical heating systems, weather and temperature-dependent factors also play an important role.

While forecasting the electrical demand therefore has been a challenge for some time, the significant infeed of electrical power from RE led to an indisputable complication of the management of the power grid. As the most widespread forms of RE have intermittent generation characteristics, not only the load, but also the production of the power has to be predicted. This process is called *power forecasting*.

Conventionally, the power demand has been seen as a largely static process (i.e., it can barely be influenced). The power generation therefore has to react to the power demand. Some effort in the industry and research community is focused on the intelligent balancing of power production and power demand. Besides other functions, these *smart grids* coordinate the use of electrical energy by controlling the points in time when electrical consumers are operated. Thus, the power demand is no longer a static process, but can be modified depending on the grid conditions. It is thereby possible to operate electric consumers at those points in time when a lot of RE power is available. This leads to an increased efficiency in power production. The realization of smart grids, however, is a lengthy process, as a significant number of electric consumers have to be equipped with the technology for smart grids. Furthermore, universal protocols for the cooperation of smart devices have to be defined. Additionally, smart grid technology mainly profits from the availability of large-capacity energy storage, which most likely will not be realized at least until the broad availability of connected electric vehicles. In the following, we will concentrate on the process of power forecasting.

The production of energy from RE power plants is a time-varying process. A forecasting system therefore should be able to create a (time-dependent) forecast over time. Depending on the time between the creation of the forecast (the *forecasting origin*) and the maximum time for which a forecast is made (the *forecasting horizon*), the forecast is of use for different actors in the industry. For instance, a forecast can be of use for the planning of the power reserve to

ensure grid stability, for electricity trading activities, or for power plant operation planning, also known as unit commitment. For RE, the power production is mainly influenced by the present weather situation. Many forecasting systems are therefore based on a prediction of the future weather situation. Forecasting a future state is a difficult task in general, especially when the prediction has to be created for an uncertain future, which clearly is the case for weather forecasting. Forecasting the future generation of produced power therefore is a highly complex problem with huge economic importance. Sophisticated power forecasting algorithms are needed in order to precisely forecast the development of generated power. The actual demand for highly-precise forecasting algorithms in the power industry is emphasized by a recent report carried out by the transmission system operator Tennet [67] that claims that Tennet has paid more than one billion euros in 2017 for emergency actions to guarantee grid stability with upward tendency.

1.3 Overview of Power Forecasting Algorithms

Precise and reliable power forecasting algorithms are needed for the safe and economical operation of power grids with electrical infeed from intermittent renewable energies. Typical power forecasting algorithms include physical models and statistical approaches which include models from machine learning.

Physical models explicitly compute the generated power from a RE power plant, e.g., for a wind turbine using a wind turbine power curve. These physical models consider the relationship of different physical factors, such as wind speed, air density, and the characteristics of the wind turbine. While these models are very well understandable and yield good results under ideal circumstances, physical models often oversimplify the complex nature of the generation process (e.g., do not consider wind turbulences) and are prone to systematic errors in the weather forecast, location data, or facility data of the power plant.

Statistical approaches and approaches from machine learning (ML) do not try to understand the underlying physical process, but only model the input-output relationship of (multiple) explanatory variables or predictors (e.g., the weather forecast and additional other data) to a predictand or target (the respective generated power). Therefore, unlike physical models, these approaches require historic data from both weather forecasts and power generation as basis for the creation of a power forecasting model. However, by only modeling the input-output relationship, the underlying complex physical processes do not have to be explicitly modeled. “Classic” statistical models investigate correlations of explanatory variables and targets in order to create forecasts. Autoregressive and averaging techniques for time series forecasting, such as variants of the ARMA model, are popular methods in this category. Models from machine learning are related to statistical models, and include methods such as (multi-)linear regression techniques, artificial neural networks, and support vector-based methods. Methods from machine learning often are *black box* methods, which means that the processes within the forecasting model (though they technically can be observed) are so complex that they are not intuitive for a human to interpret. However, while some ML methods may have weaknesses regarding explicability, their complex structures may allow for more precise forecasts.

Research and industrial practice has shown that the aggregation of multiple models is able to improve the overall forecasting accuracy for power forecasting tasks (e.g., [206, 226, 227]). These aggregated forecasts are called ensembles. Ensembles can be formed in a number of ways, for instance by aggregating multiple weather models, or power forecasting models.

Power forecasting algorithms can either issue a (deterministic) point forecast, or a (probabilistic) distribution forecast. Both forms of forecast have their eligibility, the choice of the forecasting paradigm depends in many cases on the particular task and the user of the forecast result.

Given the fact of permanently increasing installed capacity of renewable energies in the power grid, the challenges for the power grid operation and the electricity market soar, as, though the relative forecasting error may remain constant, the *absolute* effects of the errors increase (at least if the errors are correlated). If the overall grid operation is more dependent on RE power plants, the quality of forecasting algorithms has to improve in order to guarantee stable grid operation. Therefore, there is a demand for more sophisticated forecasts with higher accuracy and a better ability for uncertainty assessment. This thesis aims at addressing these questions based on a thorough analysis of error metrics which are appropriate for power forecasting. As an application example, forecasting the power generation of *wind turbines* is chosen. In the following section, the challenges and goals of this thesis are outlined in detail. Therein, the aspects of the creation and evaluation of more sophisticated forecasting algorithms using model combination techniques regarding deterministic and probabilistic forecasts are considered specifically.

1.4 Challenges and Goals of this Thesis

This section describes the broad challenges and goals of this thesis. Therein, the principal research questions are laid out and described.

1. What are appropriate metrics for quality assessment of forecasting algorithms for deterministic and probabilistic forecasts?

The clear definition of suitable error metrics is the basis for performance assessment and forecasting model comparison. While there are a number of established error metrics for power forecasting which have been transferred from conventional time series forecasting, some characteristics of power forecasting make certain error scores more attractive than others. Depending on the desired application, error measures should be able to resolve different properties of forecasting models and wind turbines. For instance, for wind farm expansion planning error metrics should be able to capture the expected average performance over time. For power forecasting *model comparison*, however, error measures should be able to exhibit two properties.

- Error measures should be able to *discriminate* well performing forecasting models from weaker models. This is important for model selection of a set of forecasting models which can potentially be used for a given time period and the same forecasting task.

- Error measures should be able to perform an *abstraction* from the difficulty of the forecasting task. Thereby, the comparison of the forecasting model performance of a number of forecasting models on different time periods and wind farm locations with different predictability is better possible.

So far, there has been no critical analysis of performance measures for power forecasting regarding these two properties. As part of this thesis, the behavior of the most common performance measures is analyzed in the area of deterministic forecasts.

For probabilistic forecasts, the performance assessment is even more challenging, as probabilistic forecasting systems have to fulfill the two properties of *sharpness* and *reliability* in contrast to point forecasts, which solely have to create forecasts that are close to the actual observations. Reliability refers to the correct assessment of the model spread, while sharpness refers to how narrow a predictive distribution is. Error measures for probabilistic forecasts should be able to evaluate those two properties of probabilistic forecasts.

These research questions are addressed in Section 3 for deterministic forecasts, and in Section 6 for probabilistic forecasts.

2. How can the strengths of single power forecasting models and weather models be found?

Research has shown that the combination of forecasting models which can be weather forecasting models or power forecasting models to ensembles can improve the forecasting accuracy. The single models that form the ensemble may have systematic strengths and weaknesses depending on external factors, such as the input of explanatory variables. For RE forecasting, these external factors are in particular different weather situations or varying lead times (time between forecasting origin and the point in time for which the forecast is created). Finding those strengths is non-trivial, as they depend on multiple factors. The following steps are required for a system that exploits the strengths of individual models depending on the particular forecasting situation.

- A method for determining the lead time-dependent quality of power forecasting models has to be developed.
- A technique for assessing the forecasting quality has to be found, which depends on the form of explanatory variables, such as particular form of weather situation, or facility data of the wind farm.
- The technique has to be robust against overfitting and should include a smoothing of the weights, e.g., by including knowledge about the overall performance of an algorithm.

A novel ensemble technique that addresses this research question is proposed in Section 4.

3. What is an optimal weighting strategy of ensemble members for power forecasting models and weather models?

The methods that form an ensemble are in most cases combined using a weighted aggregation of the models. This is typically constructed using a weighting or a gating approach. In a weighting approach, the ensemble technique assigns an overall static weight to each ensemble member which is then used to form the combined ensemble forecast. In this approach, ensemble members *cooperate* in creating the overall forecast. An alternative method for ensemble combination is using a dynamic *gating* of the ensemble members. In this variant, the ensemble members try to win a *competition* in which the forecast of the winning forecasting algorithm is accepted as the ensemble forecast. The challenge of this technique is to define an appropriate gating function, i.e., to find a criterion on which to perform the competition.

While these aggregation methods are able to better utilize the strengths of each forecasting model in comparison to a simple averaging of models, the possibilities of model combination still are not used to their full extend. In the basic framework of cooperation and competition, methods are not able to be aggregated in a combined approach which uses both properties in the sense of a *coopetition* of forecasting models. A research question therefore is how a

number of ensemble members can be combined to optimally use the information which are contained within each ensemble member to increase the forecasting accuracy. Optimality in this case means the capability to create accurate weights for a number of cases while only introducing a minimum number of parameters. This research question is interwoven with research question 2., as this form of combination should be able to be used with regard to a number of different external factors such as lead time or weather situation. This research question is also addressed in Section 4.

4. How can probabilistic forecasting models be combined optimally in an ensemble?

The forecasting of renewable energies is based on weather forecasts, while weather can be regarded as a stochastic process. A weather forecast therefore always is a forecast for an uncertain future. As the power generation depends on the weather situation, the power forecast consequently also has an uncertainty attached. Deterministic point forecasts are not optimally suited for the estimation of the forecasting uncertainty, therefore, probabilistic forecasts (or distribution forecasts) are increasingly employed to retain optimal decision-making performance under uncertain conditions.

The combination of probabilistic power forecasts, analogously to deterministic forecasts, can increase the probabilistic forecasting accuracy. Based on the combination scheme which is defined in research question 2., this research question aims at defining a framework for ensemble member combination for probabilistic forecasts. The proposed scheme has to fulfill additional requirements which are requirements to probabilistic forecasts, such as reliability and sharpness. This research question is addressed in Section 7 by proposing an extension of the technique described in Section 4 for probabilistic forecasts. In order to aggregate probabilistic forecasts, a unification of different forms of probability representation has to be performed which is analyzed in Section 5.

1.5 Overall Concept and Innovation

This section describes the overall concept and the innovation of this thesis.

The evaluation of power forecasting algorithms is the basis for model comparison and wind farm evaluation. In many cases, standard metrics are used for the evaluation of deterministic point forecasts. More advanced metrics, however, may be better suited for the investigation of specific properties of forecasting algorithms. In a first step, the most common metrics for performance assessment of deterministic forecasts are analyzed. A novel categorization of error scores is proposed, which categorizes error scores by their basic error measures (e.g., absolute errors, squared errors), and by their form of normalization. The advantages of each categorization variant are highlighted based on a number of artificial and real-world experiments. The suitable error score for a target application can then easily be chosen based on the individual advantages.

The use of probabilistic forecasts can be more optimal for certain decision-making tasks in power forecasting. While probabilistic forecasts have high utility for these tasks, their evaluation is not trivial. Probabilistic forecasts can be represented in a number of ways, e.g., as density functions, quantile forecasts, or intervals, which can be computed from single forecasting models or different ensemble types, using a wide variety of different approaches. There exist a multitude of *scoring rules* which aim at evaluating probabilistic forecasts. A probabilistic forecasting system should issue *reliable* and *sharp* forecasts, both of which are statistical properties which can only be evaluated on a series of forecast-observation

pairs, leading to an overall complex evaluation task. Furthermore, as scoring rules typically are specialized on the evaluation of a particular form of uncertainty representation, the comparability of probabilistic forecasts is hindered. Therefore, an approach to create a common form of uncertainty representation for probabilistic forecasts is proposed. This allows for the evaluation of probabilistic forecasts with different uncertainty representation and also is a prerequisite for the *combination* of probabilistic forecasts.

Based on a thorough discussion of the evaluation metrics for both deterministic and probabilistic forecasts, a novel ensemble scheme called *Coopetitive Soft Gating Ensemble* (CSGE) is proposed, which aims at improving the forecasting accuracy of a set of forecasting algorithms through model combination. The technique is designed to work flexibly on an arbitrary number of weather forecasting models (WM) and power forecasting models (PM), which are constructed in a hierarchical fashion as shown in Fig. 1.4.

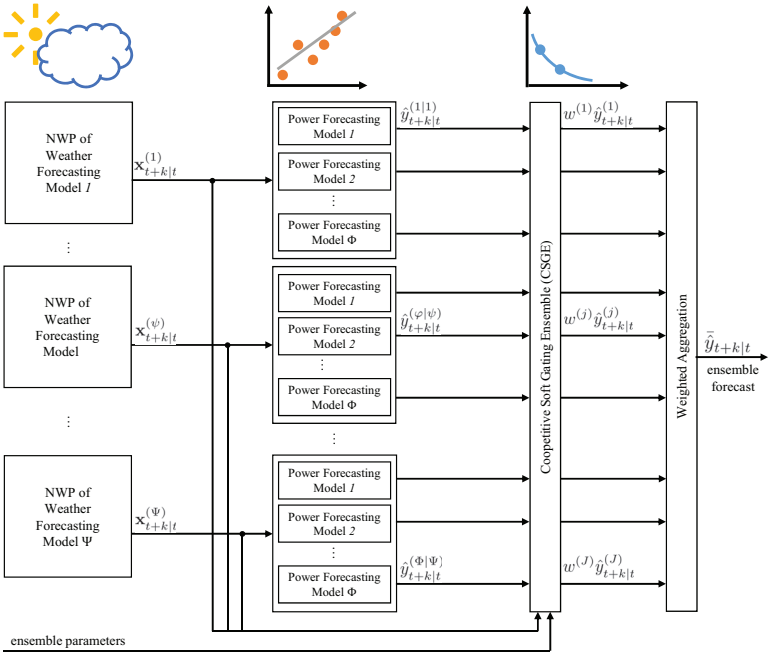


Figure 1.4: The proposed Coopetitive Soft Gating Ensemble scheme. A detailed explanation of the CSGE technique and this figure can be found in Section 4.4.

The idea is as follows: A number of PM are each trained on a single WM to create a number of deterministic forecasts. Then, each PM forecast based on a particular WM is weighted using a novel weighting technique which is called coopetitive soft gating. The idea of the soft gating technique is to gradually weight a number of ensemble members according to their historically observed performance in various situations. It thereby combines the aspects of ensemble weighting and ensemble gating. The coopetitive soft gating function is used to weight each PM-WM combination regarding the three weighting aspects global weighting, local (weather situation-dependent) weighting, and lead time-dependent weighting. These

weights are computed for each WM and for each PM, leading to a combination of six weighting factors in total per PM-WM combination. The technique is designed to work flexibly with deterministic and probabilistic forecasting algorithms. In a series of experiments it is shown that the proposed ensemble technique is able to yield superior results in comparison to other forecasting algorithms for both deterministic and probabilistic forecasts on different forecasting time horizons, such as the day-ahead or the intraday forecasting horizon.

1.6 Structure of this Thesis

In Section 2, the theoretical fundamentals in meteorological sciences, wind power forecasting, and ensemble methodology are explained. The state of the art in power forecasting algorithms and ensemble techniques is highlighted. Further details of the area of probabilistic forecasting and forecasting evaluation methodology are given. Two application examples for power forecasting algorithms are described. The section closes with a discussion of the need for research in the area of evaluation methodology and ensemble techniques.

Section 3 describes the area of forecast evaluation for deterministic point forecasts. A novel categorization of error metrics is laid out which simplifies the process of choosing the appropriate error metric for a forecasting task. In an experimental evaluation, the behavior of the presented scores is investigated regarding the properties of discrimination of models with different forecasting quality and abstraction from the difficulty of the forecasting task.

In Section 4, a novel ensemble methodology for the combination of multiple weather and power forecasting models is introduced. The technique is able to aggregate a set of forecasts using a multi-factor weighting technique. The overall weighting therein includes weighting components which depend on the lead time and on the particular weather situation. The ensemble technique is constructed to be able to combine a set of deterministic forecasting models.

Section 5 introduces a scheme for the unification of the evaluation of probabilistic forecasts. Probabilistic forecasts typically are expressed in a number of ways, such as prediction intervals, quantiles, or density functions, which are not directly comparable. In order to improve the comparability of different probabilistic forecasting algorithms, a conversion of their representation to a common representation is proposed. This is also a prerequisite for the aggregation of probabilistic forecasts in an ensemble.

Section 6 investigates the area of evaluation of probabilistic forecasts. After highlighting the most popular evaluation metrics for probabilistic forecasts, the characteristics of these scoring rules are investigated in detail. This investigation includes the decomposition of scoring rules.

Section 7 applies the findings of Sections 5 and 6 to construct the ensemble technique proposed in Section 4 in a probabilistic framework. Therein, a set of probabilistic forecasts is combined to create a refined probabilistic forecast. In the evaluation, probabilistic forecasts based on multiple weather forecasts are used for day-ahead and intraday forecasting time spans. The analysis also includes an investigation of the characteristics regarding the reliability and sharpness properties.

Finally, Section 8 summarizes the main results of this thesis and gives an outlook to future directions of research.

1.7 List of Relevant Publications

The following publications directly emerged from the work on this thesis.

- A. Gensler, B. Sick, and S. Vogt. A review of deterministic error scores and normalization techniques for power forecasting algorithms. In *Proceedings of the 8th IEEE Symposium Series on Computational Intelligence (SSCI16)*, pages 1–9, Athens, Greece, 2016, [86]
- A. Gensler, B. Sick, and V. Pankraz. An analog ensemble-based similarity search technique for solar power forecasting. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC16)*, pages 2850–2857, Budapest, Hungary, 2016, [85]
- A. Gensler and B. Sick. Forecasting wind power – an ensemble technique with gradual coopetitive weighting based on weather situation. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN16)*, pages 4976–4984, Vancouver, Canada, jul 2016, [81]
- A. Gensler, S. Vogt, and B. Sick. Metaverification of Uncertainty Representations and Assessment Techniques for Power Forecasting Algorithms including Ensembles. *Renewable & Sustainable Energy Reviews*, 96:352–379, 2018, [88]
- A. Gensler and B. Sick. A Multi-Scheme Ensemble Using Coopetitive Soft Gating With Application to Power Forecasting for Renewable Energy Generation. *ArXiv e-prints*, 1803.06344:1–22, 2018, [84]
- A. Gensler and B. Sick. Probabilistic Wind Power Forecasting: A Multi-Scheme Ensemble Technique With Gradual Coopetitive Soft Gating. In *Proceedings of the 9th IEEE Symposium Series on Computational Intelligence (SSCI17)*, pages 1803–1812, Honolulu, USA, 2017, [83]

In [86], the role of deterministic error measures is analyzed. The results of Section 3 can be found in this article. The outcomes of Sections 5 and 6 regarding the forms of representation of probabilistic forecasts and their evaluation is described in [88]. For the design of the proposed ensemble technique of Section 4, the foundational work regarding the weighting function and the principle of weather dependent weighting has been shown in [81]. More insights on the similarity measurement of weather situations regarding feature selection and weighting is given in [85]. An optimization strategy and the inclusion of the lead time-dependent weighting within the ensemble is described in [84]. Finally, the probabilistic realization of the ensemble technique that forms Section 7 is shown in [83].

Furthermore, there are some publications which are related in a broader sense for this thesis. In [79], a number of (deep) neural network architectures are investigated regarding the forecasting of photovoltaic power plants. Other forecasting applications such as the path prediction of vulnerable road users are analyzed with the joint use of neural networks and combinations of systems of orthogonal basis polynomials in [100]. Properties of these used polynomials are inspected in [78], analyses regarding their run-time are performed in [77], the time series classification ability is studied in [87]. Further applications of the polynomial approximations are the detection of events in time series that has been performed in [82] based on the measures introduced in [80]. These events could also be defined to be useful in the context of power forecasting, such as for the detection of extreme meteorological events.

Chapter 2

Theoretical and Methodical Fundamentals

This chapter details the theoretical and methodical fundamentals in the area of power forecasting. The nomenclature for power forecasting is laid out in Section 2.1, the area of numerical weather predictions (NWP) and meteorological ensemble prediction systems is detailed in Section 2.2. Wind turbines are described in Section 2.3, followed by an introduction to wind power forecasting in Section 2.4. A literature overview of deterministic power forecasting models is given in Section 2.5, ensemble principles and literature is detailed in Section 2.6. An overview of the state of the art in probabilistic forecasting techniques is given in Section 2.7. An overview of performance assessment metrics is given in Section 2.8. This chapter closes with some application examples of power forecasting algorithms in Section 2.9 and an analysis of the need for research in Section 2.10.

2.1 Power Forecasting Nomenclature and Overall Forecasting Process

This section describes the most common terms for power forecasting and summarizes the commonly used variables. Table B (on page 172) also summarizes the nomenclature.

Power forecasting deals with forecasting the power generation denoted as y of intermittent RE power plants. The symbol $\hat{y}$ indicates a deterministic point forecast of the power generation. On a wide variety of applications, explanatory variables are commonly utilized for creating a forecast. For power forecasting of renewable energies, these explanatory variables typically are a numerical weather prediction (NWP) denoted as $\mathbf{x}$. A forecasting model f can then use the explanatory variables $\mathbf{x}$ to create a forecast $\hat{y}$. For the evaluation of forecasting algorithms, the forecasts $\hat{y}$ are typically compared to a number of actual power measurements o (observations). Probabilistic forecasts create a predictive density function over the power generation y which is denoted as $\hat{p}(y)$.

In time series forecasting, there has to be a clear definition of the possible points in time frequently referred to, which is visualized in Fig. 2.1. The time from which is being forecasted is called t, the *forecasting origin*. From this origin, a forecast for a number of *forecasting time steps* called $t+k$ is issued with varying *lead time* denoted as k, which describes the time between the forecasting origin and the forecasting time step. Depending on the desired application, a forecast is performed for a number of lead times

$$k \in \left(k_{\min}, k_{\min} + 1{\cdot}\Delta, k_{\min} + 2{\cdot}\Delta, \ldots, k_{\max}\right). \tag{2.1}$$

The lead times $k_{\min}$ and $k_{\max}$ as well as Δ, which is a fixed time increment, can be chosen

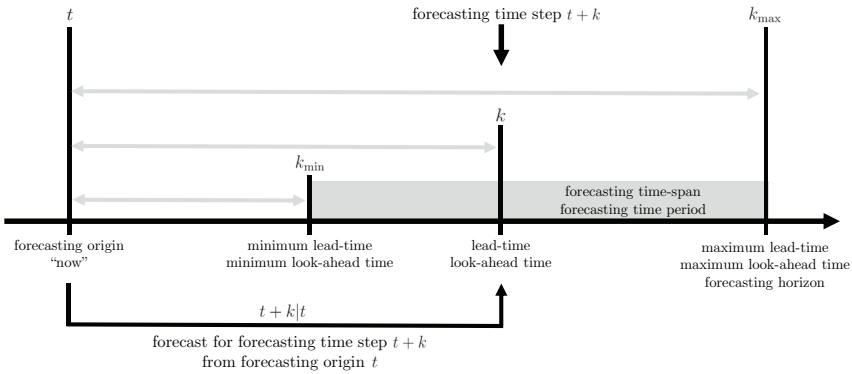


Figure 2.1: Nomenclature of important points in time for power forecasting.

arbitrarily. For typical applications, such as the hourly day-ahead forecast, the forecasting time steps are chosen to $k_{\min} = 25$ h, $k_{\max} = 48$ h, $\Delta = 1$ h. For an hourly intraday forecast, on the other hand, typical borders are $k_{\min} = 1$ h, $k_{\max} = 24$ h, $\Delta = 1$ h.

A forecast is performed in the *forecasting time span* between the *minimum lead time* called $k_{\min}$, and $k_{\max}$, the *forecasting horizon*. To refer to a forecast which is issued from a forecasting origin t for a particular forecasting time step $t+k$, the notation $t+k|t$ is commonly used. A typical deterministic power forecast therefore can be denoted as

$$\hat{y}_{t+k|t} = f(\mathbf{x}_{t+k|t}|\boldsymbol{\theta}), \tag{2.2}$$

where f is the power forecasting model and $\boldsymbol{\theta}$ are the governing model parameters in this case.

The overall data flow in a power forecasting system is shown in Fig. 2.2. The process can be grouped into the process of model creation and model application. In the model creation phase, models are either created using human expert knowledge, such as in physical models, or using learning models, e.g., statistical or machine learning models. The latter use historic NWP measurements $\mathbf{X}_{\mathrm{H}} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and corresponding historic power measurements $\mathbf{y}_{\mathrm{H}} = (y_1, \ldots, y_N)$ to train a model given a model structure and a set of hyperparameters. Some of those models may also use facility data as parameters. Physical models, on the other hand, are mostly based on facility data to create the wind turbine power curve. For all model types, the result of model creation is a power forecasting model which can be used to create power forecasts given an input of predictors, such as NWP forecasts.

Having created a power forecasting model, it can then use NWP input data $\mathbf{x}_{t+k|t}$ for the desired forecasting time step to create a power forecast $\hat{y}_{t+k|t}$. Optionally, some power forecasting models use a feedback of previous power forecasts (e.g., $\hat{y}_{t+k-\Delta|t}$) to estimate the current power generation. This is particularly frequent for forecasting for short lead times. In a final stage, the created forecast is used for model evaluation or in actual power systems operation.

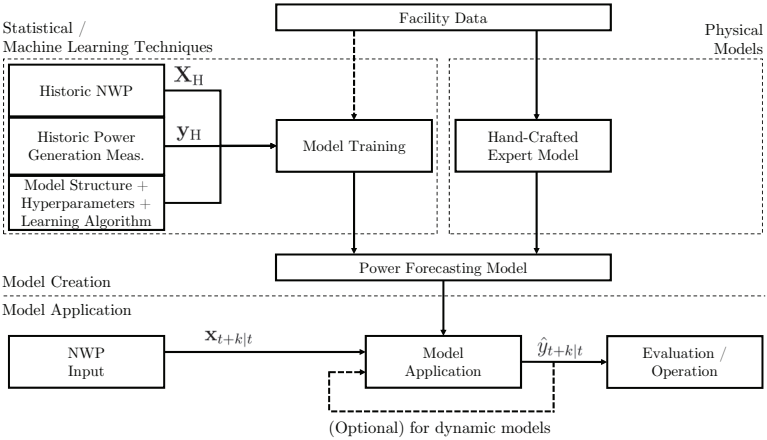


Figure 2.2: Overall Data Flow in a Forecasting System.

2.2 Numerical Weather Predictions

The mathematical models which create the weather forecasts for a particular forecasting time step are called numerical weather predictions (NWPs). These predictions are able to forecast a wide variety of meteorological variables, for instance, wind speeds, humidity, precipitation, or the formation of clouds. For power forecasting NWPs are of relevance, as for many time horizons, power forecasting algorithms are based on a weather forecast.

The earth's atmosphere is a (heterogeneous) fluid. The idea of the creation of an NWP is to measure the state of the fluid at a particular point in time and use a model to predict the future state of the fluid at some time in the future. An NWP is typically processed by a weather model provider using computer simulations. The NWP generation can be seen as a two-step process, which is visualized in Fig. 2.3.

In the first step, the initial atmospheric conditions are measured as starting point for the simulation. This process therefore is called *initialization*. The initialization process is performed using topographic terrain maps, weather satellites, radiosondes in weather balloons, measurement buoys, but also using measurements from aircraft and ships along frequent routes [6]. The World Meteorological Organization (WMO) aims at standardizing the meteorological observation practices around the globe, e.g., with the hourly METAR weather report [172], or the six-hour SYNOP observations [232].

Based on the initialization, the *dynamics* of the atmosphere are modeled at given locations and altitudes. The future state of the atmosphere is then computed using a set of nonlinear partial differential equations, the so-called primitive equations. The most important elements of primitive equations contain

- the continuity equation, which represents the conservation of mass,
- the conservation of momentum, that describes the hydrodynamical flow, and
- the thermal energy equation, that describes the temperature flow to temperature sinks,

which are described, e.g., in [112, 234]. Having a measured state of the atmosphere from the initialization process, and knowing the process fluid dynamics, the process of weather

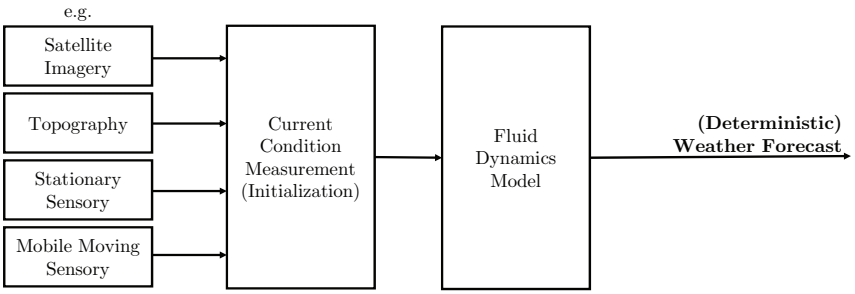


Figure 2.3: Process of the generation of a numerical weather prediction.

forecasting may seem like a deterministic process, but it is not. First, the current state of the atmosphere cannot be measured with unlimited precision. Second, the differential equations can typically not be solved analytically, therefore, approximate solutions are computed. Fluid dynamics furthermore exhibit chaotic behavior, therefore, even small errors grow with time and double about every five days [44], limiting the predictability on longer time horizons. Weather (forecasting) therefore is a *stochastic process*, i.e., the input measurements have an uncertainty attached, furthermore noise is injected into various stages of the forecasting process in an ensemble prediction systems which is detailed below. It therefore only makes sense for current weather models to predict up to a certain number of days, e.g., six days for the UKMET weather model [163] or ten days for the ECMWF HRES [61] model. There are a variety of weather models, which differ in the forecasting horizon, the frequency of model runs, the temporal and spatial resolution, and which meteorological quantities are predicted.

Weather models are computed on different scales in order to resolve different weather phenomena. This corresponds to the *spatial resolution* of a weather model. An overview of the most relevant meteorological scales is shown in Table 2.1.

Table 2.1: Scales of Atmospheric Processes [180].

Scale Name	**Grid Length**	**Forecasting Horizon**	**Resolution of**
Synoptic Scale[4]	2000 km +	1 day - 1 month +	Low and High Pressure Areas Precipitation Areas Weather Fronts
Mesoscale (Alpha)	200 km - 2000 km	1 day - 1 month	Squall Lines Mesoscale Convective Complexes Tropical Cyclones
Mesoscale (Beta)	20 km - 200 km	1 h - 1 day	Sea Breezes
Mesoscale (Gamma)[5]	2 km - 20 km	1 min - 1 hour	Thunderstorm Convection Complex Terrain Flows
Microscale[6]	up to 2 km	up to 1 min	Cloud Puffs Small Cloud Features Turbulences

[4] also: Large Scale, or Cyclonic Scale [5] also: Storm-Scale [6] also: Misoscale

Depending on the scale of the weather model, it can be used to resolve different weather aspects. In general, small atmospheric scales correspond with detailed high-frequency forecasts. The chaotic nature of fluid dynamics causes these forecasts to only be able to be useful on short forecasting horizons. With increasing grid length, longer-term trends can be

forecasted.

A *weather model run* describes the computation and operational delivery of a weather forecast. For each weather model run, the most current available weather observations are incorporated in the forecast. The higher the number of model runs, the more current the observations are on average. In particular for short-term forecasting this is a desirable property. The frequency of weather model runs is chosen by the weather model provider. For instance, the DWD ICON model runs four times per day [50]. Each weather model run therein has its own forecasting origin.

While weather models compute their model dynamics on fine time increments for the computer simulation, the results typically are represented in a more coarsely sampled *temporal resolution*. This resolution typically corresponds with the meteorological scale and consequently the forecasting horizon. For synoptic scales, forecasts are often in the range of hours, while on micro-scales higher frequencies are provided. The temporal resolution of mesoscale forecasts are in the range of minutes up to hours.

Some weather models are not computed for the entire globe, but for certain regions only. This greatly reduces the required computing power, and may yield better results, if, for instance, more dense observations can be acquired in the smaller area. However, these models often have to explicitly include a global weather model that influences the weather within the considered area (from the "edges").

The process of NWP generation in this section describes the creation of a deterministic NWP forecast. In order to better model the uncertainty of the weather prediction, meteorological ensemble prediction systems (EPS) are used which are described in more detail in the following.

Meteorological Ensemble Prediction Systems

A popular technique to model the stochastic process of weather forecasting is using *ensemble prediction systems* (EPS). The idea of an EPS is to vary the initial conditions and the fluid dynamics model (and optionally the global weather model, e.g., in the case of the DWD COSMO-EPS model [204]) in order to create a number of varying NWPs. The process is visualized in Fig. 2.4. In the figure, the upper branch visualizes the conventional deterministic forecasting process, as also shown in Fig. 2.3, yielding a deterministic point forecast. For the generation of an EPS, a *perturbation* of the initial state of the weather model as well as for the fluid dynamics model parameters is introduced.

As indicated in the previous section, the current state of the atmosphere cannot be measured with unlimited precision. Therefore, a perturbation of the initial states aims to create an estimation of the actual (non-observable) state of the atmosphere by adding a form of noise (or perturbation) to the measurements, yielding an *initial condition ensemble*.

To introduce further spread into the model to better approximate the true future state of the atmosphere, the parameters of the fluid dynamics model are also altered by a perturbation. Each of the members of the initial state ensemble is processed using a number of fluid dynamic model parameterizations. This leads to the creation of the final *weather forecast ensemble* of the EPS.

An EPS aims at creating a model of the underlying (non-observable) probability distribution of possible weather outcomes. The members of the weather forecast ensemble can be seen as samples drawn from this probability distribution. In principle, this distribution is continuous, however, the continuous structure can hardly be realized in the simulation process itself. The multivariate nature of the forecasting process further complicates the

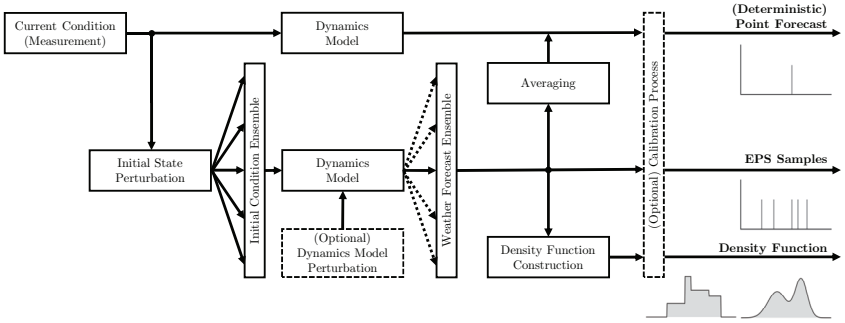


Figure 2.4: Process of NWP creation from weather measurements.

direct computation of a probability distribution. While there are initiatives for dense sampling of the underlying probability distribution using ultra large ensembles directly (e.g., [156]), the weather forecasting ensemble is sometimes post-processed to create density functions. Post-processed density functions can be continuous using ensemble dressing techniques [218] or discrete by creating a stepwise cumulative density function (cdf) from the drawn samples [108]. Another possibility is to again create a refined deterministic forecast from the weather forecast ensemble by averaging or median filtering [251].

The goal of an EPS is to create a model spread that has a correct model *dispersion*, i.e., it does not have too little spread (underdispersive) nor too much spread (overdispersive). Methods for the investigation of model dispersion (e.g., probability integral transform (PIT) diagrams) are detailed in Section 6. If the raw weather forecast ensemble has some form of systematic error, it can be corrected using *calibration techniques*. Typical methods for the calibration of weather forecasts can be simple bias correction, but also more advanced methods, such as ensemble model output statistics (EMOS) [96], Bayesian model averaging (BMA) [199], or ensemble copula coupling (ECC) [217]. Furthermore, in the power domain some popular calibration techniques include sigmoid calibration [195] or isotonic regression [255].

EPS forecasts create a forecast for a particular forecasting time step. A generalization of the EPS methodology to tempo-spatially consistent *scenarios* is proposed in [192]. A scenario is a forecasted possible realization of the development of a weather situation. Scenarios are in many cases computed from probabilistic forecasts while considering the interdependence structure of the forecasting errors among multiple lead times.

2.3 Wind Turbines

Wind turbines are a type of RE power plant that generate electrical energy from kinetic wind power. Wind turbines exist in different variants in vertical and horizontal axis types, and in different electrical capacities. The most popular form, horizontal-axis wind turbines (HAWT), are placed on top of a tower. The rotor shaft and the electrical generator sit on top of the tower, and are pointed into the wind in order to generate electrical energy. An example for a HAWT is shown in Fig. 2.5.

Figure 2.5: Example of a HAWT wind turbine.[1]

The generated power y of wind turbines can be computed from the physical formula

$$y = \frac{1}{2} A \rho c_p v^3, \tag{2.3}$$

where A is the rotor area which can be computed by the rotor blade radius r with $A = \frac{1}{2}\pi r^2$, ρ is the air density, c_p is a aerodynamical rotor efficiency coefficient (or drag coefficient), and v is the wind speed. As can be seen from the formula, given a wind turbine (and thus fixed area A and coefficient c_p), the main influencing factor for the power generation is the wind speed v. An exemplary wind turbine power curve is shown in Fig. 2.6.

As can be seen from the figure, below a certain wind speed, the wind force is not able to overcome the friction of the wind turbine, and therefore, does not generate any power[2]. Above the *cut-in wind speed* (or turn-on wind speed), the wind turbine starts to produce energy in the partial load range. In this range of wind speed, the wind turbine blade's pitch angle is set for maximum drag. At the *nominal wind speed*, the wind turbine approaches the nominal power generation capacity of the wind turbine. If the wind speed is further increased, the pitch angle of the wind turbine blades are increasingly bended to reduce drag (the aerodynamical rotor efficiency coefficient c_p is reduced), the value of c_p thus is a function of the wind speed v itself. Thereby, the maximum rotation speed of the wind turbine generator is not exceeded. If the *cut-out wind speed* is exceeded for a certain time period, the wind turbine blades are turned in feathering position (90° pitch) to create minimum drag. This process is supplemented by brakes. A wind speed not shown in the figure is the *survival wind speed*, which is the maximum wind speed the wind turbine is designed to resist. Depending on the expected maximum wind speeds within a particular region, wind turbines are designed for different survival wind speeds.

In practice, wind turbines are often aggregated into arrays, which are referred to as *wind farms*. These wind farms typically have a common point of infeed into the power grid. Often,

[1] Image from pixabay.com under Creative Commons CC0 Public Domain license. URL: `https://pixabay.com/en/landscape-wind-turbine-sky-blue-1814599/`. Last accessed 2018-01-03.

[2] Other reasons include that the internal power consumption of the wind turbine in operation would not justify the operation, therefore brakes supplement the halting process.

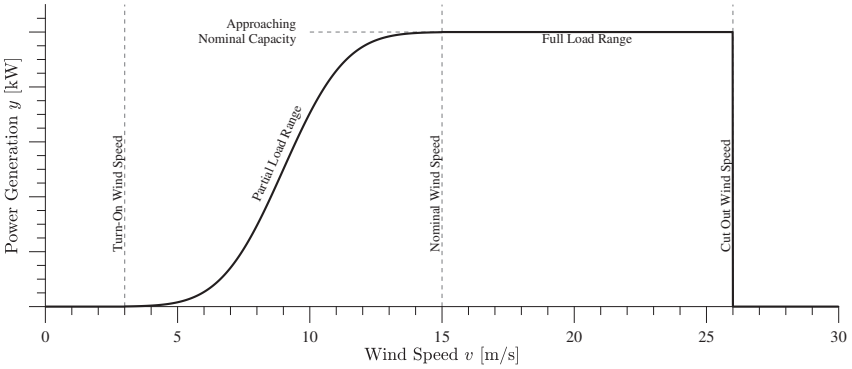


Figure 2.6: Exemplary wind turbine power curve depending on the wind speed.

wind farms are also aggregated logically into wind farm *portfolios*. The goal of portfolios is to enable a better predictability of the portfolio in the forecasting process, as the (relative) forecasting error typically decreases with increasing distributed power capacity. This effect becomes particularly noticeable when the wind parks are scattered over a larger area where error canceling effects occur.

One could argue that the task of wind power forecasting can be reduced to a wind speed and air density forecasting problem. While this is true in principle and is partly applied in the industry when using physical forecasting models, the actual power generation differs significantly when naively using the NWP forecast to estimate power generation using Eq. 2.3. This can be due to the dependence of the wind speed from the angle in which the wind arrives and other factors, such as, e.g., the topology which may have influence on the actual wind speeds. Furthermore, shadowing effects within a wind farm may occur, where one wind turbine is located in the wind shadow of another wind turbine. These and further effects are discussed in Section 2.4.

2.4 Wind Power Forecasting

The production of energy from RE power plants is a time-varying process, a forecasting system therefore should be able to create a (time-dependent) forecast over time. Depending on the time between the creation of the forecast (the forecast origin) and the maximum time which is being forecasted (the forecasting horizon), the forecast is of use for different actors in the industry. A categorization of different forecasting horizons with possible applications is laid out in Table 2.2.

As can be seen from the table, possible applications of weather forecasts vary widely depending on the forecasting horizon. The explanatory variables for the forecasting process also happen to be very different. For instance, for very short-term horizons, the current power generation of the wind farm is of critical importance for accurate predictions. Here, persistence-related methods and dynamic models (models which have a recurrent feedback of the predictand as predictor) play a critical role. Short-term and medium-term time horizons arguably have been the most well-researched forecasting horizons. For forecasts of this type, numerical weather predictions are the most important predictors. Typical methods for this

Table 2.2: Relevant forecasting horizons in power systems [259].

Forecasting Horizon	Time Scale	Application
Very short-term	Seconds to minutes	Wind turbine control Power system frequency control
Short-term	Minutes to days	Economic dispatch Power reserve planning Day-ahead electricity market
Medium-term	Days to weeks	Unit commitment Maintenance scheduling
Long-term	Weeks to months or years	Wind power planning Power systems planning

type of forecasting horizon are wind turbine power curves, machine learning models, or techniques from statistics. Popular time horizons in operational practice, such as the intraday or day-ahead forecast, are also part of this category. Long-term forecasting is often performed using long-term seasonal climatology models, as weather forecasts cannot be created reliably for this forecasting horizon. In this thesis, the focus is in particular on the short-term horizon, i.e., the intraday and the day-ahead forecast.

2.4.1 Power Forecasting in Operational Practice

In operational practice, the intraday and the day-ahead forecast are a special case of forecasting. As previously mentioned, for the hourly day-ahead forecast, the time step parameters are $k_{\min} = 25$ h, $k_{\max} = 48$ h, $\Delta = 1$ h, for hourly intraday forecasts, the parameters are $k_{\min} = 1$ h, $k_{\max} = 24$ h, $\Delta = 1$ h. In many markets, the time increments are chosen even shorter, e.g., to $\Delta = 15$ min. Therein, the computation of the forecasts is based on the weather model (WM) runs (see Section 2.2 for details). The power forecasting models typically incorporate the most recent weather model run as input for the forecast. An overview of the operational forecasting process is given in Fig. 2.7.

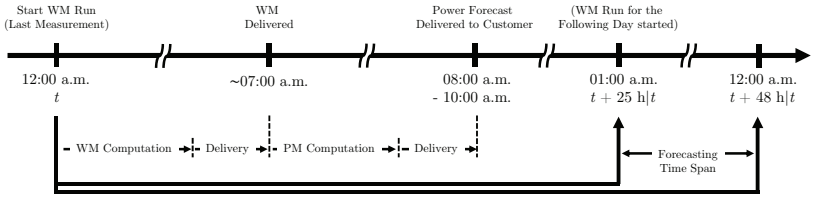


Figure 2.7: Example of operational delivery of power forecasts using the ECMWF global weather model.

The figure shows the typical time spans for an operational day-ahead forecast. In the following, the ECMWF IFS model is taken as an example to make the time horizons more concrete and thus understandable. The processing chain starts with the computation of the NWP forecasts. At that time, the most recent weather condition measurements are used as initial condition determination of the weather model. Many popular weather models (including the ECMWF IFS model) perform a weather model run at 12:00 a.m. UTC (coordinated universal time). Therein, the NWP forecasts for a certain time span are computed,

for the ECMWF IFS model this is up to ten days. This is a highly computationally expensive process that takes several hours on mainframe computers. Furthermore, the result of the computation is extensive, as depending on the model the global or regional atmospheric conditions are individually stored for a number of time steps. The transfer of the data from the weather model provider to the power forecasting institution therefore also takes some time. For the ECMWF IFS model, the overall process takes about seven hours. Having the weather forecast (and a trained power forecasting model), the power forecast can be created. After the power forecast computation (and optionally the verification by a human forecasting expert), the forecast is provided to a client. The power forecast is a much more compact result in comparison to the weather forecast. Typical day-ahead forecasts are delivered between 8:00 a.m. and 10:00 a.m. local time. The power forecasts therein have been computed for the whole following day from 01:00 a.m. to 12:00 a.m. The client can then plan his activities (such as economic decision-making processes) for the following day depending on the information of the power forecast.

As is visible for the figure, the forecasting system is a *rolling forecast*, which means that the time periods in the forecasting process overlap. For each additional weather model run during the day, more recent forecasts can in principle be created (which is made, for instance, for the intraday forecast). For the day-ahead power forecast, the weather computation is started anew at 12:00 a.m. every day. Therefore, the time periods overlap. The power forecasts, on the other hand, do not overlap for day-ahead forecasts. However, for the intraday forecast overlaps are common, since new intraday power forecasts are created multiple times per day.

2.4.2 Deterministic Point Forecasting and Probabilistic Distribution Forecasting

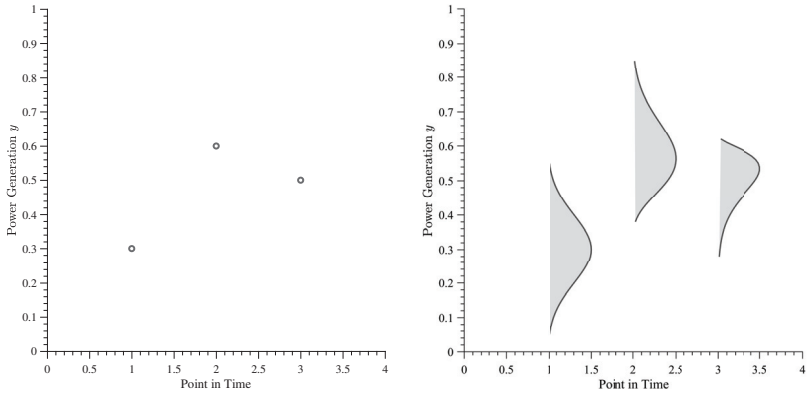
There are two basic principles for issuing forecasts for future points in time. These basic principles are point forecasts and distribution forecasts.

A point forecast, also called deterministic forecast, computes a defined "crisp" point estimate for a future point in time. This is the result of a predictive function f that computes a forecast $\hat{y}$ with

$$\hat{y}_{t+k|t} = f(\mathbf{x}_{t+k|t}, \boldsymbol{\theta}), \tag{2.4}$$

where $\mathbf{x}_{t+k|t}$ are the explanatory variables, which are typically NWP forecasts, and $\boldsymbol{\theta}$ are the model parameters. Point forecasts have the appeal of an easily interpretable result and compatibility with the theoretical framework of common machine learning approaches which return a single numerical value. If the process to be forecasted is of deterministic nature and can in theory be predicted with unlimited accuracy, then point forecasting clearly is the correct theoretical approach. An example of point forecasts issued at every whole-numbered point in time is shown in Fig. 2.8.1.

If, however, the process for which a forecast will be created is stochastic (which is the case for forecasts based on weather forecasts), then the more appropriate framework is one which issues predictive distributions rather than points. These distributions should ideally reflect the uncertainty of the underlying stochastic process. Predictive distributions can be expressed in a number of ways, we will denote a predictive probability density function (pdf) as $\hat{p}_{t+k|t}(y)$. A set of density forecasts for every whole-numbered point in time is laid out in Fig. 2.8.2. The density functions represent the probability of the actual "true" value being at power value y with probability density of $\hat{p}_{t+k|t}(y)$. Distribution forecasts therefore are the more general representation of a forecast. In principle, a deterministic forecast is a special case of a probabilistic forecast assuming a Dirac / Delta function at the location of



(1) Deterministic (point) forecast. (2) Probabilistic (distribution) forecast.

Figure 2.8: Schematic differences between point forecast and distribution forecast.

the deterministic forecast, i.e., when $\hat{p}_{t+k|t}(y) = \delta(y - \hat{y}_{t+k|t})$. The corresponding cumulative density function (cdf) can be denoted as as $\hat{P}_{t+k|t}(y)$ with relation

$$\hat{P}_{t+k|t}(y) = \int_{-\infty}^{y} \hat{p}_{t+k|t}(y')\mathrm{d}y'. \tag{2.5}$$

In particular the cdf representation is useful in practice, as the actual probability (denoted as P) of the actual measurement o_{t+k} being below a certain threshold $\hat{y}^{(\tau)}_{t+k|t}$ can be expressed in the form

$$P(o_{t+k} < \hat{y}^{(\tau)}_{t+k|t}) = \tau. \tag{2.6}$$

The value of $\hat{y}^{(\tau)}_{t+k|t}$ therein can be computed using the parameter $\tau \in [0,1]$ (which specifies the point in the cdf with probability mass below $\hat{y}^{(\tau)}_{t+k|t}$ and is often referred to as the τ-quantile of the cdf) and the inverse cdf in the form

$$\hat{y}^{(\tau)}_{t+k|t} = \hat{P}^{-1}(\tau). \tag{2.7}$$

One could question why deterministic forecasts have a justification at all for power forecasting, which clearly is a forecast based on a stochastic process. There are a number of reasons for both deterministic and probabilistic power forecasts having their applications. The following reasons give arguments for the use of deterministic forecasts for power forecasting:

- Probably the most important aspect is that the decentralized electricity market operates on deterministic point forecasts. For instance, power grid operators have to make guarantees about grid stability and electricity traders sell absolute quantities of energy on energy exchange markets. This is not just a historically developed phenomenon, but actually makes sense for the grid operation from a grid stability standpoint.

- While not representing the expected uncertainty of a forecast directly, the error of deterministic forecasts can nevertheless be expressed using error metrics such as mean-

absolute error (MAE), root-mean squared error (RMSE), or standard deviation of errors (SDE). A detailed analysis of error measures in the deterministic domain can be found in Section 3.

- Toolboxes for creating deterministic predictive models are widely available and technically mature. Furthermore, deterministic models are much easier to train, as they solely have to provide accurate predictions. Probabilistic forecasts, on the other hand, have to be both sharp and reliable. Probabilistic forecasts therefore eventually sacrifice sharp forecasts in order to achieve statistically sound (reliable) forecasts. In other words, deterministic forecasts may, on average, yield more precise results. More details on required properties of probabilistic forecasts can be found in Section 6.

- The idea and working principle of deterministic models is easier understandable and can therefore be offered to a wider audience and people that are not domain experts.

Then again one could argue why probabilistic forecasts at all provide a benefit over deterministic forecasts. The main advantages of using probabilistic forecasts are detailed in the following:

- Distribution forecasts allow for an assessment of the uncertainty of a particular forecast not only on a statistical basis, but for every forecast on an individual basis. Distribution forecasts can issue heteroscedastic uncertainty (i.e., varying expected error in different weather situations and on varying lead times), whereas deterministic forecasts assume homoscedastic behavior (same expected error in each situation) of the forecasting error.

- Distribution forecasts can be used for optimal decision making given an uncertain future. Given asymmetrical costs of over- or underestimation, probabilistic forecasts can be used to retain optimal performance given an uncertain future. More details on optimal decision making can be found in Section 2.9.2.

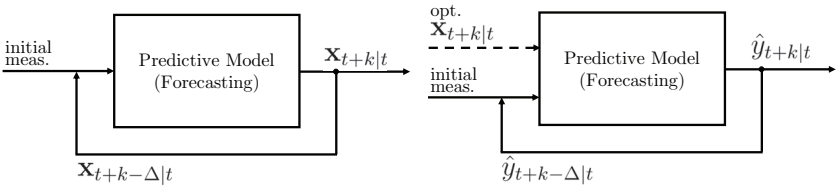
- The tails (i.e., locations with marginal probabilities) of the probability density function can be used for the investigation of extreme errors which may be of importance for guarantees of grid stability. An application example for the estimation of the reserve capacity is laid out in Section 2.9.1.

In conclusion, both deterministic and probabilistic forecasting techniques have their eligibility in operational practice. The particular choice of the suitable technique depends on the area of application, the target audience, and the willingness to dedicate the resources into the process of deploying and maintaining an operational probabilistic forecasting system.

2.4.3 Time Series Forecasting and Predictive Regression

Power forecasting is often seen as a time series forecasting task. While this is correct for the overall prediction task, the single stages in the forecasting process can be described in more detail and are referred to under different terms.

Time series forecasting is a subdomain of regression for predicting the future state of a target variable depending on the (set of) past observations. Therein, typically the predictand of a predictive model is used as input for this model to forecast the *consecutive* time step, possibly with the help of further explanatory variables. These types of algorithms therefore are iterative in many cases when creating multi-step ahead predictions. The task of time series forecasting is shown on a schematic level in Fig. 2.9.



(1) Forecasting of an NWP forecast which is a multivariate variable. The predictand of the previous forecasting step is used as input in the consecutive step. For the initialization, actual measurements are used.

(2) Forecasting of a univariate variable (e.g., the power generation). The predictand of the previous forecasting step is used as input in the consecutive step. Other explanatory variables may be used optionally.

Figure 2.9: Schematic representation of time series forecasting for univariate and multivariate variables with examples.

Typically, the model input and model output are the same quantity (e.g., for power forecasting, both model input *and* model output are the power generation). For power forecasting, very short-term models are often conventional time series forecasting models. They take the (set of) current power measurements and extrapolate them into the future, potentially with the help of further explanatory variables. These models therefore are also referred to as dynamic models, as the prediction of the models depends on the current internal state of the model itself. This example is shown in Fig. 2.9.2. The NWP forecasting process also is a multivariate time series forecasting process, such as shown in Fig. 2.9.1.

"Classic" regression is defined by a mapping of a set of explanatory variables to a continuous target variable. In many cases, no temporal information are considered in the regression process (though temporal information may be encoded, e.g., in a sliding time window for regressive models, see Section 2.5.2). This form of prediction is shown in Fig. 2.10. In power forecasting, the transformation from an NWP to a power generation value is a regression task.

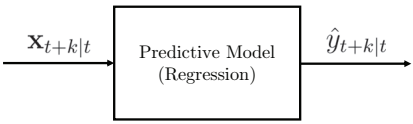


Figure 2.10: Schematic representation of regression. The mapping between explanatory variable and target is created for the same point in time.

The overall process in power forecasting for short and medium-term time horizons is a two-step process, the first of which is the NWP generation, whereas the second is the forecast of the power generation. According to the nomenclature laid out above, this is a combined time series forecasting and regression task. The process is visualized in Fig. 2.11, where the first step creates an NWP forecast $\mathbf{x}_{t+k|t}$, whereas the second step of the process converts the predicted NWP forecast to a power generation forecast $\hat{y}_{t+k|t}$. As in many cases the NWP is assumed to be given (as the actual computation of an NWP is very resource-intensive), the forecasting task is one in the sense of a predictive regression. This predictive regression can be formulated to create a point forecast or a distribution. More details on these forms of forecasting are detailed in the following section.

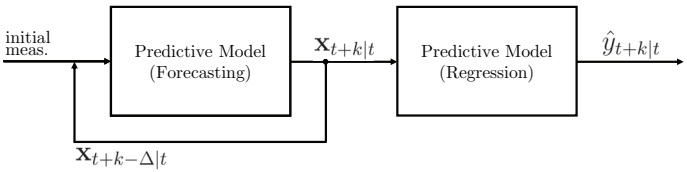


Figure 2.11: Schematic representation of power forecasting. The process consists of a forecasting step (the creation of the NWPs) and a regression step to create the mapping from NWP to power generation.

2.5 State of the Art in Deterministic Power Forecasting Models

The area of forecasting in the renewable energies sector is a well researched area with over 1000+ publications[3]. As it is impossible to give insights to every work which was carried out in detail, this section aims to give an overview of forecasting algorithms in the area. A number of surveys detail the problem domain and algorithms for power forecasting. In [66], forecasting methods are categorized in physical models, statistical and learning approaches, and meteorological ensemble methods. The authors furthermore give insight to common evaluation metrics and available weather models. The authors of [221] introduce a categorization by the forecasting horizon and by the nature of the forecasting model. The main models are identified to be physical and statistical approaches, other mentioned categories include the persistence method and hybrid structures. This categorization is also utilized in a more recent literature review [37]. Some surveys furthermore focus on wind power forecasting methods [42, 57, 145, 146, 241, 260]. These surveys group the existing methods in categories that are similar to the ones mentioned above. An overview of particular weather situations, such as wind ramps, is given in [74]. The authors categorize existing approaches for ramps in deterministic and probabilistic techniques.

Particular attention is drawn to solar radiation forecasting using machine learning methods in [237]. The authors categorize the methods in data pre-processing techniques, as well as supervised, unsupervised, and ensemble techniques. Another recent review of photovoltaic power forecasting highlights the economics of power forecasting [3]. The survey also highlights some of the most important forms of power forecasting, which again are categorized in physical and statistical methods. Therein, statistical methods are subcategorized in regressive approaches and techniques from machine learning.

An overview of combination techniques for short-term power forecasting is given in [224]. The authors categorize combination techniques as weighting based approaches (which are very similar to ensemble techniques), data pre-processing techniques, parameter selection techniques, and post-processing techniques. This categorization is very similar to the diversity principles of ensembles, see Section 2.6.2. More details on the area of power forecasting using ensemble techniques is given in Section 2.6.3.

Some articles deal with forecasting for particular forecasting horizons, e.g., (very) short-term [43, 89, 186], mid-term [164], and long-term [45, 117]. In the following, we want to detail the state of the art in power forecasting approaches by their methodical origin in more detail.

[3] From: Tao's Recommended Reading List for Energy Forecasters, September 2013. URL: `http://blog.drhongtao.com/2013/09/reading-list.html`. Last accessed 2017-06-08.

2.5.1 Physical Models

Physical power forecasting models estimate the power generation by estimating the physical behavior of a RE power plant. Essentially, they estimate the photovoltaic or wind turbine power curve (which, in turn is based on the formula of Eq. 2.3) based on an NWP forecast. However, other factors, such as the topography, elevation, and wind shadow effects, also play an important role. As physical models can easily introduce systematic errors in the prediction, the forecasts are frequently corrected using a statistical calibration method such as model output statistics (MOS) [90].

A comprehensive overview of the area of short-term forecasting using physical models is given in [89]. Therein, a physical model which considers topography and elevation is described. The role of weather simulation for power forecasting is investigated in [144]. It is concluded that detailed simulations which require high computational effort are needed for accurate power forecasts. The authors of [143] proposed a model that considers the roughness of the ground and includes models for obstacles. A combined system of a physical model with MOS is presented in [173] that considers elevation with digital topography maps. Orography effects are considered in a day-ahead model presented in [60]. A comparative study of the performance of physical models with neural networks and the persistence method is performed in [142]. It is concluded that on certain sites with high mean winds, physical models can be very capable on shorter time horizons. However, the persistence also exposes high quality on those time horizons up to 6h. In [113], eight case studies highlight the performance of physical wind power forecasting techniques with focus on aspects such as transmission planning and balancing cost.

While physical models have the appeal of not requiring any historic data nor power measurements during operation, the overall method requires very specific knowledge of the physical model, the turbine types, and the site location. For a flexible deployment, physical models therefore are not recommendable. Their main area of employment therefore often remains the power forecast of newly constructed RE power plants without any historic data. Furthermore, they may be needed in the future due to regulatory reasons where the use of "black box" models is not permitted (such as it is also required, e.g., for systems in aviation [41]). Most of the research nowadays therefore focuses more on statistical models, an overview of which is given in the following section.

2.5.2 Conventional Regressive Models

In [3], statistical models are categorized in regressive models and models from machine learning, the latter of which often are referred to as artificial intelligence approaches. Regressive models predict the future state of a time series (the predictand) based on the state of a number of input predictors. A categorization of regressive models can be performed using the categories of stationary and non-stationary methods, and non-exogenous and exogenous models. Most of the standard regressive approaches are detailed in the standard reference of [18]. A categorization of regressive models is given in Table 2.3.

Stationary models assume constant long-term statistical properties of the time series to be forecasted. The most simple *stationary linear models* is the moving average (MA) model, which can be used when no trend in the data is visible. More extended versions use multiple moving averages with different lengths to model a linear trend in the data, which is used for forecasting RE power plants in [148]. A well known model is the ARMA model, which is used in a case study for PV forecasting in [39]. ARMA extends the properties of the MA model with an error term that incorporates past values. AR, MA, and ARMA only use the past observations

Table 2.3: Overview of the most common regressive power forecasting models.

	Non-Exogenous	**Exogenous**
Stationary	MA [148]	ARX [5]
	AR [11]	ARMAX [148]
	ARMA [39]	NARX [32]
Non-Stationary	ARIMA [184]	ARIMAX [40]
		NARIMAX [213]

of the predictand (the power generation) to forecast the future state of the time series. The inclusion of independent predictor variables (often called exogenous variables in the context of regressive models) to an AR model leads to the autoregressive eXogenous model (ARX). For the ARMA model, the analogous model is the ARMAX model. For power forecasting, these exogenous variables typically are NWP data. Two case studies for power forecasting using an ARX model are shown in [5], an ARMAX model is used in [148]. An adaption of the AR model to probabilistic power forecasts is presented in [11]. *Non-stationary models* model a long-term trend in the data. A popular approach is the autoregressive integrated moving average (ARIMA) model. A study compared the performance of the ARIMA model to other forecasting approaches [184]. Some authors extended non-stationary models with seasonal components [17]. A case study for the long-term electricity load forecasting using an ARIMAX model can be found in [40]. Other approaches include *nonlinear models*, e.g., nonlinear AR models (NARX), the nonlinear ARMAX model (NARMAX), and the NARIMAX model. As is to be expected, the models which do incorporate exogenous models typically outperform models which do not, as has been investigated by [5, 148]. However, the case studies also show that models from machine learning typically outperform regressive models in many cases. In general, regressive models and in particular models which do not use exogenous variables are relatively simple in their structure and therefore are mainly used for (very) short-term forecasting.

2.5.3 Machine Learning Models

Forecasting models from machine learning often are described as “black box” models which learn the relationship of input variables (predictors) to a target predictand from historic data. In power forecasting, machine learning models are also frequently referred to as models from artificial intelligence (e.g., in [37, 146, 221]). There are a number of popular machine learning techniques for power forecasting.

According to [3], the most widely used techniques for power forecasting are *artificial neural networks* (ANN). An ANN is composed of a set of artificial neurons, which is a mathematical function that is loosely inspired by the working principles of biological neurons. In an ANN, the artificial neurons are interconnected and typically organized in a layered structure. An overview of the area of ANN for power forecasting and a variety of neural network structures is given in [121]. The forecast quality of multi-layer perceptrons (MLP) is compared to other neural network structures in [247] using particle swarm optimization techniques, while recurrent neural networks are investigated in [7]. Time delay neural networks for power forecasting are investigated in [125]. An automated specification method for ANN using particle swarm optimization (PSO) is described in [131].

In recent years, ANN have regained additional attention in research as deep learning emerged [9]. Research in deep learning focuses on a different set of network structures, e.g.,

autoencoders, deep belief networks, and long short-term memory networks (LSTM) for tasks such as data encoding, information extraction, or time series forecasting [49]. LSTM networks are a subcategory of recurrent neural networks and use additional memory cells to be able to store states [111]. Deep learning architectures has recently been used to forecast renewable energies. In [197], the performance of LSTM networks for wind power forecasting has been demonstrated and compared to PCA based methods. Deep belief networks are used in [223] to predict wind power and [133] uses stacked autoencoders to predict short-term wind speed. A comparison of deep learning techniques for solar power forecasting has been investigated in [79]. The authors of [118] compare the performance of deep learning methods to support vector regression (SVR) and extreme learning machines (ELM) on wind farm data.

For classification tasks, *support vector* based methods, such as support vector machines (SVM), are seen as one of the most capable classifiers. The idea of SVMs is to fit a hyperplane in a feature space to separate data with different classes. Support vector regression methods (SVR) are the counterpart to SVMs for regression problems. In contrast to SVMs, the hyperplane is fitted to represent the data in the most accurate way. SVR was used for interval forecasting of the power generation in [203]. A study of photovoltaic power prediction on multiple time horizons using SVR techniques is investigated in [46].

Nearest neighbor methods are common methods in machine learning. The idea of the nearest neighbor method is to find similar data points (or time series segments) to a query data point in known data and aggregate the predictand of the found similar points. For power forecasting, these techniques are commonly referred to as analog ensembles. Examples of analog ensemble techniques for power forecasting can be found in [48, 85]. More details on analog ensembles can be found in Section 2.6.4.

While there are also other machine learning techniques, such as piecewise linear regression techniques, logistic regression, and single decision tree-based methods, their performance typically cannot keep up with more complex models, which has, e.g., been analyzed in [152].

Ensemble techniques have proven capable of successfully increasing the forecast accuracy of a number of individual models. Ensembles are also located within the realm of machine learning techniques. A survey on ensemble techniques for power forecasting can be found in [206]. A detailed review of ensemble techniques and their application for power forecasting can be found in Section 2.6.

2.5.4 Baseline and Hybrid Methods

Besides the categories of the forecasting categories mentioned above, there are other techniques which cannot be assigned to one of the above categories. A common baseline technique for very short-term forecasting is the persistence method that creates the forecasts by assuming the same power generation as the most current actual power measurement in the future. While this technique appears to be very simple, it is actually competitive on very short time horizons [221]. On the other hand, for very long forecasting horizons, the climatological forecasting model is the standard model. The climatological method takes the long-term average value as prediction. In many cases conditional climatological models, such as seasonal climatological models [183], are constructed to create a more precise forecast.

In the area of hybrid models, the authors of [201] present a model combining machine learning techniques with turbine power curve models. The machine learning techniques can thus be seen as a correction factor on the physical power model. More recent models of this type for PV forecasting can be found, e.g., in [56, 75]. In [47], a neuro-fuzzy model

is applied to very short-term horizons. A forecasting technique based on Kalman filters is described in [258]. In the article, the filtering technique is combined with a regressive ARIMA model. Several approaches focus on combining regressive and machine learning models, such as ARIMA and ANN techniques [202], SARIMA with SVM [17], ANN with NARX [236], or ARIMA, ANN, and fuzzy techniques [252], to name a few. These approaches focus mostly on the very-short-term horizon.

2.6 Ensemble Principles and Architectures

As mentioned in Section 2.4, the aggregation of predictive models in ensembles can improve the forecasting accuracy. This section details the base working principles of ensembles and their application in the power forecasting domain. Section 2.6.1 lays out the principle of the aggregation of single predictive models to an ensemble. Furthermore, the explanatory reason of why ensembles work, which is founded in bias-variance-covariance decomposition [214], is detailed. In Section 2.6.2, the state of the art in ensemble techniques is described and the main principles of ensemble construction are given. The main principles for ensemble creation are categorized by their form of diversity. The most important diversity forms are data diversity, parameter diversity, and structure diversity. Ensembles for power forecasting from multiple weather models are detailed in Section 2.6.3. Therein, ensembles prediction systems (EPS), multi-model ensembles (MME), and time-lagged ensembles (TLE) are described. Finally, analog ensembles (AE) are explained in Section 2.6.4.

2.6.1 Ensemble Fundamentals

An *ensemble* is an umbrella term that refers to the aggregation of predictions of a number of single models to an overall prediction. Ensembles are also called committees (e.g., in [13]). The basic idea of ensembles is to create an aggregated prediction that is more accurate than the single predictions which form the ensemble in the first place. The most central formula for the aggregation of single model predictions to an overall prediction $\bar{f}(\mathbf{x})$ with explanatory variable $\mathbf{x} \in \mathbb{R}^D$ which most ensemble techniques share can be denoted as

$$\bar{f}(\mathbf{x}) = \sum_{j=1}^{J} w^{(j)} \cdot f_j(\mathbf{x}), \tag{2.8}$$

where J is the number of ensemble members (or *base predictors*) indexed by j and $f_j(\mathbf{x})$ is the prediction of ensemble member j with weight $w^{(j)} \in [0,1]$, which should comply to

$$\sum_{j=1}^{J} w^{(j)} = 1 \tag{2.9}$$

in order to not introduce biased predictions when having unbiased base predictors. There are two basic approaches of setting weights:

- *Cooperation / Weighting*: One possibility to create an ensemble forecast is by letting the single ensemble members cooperate in creating the final point estimate. In the easiest case, the weights $w^{(j)} \in [0,1]$ can be chosen equally, i.e., $w^{(j)} = \frac{1}{J}$. Other possibilities are, e.g., to set them proportional to their overall average forecasting quality, if known. The weight values typically are static in this technique, they do not change after being set.

- *Competition / Gating*: In this approach, in each situation one model succeeds in competing against the other models, i.e., $w^{(j)} \in \{0,1\}$ and $\sum_{j=1}^{J} w^{(j)} = 1$. The challenge of this approach consequently is in deciding which power forecasting model should win the competition for a particular forecast. The weight values are dynamic in this technique, they vary depending on some defined criterion.

Combinations of these two basic approaches are part of ongoing research and will be highlighted in the following sections.

Two prerequisites which are required for ensembles to work is that the single ensemble members are *skillful*, and *diverse*, as laid out in [54]. Skillful in this case means that the performance of each ensemble member has to be better than random guessing. Diverse, on the other hand, means that the errors of each model are uncorrelated in the ideal case. This means that when having a set of predictions, if one model makes an error, there may be a chance that the other models do not make this error in the particular case, and the majority vote may still result in issuing the correct prediction. This effect is related to *bias-variance decomposition*, which is the main rationale of why ensembles work. Bias-variance decomposition is explained in detail, e.g., in [13, 138]. On a coarse level, this decomposition describes the attribution of expected error (loss) to

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}. \tag{2.10}$$

Assume a set z of predictors $\mathbf{x}$ and target observations o with $z = \{(\mathbf{x}_1, o_1), \ldots, \mathbf{x}_N, o_N)\}$ drawn from the "true" distribution $p(\mathbf{x}, o)$ (typically unknown in practice) with zero noise for the sake of simplicity. The set z is used to create the function $f(\mathbf{x})$ (abbreviated as f where unambiguous) to predict o from the predictors $\mathbf{x}$. The model should not fit z with zero error to avoid overfitting as z is only a sample of $p(\mathbf{x}, o)$. The least-squares loss function can then be written as

$$\mathbb{E}\left[(f-o)^2\right] = \underbrace{\left(\mathbb{E}[f]-o\right)^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}\left[(f-\mathbb{E}[f])^2\right]}_{\text{variance}} \tag{2.11}$$

with expectation $\mathbb{E}$. This formula quantifies the relationship of bias and variance for a *single* model and describes the generalization characteristics of the model. Bias and variance have to be balanced in order to achieve optimal model accuracy.

Assuming a set of single models f_j is each created by independently drawing from $p(\mathbf{x}, o)$, the ensemble model $\bar{f}(\mathbf{x})$ (abbreviated as $\bar{f}$) can be computed using Eq. 2.8 with $w_j = \frac{1}{J}$. For the combination of models an extended form of decomposition into *bias-variance-covariance* is described, e.g., in [25, 26, 194, 231], which is denoted by

$$\mathbb{E}\left[\left(\bar{f}-o\right)^2\right] = \overline{\text{bias}}^2 + \frac{1}{J}\,\overline{\text{variance}} + \left(1-\frac{1}{J}\right)\overline{\text{covariance}}, \tag{2.12}$$

where $\overline{\text{bias}}$, $\overline{\text{variance}}$, and $\overline{\text{covariance}}$ indicate the averaged variants of bias, variance, and

covariance over all ensemble members in the form

$$\overline{\text{bias}} = \frac{1}{J}\sum_{j=1}^{J}\big(\mathbb{E}[f_j]-o\big), \tag{2.13}$$

$$\overline{\text{variance}} = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\big[\big(f_j-\mathbb{E}[f_j]\big)^2\big], \tag{2.14}$$

$$\overline{\text{covariance}} = \frac{1}{J(J-1)}\sum_{j=1}^{J}\sum_{i\neq j}\mathbb{E}\big[(f_j-\mathbb{E}[f_j])(f_i-\mathbb{E}[f_i])\big]. \tag{2.15}$$

As can be seen from the equation, the overall error is composed of three components. A bias component measures the average deviation of the ensemble members to the observations, a variance component measures the average variation to the expected value of each ensemble member, and a covariance component measures the pairwise variation between the ensemble members. As laid out in [25, 207], the goal ideally is to decrease the covariance term (which can be smaller than 0) while not increasing the (positive-valued) bias and variance, to decrease the overall squared error. As can also be observed is that when increasing the number of ensemble members J, the importance of covariance in relation to the variance term increases. In a nutshell, the formula indicates that using a number of low correlated models can improve the overall prediction accuracy.

In the perfect case of unbiased and completely uncorrelated errors, i.e.,

$$\mathbb{E}\big[(f_j(\mathbf{x})-o)\big] = 0, \tag{2.16}$$

$$\mathbb{E}\big[(f_j(\mathbf{x})-o)(f_i(\mathbf{x})-o)\big] = 0,\ \ i\neq j, \tag{2.17}$$

the squared error of the ensemble $E_{\text{ens.}}$ can then be reduced to

$$E_{\text{ens.}} = \frac{1}{J}E_{\text{avg.}}, \text{ with} \tag{2.18}$$

$$E_{\text{ens.}} = \mathbb{E}\big[(\bar{f}(\mathbf{x})-o)^2\big], \tag{2.19}$$

$$E_{\text{avg.}} = \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\big[(f_j(\mathbf{x})-o)^2\big], \tag{2.20}$$

if the errors of the models are uncorrelated (the proof for the reduction of the error is based on ambiguity decomposition, which is described, e.g., in [25, 26]). However, this rarely is the case in practice, as errors typically are highly correlated. This puts a limit on the reduction of the prediction error, which would otherwise be reduced steadily when incorporating more models in the prediction. However, in many cases one can assume that there is at least some amount of independence of errors between the models. As described in [13] it can be shown that the squared error of an ensemble does not exceed the sum of average errors of the ensemble members in the form

$$E_{\text{ens.}} \leq E_{\text{avg.}}. \tag{2.21}$$

The diversity principle still has to be fulfilled, which is why high variance – low bias models typically yield the best results in practice.

2.6.2 Basic Ensemble Techniques and Construction Principles

A number of reasons for the success of ensemble techniques are laid out in [54]. The authors argue that ensembles help to overcome some of the "practical" problems of creating a predictive model, which can be categorized in *statistical*, *computational*, and *representational* reasons. Statistical reasons are seen to occur when the training data set is too small to model the observed process. By averaging models that are trained on subsets of the available data, the average hypothesis of the models is assumed to yield more accurate predictions. Computational reasons include the effects of terminating model training in local optima (e.g., when using stochastic gradient descent in neural networks), which can better be overcome by repeated training of models. Representational reasons involve the lack of model flexibility that may render the model incapable of modeling the complex structure of the problem. Through combination, a higher combined flexibility is achieved (e.g., if different models are used for different areas of the feature space). This effect is also tightly coupled to the divide and conquer principle [72]. Strength correlation [21], stochastic discrimination [135], and margin theory [216] are also argued to be reasons why ensembles work. They have been shown to be equivalent to a bias-variance-covariance decomposition in [194].

Ensemble classifiers and in particular variants of the AdaBoost algorithm are investigated in [210]. Therein, ensemble methods are constructed by a scheme of building blocks consisting of the data set manipulators, base inducers, diversity generators, and combiners. A similar taxonomy of building blocks for ensembles is also proposed in [209]. Based on the outlined vocabulary of the above articles, the authors of [102] categorize ensemble methods that are built upon base techniques which are identified as bagging, random subspaces, random forests, and rotation forests.

A survey on ensembles for regression can be found in [162]. The authors categorize the ensemble generation process into the three phases of generation, pruning, and integration. Therein, ensemble techniques are categorized in constant and non-constant weighting functions (cf. weighting and gating, see Section 2.6.1). A survey with focus on ensemble variants of particular machine learning models is shown in [261]. Therein, an overview of ensemble generation from multiple neural networks is given. In particular, variants of boosting, bagging, and cross-validation ensembles are described.

Some ensemble surveys highlight specific application areas, such as bioinformatics, which is investigated in [254]. The survey particularly focuses on the application of ensemble techniques to high-dimensional data of microarrays to measure gene expressions and focuses in particular on random forests. In a recent survey [207], a joint overview of regression and classification ensembles is given. The authors categorize ensembles based on their form of diversity, which directly relates to the main concept of ensemble member diversity that is required for ensembles to work as laid out in Section 2.6.1. The main diversity principles are identified to be data diversity, parameter diversity, and structure diversity.

In [206], ensemble methods for forecasting renewable energies are investigated. As this article is highly related to some of the contents of this thesis, it is described in more detail. The authors highlight the concepts of cooperative and competitive ensembles. Competitive ensembles are seen to be constructed from slightly different initial conditions or different parameters. To create the ensemble forecast, the base models are aggregated by averaging after a pruning step of weak models. The data and parameter diversity principle are seen to be used for the construction of competitive ensembles. Cooperative ensembles are described to divide the overall prediction task into smaller sub-tasks which are then solved individually. Cooperative ensembles are categorized in pre- and post-processing techniques. For preprocessing techniques, the data set is divided into subsets, each of which is predicted by

an individual forecasting model. The final prediction is then created by summarizing the outputs of the models. Post-processing techniques are seen as multi-step forecasts, where the forecast is created by consecutively applying a second technique that models the residual of the first forecast. However, the role of ensembles of weather forecasts (detailed in this thesis in Section 2.6.3) remains largely disregarded in [206].

As mentioned in [162], the number of approaches for the creation of regression ensembles is too large to sum them up in detail. Therefore, the main principles for the creation of ensembles using the diversity principles is highlighted in the following. Alongside the description of the diversities, examples of the working principles of the diversity are given using a set of figures (Fig. 2.12 – Fig. 2.14). For the sake of easier understanding, these diversities are given in the context of a classification problem.

Data Diversity

The most widely used principle to create diversity in an ensemble is to introduce a variation of the underlying training data. This form of variation is called *data diversity*. The main principle of data diversity approaches is to use a different data set for the model training of each ensemble member. This diversity is achieved using sampling strategies or subspace methods. Probably the most popular sampling method is bagging [19] which is short for bootstrap aggregation. The principle of bagging is to create subsets of the data set through bootstrapping, a predictive model thus is trained on each subset of data. An example of the working principle of bagging is shown in Fig. 2.12, which shows a number of classifiers, in this case a linear support vector machine (SVM), which were trained from a different subsampled data set (constructed using bootstrapping) indicated by the colors. As can be seen, while the classifiers roughly show the same behavior, diversity is introduced.

A second popular method is boosting [215], which iteratively boosts the importance of misclassified samples of a trained predictive model to create a better classification in the following step. Regression variants of the method which originally was introduced for classification do exist, e.g., least-squares boosting [70]. The principle of subspace methods is to create subsets of available features to independently train a base predictor. In [110], the concept of subspace methods for decision trees is introduced. The principles of sampling and subspace methods are combined in random forests [21]. Random forests are in many cases considered to be the best "off the shelf" general purpose machine learning models (e.g., in [64]), which, however, is part of an ongoing discussion [238]. A popular alternative approach is the mixture of experts model (e.g., [4, 123]), which uses a gating function to train single base predictors on particular regions of the subspace, e.g., using expectation maximization. Other more rare forms of ensembles modify the outputs of the ensemble rather than the inputs by introducing artificial noise in the ensemble. Two approaches introduced, e.g., in [20], are output flipping and output smearing.

Parameter Diversity

An alternative form of diversity creation is using parameter diversity techniques. The basic idea is that the hyperparameters of a predictive model are varied to create a number of different base predictors. Varying the parameters allows a model to work with different degrees of fit which introduce diversity. Figure 2.13 visualizes the ensemble construction principle through base predictors which expose parameter diversity. In the example, a number of SVMs with polynomial kernels are trained on the same data set. A parameter (the polynomial degree in the presented case) is varied.

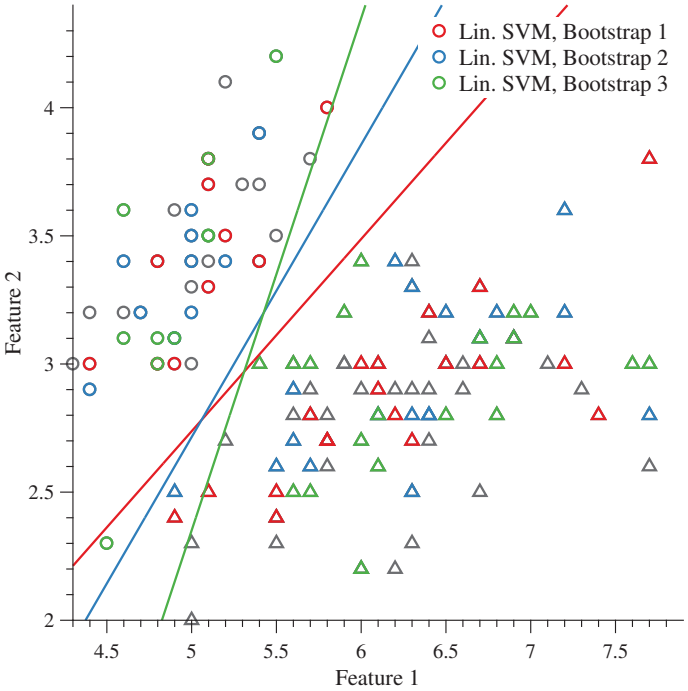


Figure 2.12: Visualization of the data diversity principle using bootstrapping.

A popular parameter diversity approach is multiple kernel learning (MKL). An overview of the area and comparison of MKL approaches is given, e.g., in [101]. The principle is that kernel functions with different parameters are used to refine the overall prediction. In the article, the authors expressed that a nonlinear combination of linear kernels or the linear combination of more complex kernels (e.g., Gaussian kernels) yields the most improvements regarding the predictive quality. MKL approaches can be categorized into cooperative and competitive approaches. Other categorizations can be made by the way of weighting (unweighted, linear, nonlinear), and the way of model learning (e.g., heuristic, optimization, Bayesian). Other reviews on MKL algorithms focus on applications such as object recognition [27], or multi-label learning [257]. Some authors refer to parameter diversity approaches as *intermediary combination* of models [178] (as the diversity is introduced in an intermediary step in contrast to the early combination through concatenation of features or late combination of only the classifier outputs). In the meteorological domain, ensemble prediction systems (EPS), see Section 2.2, can be seen as parameter diversity ensembles, as the initial conditions (measurements) and perturbation parameters are varied in order to introduce a model spread.

Structure Diversity / Heterogeneous Ensembles

Heterogeneous ensembles consist of different types of predictive models which serve as base predictors for the ensemble. This form of combination contains diversity regarding the

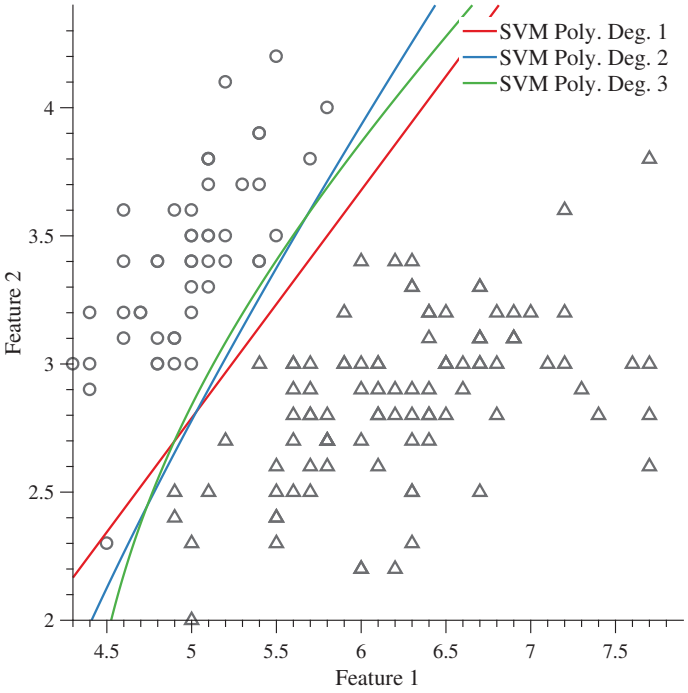


Figure 2.13: Visualization of the parameter diversity principle.

structure of the ensemble members. Figure 2.14 shows the working principle of a heterogeneous ensemble. In the example, the base predictors are an SVM with polynomial kernel, a decision tree, and a nearest neighbor algorithm. As can be seen from the figure, the decision boundaries of the base predictors greatly vary.

An overview of the area of heterogeneous ensembles for regression is given in [162]. Parameter and structure diversity ensembles are combined in [211] by including heterogeneous base predictors with multiple parameter combinations. In [242], ensemble learning using heterogeneous ensembles with multiple strategies is investigated. A pruning method for the selection of base predictors in heterogeneous ensembles is proposed in [34]. With respect to meteorological sciences, multi-model ensembles (MME), as detailed in Section 2.6.3, can be interpreted as structure diversity ensembles, as they may use a different algorithmic structure, or different number of considered meteorological variables during power forecasting model computation. Structure diversity ensembles sometimes have a fluent transition to parameter diversity approaches. For instance, the variation of a neural network regarding the size and number of hidden layers technically is a hyperparameter of the neural network model, but enables a whole different class of possible computations in the sense of a different structure of the predictive model.

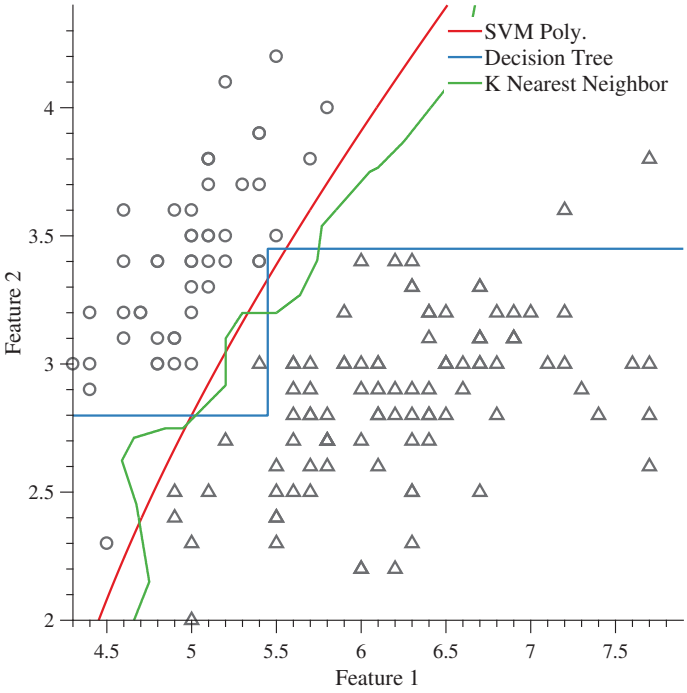


Figure 2.14: Visualization of the structure diversity principle.

Other Diversity Types

In [207], a number of other diversity types are laid out, which, however, can be seen as secondary to the main principles laid out above. The divide and conquer principle is seen to be a source of diversity [205]. Multi-objective optimization [36] may induce multiple forms of diversity among multiple base predictors along the Pareto optimal front. As pointed out in [207], diversity may be created using fuzzy ensembles [122].

2.6.3 Ensembles for Power Forecasting

In principle, any of the ensemble principles mentioned in the previous sections can be applied to the area of power forecasting. However, as the main explanatory variables for power forecasting algorithms are numerical weather predictions (NWP) on many time horizons, ensembles of meteorological weather forecasts form an own category of ensembles which are specific to the area of meteorology and are thus relevant for power forecasting. This section describes how ensembles from meteorological sciences can be applied to create a refined power forecast. The main types of ensembles from meteorology are

- ensemble prediction systems (EPS),
- multi-model ensembles (MME), and

- time-lagged ensembles (TLE).

Table 2.4 summarizes possible ensemble types by the number of weather and power forecasting models involved. Conventional ensembles from machine learning (as described in the previous section) are abbreviated as *power forecasting model ensembles* (PFE), as the single predictive models create a power forecast in this case.

Table 2.4: Categorization of ensemble types by number of weather and power forecasting models and number of forecasting origins.

Ensemble Model Type	Weather Forecasting Models	Power Forecasting Models	Forecasting Origins
No Ens.	1	1	1
EPS	>1	1	1
MME	>1	1	1
PFE	1	>1	1
TLE	1	1	>1

The following sections describe the different meteorological ensemble principles in more detail. Another ensemble principle from meteorological sciences are analog ensembles, which, however, do not fall into the same methodical category as the ensembles mentioned above, they are detailed in Section 2.6.4.

Ensemble Prediction System (EPS)

Ensemble Prediction Systems (EPS), sometimes also called *single-model ensembles*, are created using a systematic variation of the initial conditions (weather measurements) and optionally of the weather forecasting model generating processes, yielding different NWP. The goal of such an EPS is to assess the possible weather outcomes by including an explicit model spread which reflects the stochastic nature of the forecasting task. More details on the creation of EPS are given in Section 2.2. It thereby in principle is a data diversity ensemble, which is, however, created through varying the parameters of the NWP generating process in the sense of a parameter diversity ensemble. These forms of ensemble forecasts are typically conducted by a weather forecasting model provider. When using an EPS for power forecasting, each of the weather forecasts is used for the power generation forecast using a power forecasting model. In case an ensemble prediction system (EPS) is used, a single power forecast is created with

$$\hat{y}^{(j)}_{t+k|t} = f(\mathbf{x}^{(j)}_{t+k|t}|\boldsymbol{\theta}). \tag{2.22}$$

For this type of ensemble, the input NWP values are changing, however, the type and parametrization of the power forecasting model often remains the same. As for many EPS forecasts all NWP ensemble members are assumed to have an equal probability of being correct, the values of $w^{(j)}$ are often chosen to be equal, i.e., $w^{(j)} = \frac{1}{J}$, if a deterministic forecast is desired. An EPS is employed in [227] to forecast electrical load using neural networks with multiple possible outcomes for the weather parameters. The study is performed for a number of lead times up to ten days. In [189], the ensemble forecast is used to predict the forecasting skill by evaluating the spread of the EPS. The EPS is also compared to lagged ensembles (explanation see below). The authors of [2] conduct a comparative study between two ensemble prediction systems regarding their forecasting accuracy for wind power. In [230], an EPS is used to create probabilistic forecasts to investigate extreme weather situations and ramp events.

Multi-Model Ensembles

Multi-Model ensembles (MME), sometimes also called *poor-man's ensembles*, refer to the combination of (typically deterministic) point forecasts of different weather forecasting model providers. In principle, each of the single forecasts is created independently as detailed in Section 2.2 and Fig. 2.3. MMEs have characteristics different from EPS, as each ensemble member yields the most likely point forecast and does not try to explicitly include a model spread. Furthermore, multi-model ensemble members can differ in their representation (e.g., different physical units or different number of NWP parameters, etc.). MMEs have characteristics from structure diversity (as the weather forecasts are created using different systems from different weather model providers), but yield a set of forecasts in the sense of data diversity ensembles. For MMEs, a single power forecast is created using

$$\hat{y}^{(j)}_{t+k|t} = f_j(\mathbf{x}^{(j)}_{t+k|t}|\boldsymbol{\theta}^{(j)}). \tag{2.23}$$

For MMEs, the NWPs $\mathbf{x}^{(j)}_{t+k|t}$ may be of a different form (or they may have a different number of dimensions). The power forecasting models consequently have a different structure f_j, thus, different model parameters $\boldsymbol{\theta}^{(j)}$ have to be chosen for each ensemble member. This does not necessarily have to be the case for an EPS. MME members often have a varying weather forecasting model quality which translates to different qualities of the power forecast. The corresponding weights $w^{(j)}$ therefore typically have different values which can be set according to the expected quality of the models (e.g., by testing the model on some historic data) in order to maximize the ensemble quality. The authors of [185] use MMEs for photovoltaic power forecasting, the creation of prediction intervals for uncertainty assessment is also investigated. An MME of four climate forecasting systems using coupled ocean-atmosphere models is investigated in [235] with particular focus on the creation of statistically sound forecasts. The performance of MMEs is compared to EPS forecasts in [262]. In some cases, multiple EPS forecasts are also combined into a multi-model super ensemble [243].

Time-Lagged Ensembles

*Time-lagged ensemble*s (TLE) use a repetitive forecast of the same absolute point in time computed from different forecasting origins to contribute to an ensemble. TLEs can be computed using only a single power forecasting model and a single weather forecasting model in the form of a data diversity ensemble. TLEs typically operate on the same weather model using the same power forecasting model. They use forecasts for the forecasting time $t+k$ from different forecasting origins $t-\Delta\cdot v_j$ in the form

$$\hat{y}^{(j)}_{t+k|t} = f(\mathbf{x}_{t+k|t-\Delta\cdot v_j}|\boldsymbol{\theta}), \tag{2.24}$$

such as shown in Fig. 2.15. The value of $v_j \in \mathbb{N}^+_0$ denotes the amount of lag of the j-th ensemble member. Typically, the corresponding weights $w^{(j)}$ are chosen in a form that smaller values of v_j have a higher weight (as the amount of time-lag is smaller for those forecasts, which typically correlates with an increased precision of the forecast).

The authors of [165] use lagged ensembles for the aggregation of temporal high-resolution forecasts to achieve the effect of spatial averaging. Lagged ensembles are also frequently used to assess the uncertainty of a forecast, e.g., using risk indices [188, 189].

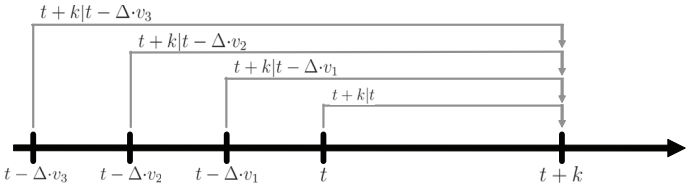


Figure 2.15: Schematic representation of the construction principle of a time-lagged ensemble.

2.6.4 Analog Ensembles

An analog ensemble is a forecasting algorithm which searches historic NWP data to create a power forecast. In a first step, similar situations ("analogs") to the predicted weather situation at time $t+k$ are searched in a historic data set. Afterwards, the corresponding historic power measurements of the analogs are aggregated to create a power forecast using Eq. 2.8. Figure 2.16 visualizes the working principle of analog ensembles.

As actual weather measurements for the desired location are rarely available, numerical weather predictions of historical situations are used. While analog ensembles aggregate multiple power measurements to an overall forecast, this technique methodically is also related to nearest neighbor techniques. The state of the art in analog ensembles is highlighted in the following.

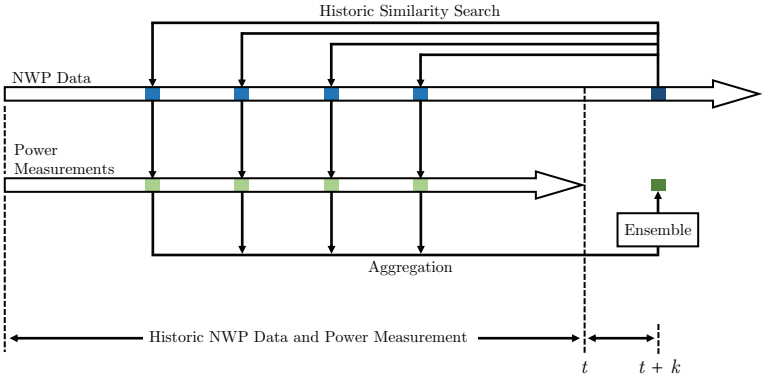


Figure 2.16: Principle of analog ensemble.

In [166], a probabilistic approach called "historical similar mining" (HISIMI) is described. One property of this technique is that probabilities for a past observation being similar to the current situation are created. The measure that defines similar weather situations in the historical data is based on a local Gaussian function. The goal in [263] is to estimate a weather situation with a finer spatial resolution given a general circulation model (GCM) in coarse resolution by statistical downscaling. Given the relationship of large-scale circulations and local weather situations as historical data, an analog method is used to find similar patterns to the current weather situation. The similarity search is performed on empirical orthogonal functions which represent the characteristics of the large-scale GCM. In [130], a model using

analog ensemble model output statistics (EMOS) is proposed to calibrate an existing ensemble forecast. EMOS needs a calibration of the coefficients on test data. To improve the calibration, a similarity search is performed to find the optimal training set.

In [48], a probabilistic approach of an analog ensemble is described. A simple Euclidean distance based similarity measure is used to find similar segments (so-called *analogs*). The corresponding observations are used to generate forecasts up to a forecast horizon of 48 h. The types of observations are weather parameters.

2.7 State of the Art in Probabilistic Forecasting

As mentioned in the previous sections, probabilistic forecasting techniques issue predictive distributions rather than point estimates to give an indication on the expected amount of uncertainty for a particular forecasting situations. As for the previous sections, this overview of the state of the art focuses on creating forecasts for continuous quantities such as the power generation (in contrast to creating probabilistic forecasts for categorical events). The representation and application has been discussed in Section 2.4.2. As a brief recapitulation, in contrast to point forecasts, probabilistic forecasts create predictive distributions $\hat{p}_{t+k|t}(y)$, e.g., in form of a parametric density function. In many cases, conventional point forecasting techniques with function $\mu(\mathbf{x})$ serve for the estimation of the conditional mean of the parametric distribution, e.g., using a normal distribution $\mathcal{N}$ with $\hat{p}_{t+k|t}(y) = \mathcal{N}(y|\mu(\mathbf{x}_{t+k|t}), \sigma)$ as a simple example.

Distributions that are continuous in the power domain are typically constructed from a (combination of several) probability distribution(s) with predefined functional form. (Mixtures of) Gaussians come to mind when thinking of predictive distributions in the continuous space. Another popular form of representations are stepwise constant (non-continuously differentiable) probability distributions in the continuous space of the predictand (e.g., the power generation for power forecasting) that assign areas of equal probability to an interval in the target value of the continuous predictand. Those distributions therefore typically are non-parametric. The most common form of representation of such a distribution is in the form of a cumulative density function (cdf) using a number of quantiles.

Probabilistic forecasts can be created using either a single predictive model as basis for the probability creation or a set of models of an ensemble. An overview of models of both types is given in Table 2.5. Other overviews on uncertainty estimation techniques for power forecasting approaches is also given in [73, 253, 259]. Whether the distribution is continuous or stepwise constant is denoted with (*C*) and (*S*), respectively, in the table.

The area of probabilistic forecasting techniques from single predictor models is further described in Section 2.7.1, the probability construction from ensembles is detailed in Section 2.7.2.

2.7.1 Probabilistic Forecasting Techniques from Single Predictor Models

Distributions from *single* predictive models can further be categorized. Parametric distributions are based on an assumption about the functional form of the probability distribution and therefore have few parameters to optimize for. While these techniques are computationally inexpensive and competitive for very short-term forecasting, the explicit assumption on the distribution shape does not always meet the distribution of the observations. Typical approaches use homoescedastic (e.g., in [18]) or heteroscedastic (e.g., [228]) uncertainty

Table 2.5: Categorization of existing representations of forecast uncertainty by their origin from a single predictor model or an ensemble method. (*C*) and (*S*) indicate a continuous or stepwise constant (non-continuous differentiable) probability distribution, respectively. EPS is the abbreviation for *ensemble prediction system*, which is set out in Section 2.2.

		Parametric	**Non-Parametric**
Single Pred.	*(C)*	Homoscedastic [18]	Kernel Density Forecast [128]
		Heteroscedastic [228]	
	(S)	-	Analog Ensemble [48]
			Quantile Regression [106, 176]
			Prediction Intervals [134, 240]
Ensemble	*(C)*	Distribution Fitting [23, 96]	Ensemble Dressing [59, 218]
	(S)	-	EPS Ensemble [108]
			Scenario Forecasting [192]
			Skill Category Forecasting [188, 189]

distributions, which refer to a non-varying or varying uncertainty within the model, respectively. The approaches are designed to work with an arbitrary parametric density function. Other parametric approaches include homogeneous (non-)linear regression [228] and heteroscedastic autoregressive models [14]. An overview of parametric distributions is given in [259]. Non-parametric uncertainty distributions do not make any assumptions of the form of the probability distribution. Therefore, they require more observation data in order to accurately model the probability distribution, especially when creating the model in a high-dimensional space of the predictor. However, they often outperform parametric distributions on longer time horizons. Distributions can be created based on historically similar weather situations. Popular techniques which use this principle are kernel density estimation (KDE) techniques [128, 196]. In KDE, a joint distribution of NWP forecast and past measurements of the predictand is constructed. For creating a forecast, a conditional distribution is constructed using the information of the NWP forecast. A multivariate variant of KDE is described in [256]. Analog ensembles [48] are a non-parametric technique that creates the probabilistic forecast non-parametrically by using a nearest neighbor search of historic measurements. Properties of analog ensembles are discussed in [12]. A third possibility is to train a power forecasting model directly to forecast *the location* of a certain cumulative density function (cdf) value within the predictive distribution. Using multiple models which each aim to create a forecast for different cdf values (the point forecast is systematically under- or overestimated), an approximation of the overall uncertainty from the single forecasts is possible. Popular techniques create the forecasting model through optimization of a modified error function during model training, e.g., quantile regression (QR) or prediction intervals (PI). The concept of quantile regression is introduced in [137]. Variants of quantile regression are, e.g., described in [176] for nonlinear QR, or in [106] for the combination of QR with fuzzy logic networks. Prediction intervals are described, e.g., in [134, 240]. PI are investigated with autoregressive models in [103] and using machine learning models [240]. In particular, PI are used with artificial neural networks [198] or extreme learning machines [239].

2.7.2 Probabilistic Forecasts from Ensemble Techniques

Furthermore, predictive distributions can be created from *ensembles*. Probabilistic forecasts using ensemble techniques are almost exclusively created as a post-processing step of a

number of deterministic forecasts that form the ensemble forecast. Therein, the origin of the ensembles typically are meteorological ensembles, i.e., EPS, MME, or TLE. Predictive distributions can be directly created using sampling from ensemble prediction systems (EPS) ensembles [108] or scenario forecasts which are the tempo-spatially consistent generalization of EPS forecasts [192]. These approaches assume equal probability of each ensemble member and the distribution is created of a set of Heaviside step functions, requiring ensembles with enough ensemble members that model the expected spread correctly. More details on the construction of probabilistic power forecasts from EPS forecasts are given in Section 5.7.1. Scenario forecasts are of use in operational decision making processes. They can be seen as samples drawn from an (unobservable) conditional forecasting probability density which considers temporal or spatial correlations. For a single time step, scenario forecasts behave very similar to EPS forecasts. The concept of scenario forecasts was first introduced in [192]. A case study of scenario forecasts for climate model data is given in [147].

Other approaches include distribution creation of ensemble members using multiple parametric density functions which are mainly designed for MME, such as ensemble dressing [218] and Bayesian model averaging [59]. These approaches are related to kernel density estimation in the sense that each deterministic forecast is assumed to have an attached uncertainty that is represented by a kernel function. A framework for the categorization of ensemble dressing approaches is given in [23]. Therein, the categories of standard kernel dressing, Bayesian model averaging (BMA), and affine kernel dressing are proposed. Case studies for BMA for multi-model ensembles is given in [150], hydrologic predictions using BMA are investigated in [59]. A case study for the prediction of regional climate change using ensemble dressing extended with temporal autocovariance is conducted in [218] and includes multivariate dressing functions and affine kernel dressing. Differences of ensemble dressing and KDE are discussed in [12]. The use of different kernel functions for bounded variables, such as power generation or wind speed, is explained in [259].

If assuming a parametric distribution on the ensemble forecasts, distribution fitting approaches can be performed, yielding a single parametric distribution as result [23]. A simple dressing approach without model training is shown in [249]. A variant of distribution fitting for probabilistic scoring rules is presented in [96]. Finally, uncertainty can be represented using skill category forecasting techniques using risk indices [188, 189]. These techniques estimate the expected error using an a-priori estimation of the expected risk of a situation using time-lagged ensembles or the ensemble spread.

This section summarizes the state of the art in the creation of probabilistic forecasts from single predictive models and ensemble models. A more thorough analysis of the mentioned forecasting algorithms is presented in Chapter 5.

2.8 Quality Assessment

This section gives an overview of the state of the art in quality assessment. In Section 2.8.1, an overview of deterministic error scores is given. Probabilistic scoring rules that are able to assess the performance of probabilistic forecasts are detailed in Section 2.8.2. Finally, the area of statistical forecast validation is highlighted in Section 2.8.3.

2.8.1 Deterministic Error Measures

Deterministic error measures quantify the performance of an issued point forecast to the actual observation. This section only briefly introduces literature on deterministic error

scores as a more detailed investigation is given in Section 3.

A number of surveys highlight the most important principles of deterministic error measures without specifying an application domain. Many error measures are based on Minkowsky distances, which are a generalization of metrics such as the mean absolute error (MAE) or the root-mean square error (RMSE), which are discussed, e.g., in [120]. The study further describes and compares measures derived from these basic error measures. In [225], the concept of summary statistics given errors for absolute (e.g., MAE) and relative error measures (e.g., the mean absolute percentage error, MAPE) is given. Furthermore, the evaluation of fixed and rolling time horizons is discussed. The concept of measures that account for the time-dependent change of the predictand (such as the MASE score) are discussed in [91]. The study presented in [38] gives a broad overview of existing error measures, presents a method for their theoretical evaluation, and gives insights on weaknesses of certain error scores.

Some surveys on power forecasting also include sections on forecasting errors. In [66], the most relevant basic measures and relevant measures for power forecasting, namely the bias, the standard deviation of errors (SDE), and the skill score are described. Correlation-based methods (such as the coefficient of determination R^2) are mentioned as well. The article furthermore compares reported RMSE values of different forecasting algorithms and data sets. Similar error measures are described in [155]. The authors also focus on the distribution of errors and give recommendations on the use of certain error scores. Another survey focuses on error scores for power forecasting and particularly highlights scores reported from ensemble approaches [206]. However, the articles mentioned above are partly inconsistent with each other, e.g., they refer to different error measures under the same name.

2.8.2 Probabilistic Scoring Rules

In the probabilistic domain, metrics for performance assessment are referred to under the term *scoring rules*. A scoring rule compares a probabilistic forecast with an observation. This section only gives a brief introduction to the area of probabilistic scores, more details and a detailed analysis are given in Section 6.

While deterministic forecasts have to be close to the corresponding actual measurements, probabilistic forecasts have to both be *sharp* (i.e., the conditional mean of the distribution is close to the measurement) and have to assess the uncertainty correctly (e.g., the width of a probability distribution) as described in [94]. The authors also give an overview to many relevant aspects of probabilistic forecast evaluation, such as the importance of a scoring rule being *proper* which is a property that ensures that the score is robust to cheating. This aspect is also further discussed in [93].

As has been shown in Section 2.7, probabilistic forecasts can be created in a number of ways. For instance, quantile forecasts are evaluated using the quantile score [10]. Interval forecasts are evaluated using the interval score [240]. Density functions, on the other hand, are evaluated using the continuous ranked probability score (CRPS) [108] (which is the continuous version of the binary Brier score [52]) or the ignorance score [212]. Alternative evaluation metrics emerge from visual verification such as the probability integral transform of the Talagrand diagram [33].

For some scoring rules, a *decomposition* has been proposed that enables a more direct investigation of properties such as reliability. [108] proposed a decomposition of the CRPS, the ignorance score has been decomposed in [229], a decomposition for the quantile score has been proposed in [10].

2.8.3 Statistical Forecast Validation

When comparing the performance of forecasting techniques, differences in the score values may be only due to chance and not due to a real difference in performance. Whether performance differences are *statistical significant* can be assessed using statistical hypothesis tests. A popular approach for the comparison of forecasts is the Diebold-Mariano test [53] that compares the forecasts of *two* forecasting algorithms on the *same* (single) time series. For the verification of *multiple* forecasting algorithms with *multiple* data sets (and optionally multiple repetitions), the Friedman test in conjunction with the Nemenyi post hoc test to assess the *ranked performance* of forecasting algorithms is proposed, e.g., in [109, 139]. In the following, we will give an overview of the working principles of the Friedman test and the Nemenyi test.

For the evaluation, a set of $m \in \{1,\dots,M\}$ different forecasting models over $s \in \{1,\dots,S\}$ data sets and $f \in \{1,\dots,F\}$ repetitions (or folds) for each combination is considered. The performance of the individual forecasting algorithms is ranked regarding a scoring function for a particular data set. For each data set the rankings can be created individually for each repetition f (leading to a set of rankings in the range $r \in \{1,\dots,M\}$) *or* over all folds (leading to a set of rankings in the range $r \in \{1,\dots,M \cdot F\}$) such as discussed in [139]. Here, we consider the ranking over each fold individually as this form of ranking is easier understandable. From this ranking, the mean rank $\bar{R}_m$ of each model m can be computed with

$$\bar{R}_m = \frac{1}{S \cdot F} \sum_{s=1}^{S} \sum_{f=1}^{F} R_{m,f,s}, \tag{2.25}$$

where $R_{m,f,s}$ is the rank r of model m on repetition f of data set s. If there is a set of models with same performance with ranks r to $r+c$, their common rank is computed with $r + \frac{c-1}{2}$. The Friedman test [71] is a nonparametric test that evaluates the null hypothesis

$$H_0 : \ \bar{R}_1 = \ldots = \bar{R}_M \tag{2.26}$$

with alternative hypothesis being

$$H_1 : \bar{R}_1,\dots,\bar{R}_M \text{ are not all equal.} \tag{2.27}$$

The Friedman test is computed with the Friedman test statistic FR with

$$\mathrm{FR} = \frac{12 \cdot S \cdot F}{M(M+1)} \sum_{m=1}^{M} \left(\bar{R}_m - \frac{M+1}{2} \right)^2. \tag{2.28}$$

From the Friedman test statistic FR, the so-called p value can be computed that indicates the likelihood of the null hypothesis (Eq. 2.26) being correct when witnessing the given set of observations (ranks). The p value is the probability under the null hypothesis of getting the observed or a larger value of the test statistic FR. It is calculated by computing the right-tailed area of the χ^2 distribution with $M-1$ degrees of freedom (number of models) from the observed test statistic value, i.e., with

$$p = \int_{x=\mathrm{FR}}^{+\infty} f_{\chi^2}^{(M-1)}(x), \tag{2.29}$$

where $f_{\chi^2}^{(M-1)}(x)$ is the density function of the χ^2 distribution. The null hypothesis conse-

quently is *rejected* if the p value is below an assumed significance level $\alpha \in [0,1]$. A typical choice of the significance level is $\alpha = 0.05$.

If the null hypothesis of equal performance is rejected, the Nemenyi post hoc test [175] can investigate which of the average ranks differ significantly. In order to have a significant rank difference, the difference of the average ranks of two models $\bar{R}_m$ and $\bar{R}_{m'}$ has to exceed the value of the critical distance (CD) which is computed by

$$\mathrm{CD} = q_\alpha \sqrt{\frac{M(M+1)}{6 \cdot S \cdot F}}. \tag{2.30}$$

The value if q_α is a factor computed from the studentized range distribution [132] depending on the significance level α. The studentized range distribution is a continuous probability distribution that is used when estimating the range (difference between minimum and maximum value) of a normally distributed set of samples when the population deviation is unknown. Typically, the result of the Nemenyi test is visualized in a plot that shows the ranks of the investigated models in conjunction with bars that indicate which plots do not show statistically significant performance differences. Such an example is visualized in Fig. 2.17.

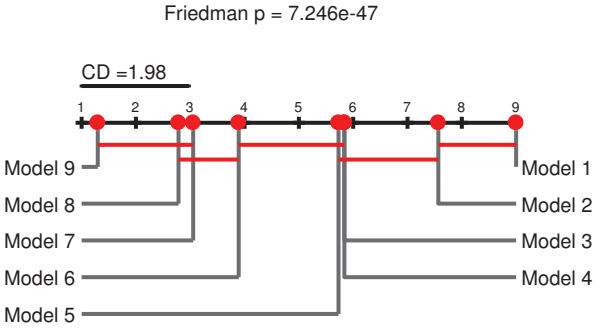


Figure 2.17: Example of Nemenyi post hoc test.

The figure furthermore indicates the result of the Friedman test on the top and the value of the critical distance. In the example, the results of a ranked performance test of 9 models is compared. As can be seen from the figure, model 9 has the lowest average rank number (thus, performs best), while model 1 has the highest average rank. As can be seen, a statistical significant difference cannot be stated between some of the models (such as model 9 and model 8). A statistically significant difference of the ranks of models cannot be proven for the models that are within the critical distance.

2.9 Application Examples

This section details two application examples that describe the use of deterministic and probabilistic forecasts for power forecasting applications. In Section 2.9.1, the process of economic dispatch and reserve capacity planning is described, while Section 2.9.2 details optimal bidding strategies for electricity traders using deterministic and probabilistic forecasts.

2.9.1 Unit Commitment, Economic Dispatch and Reserve Capacity Planning

This application example highlights how power forecasts of RE can be used for unit commitment, economic dispatch of fossil fuel combustion power plants, and the estimation of the reserve capacity for renewable energies. As the actual technical processes are very specific to the particular power grid and the regulatory framework, this example highlights the concepts on a more abstract level. As a basic principle, the power demand or electrical load (EL) has to roughly equal the produced power from fossil power plants $o^{(\text{fossil})}$ and RE power plants $o^{(\text{RE})}$ in the form

$$\text{EL}_t = o_t^{(\text{fossil})} + o_t^{(\text{RE})} \tag{2.31}$$

at each point in time t to ensure power grid stability. Load forecasting itself is a complex process and is an own area of research. A good overview on load forecasting is given in [114]. Losses in power transmission and the tolerated deviation that is uncritical for grid operation with respect to frequency protection are not included here for the sake of simplicity. The process of regulating both fossil fuel combustion and RE power plants is called *unit commitment* which describes the process of connecting and disconnecting single power plants from the power grid or regulating them if the power plant type allows for it. As can be seen from the above equation (Eq. 2.31), for an optimal control of the power output of fossil fuel combustion power plants, the determination of the expected power generation from RE power plants is necessary using a power forecast. Unit commitment is a complex process if optimizing for profit due to many constraints in the optimization process, such as fuel costs, maintenance costs, startup and shut-down costs, forced downtimes, ramp rates, and power-flow constraints in the power network [182].

The process of optimally performing unit commitment is frequently referred to as *economic dispatch* which is explained, for instance, in [157]. In a basic form economic dispatch tries to optimize the expected profit (EP) over all time steps in the way

$$\text{EP} = \sum_{n=1}^{N} \left(\hat{y}_n \cdot \hat{q}\right) \cdot U_n - \text{OC}, \tag{2.32}$$

where $\hat{y}_n$ is the predicted power generation of an individual power plant (both fossil and RE) with N being the total number of power plants in the grid. The parameter $\hat{q}$ is the expected electricity price per unit, $U_n \in \{0,1\}$ is an indicator whether the unit is committed. OC specifies the operating costs which are, besides others, composed of fuel costs, maintenance costs, and startup and shut-down costs. It therefore is a function of $\hat{y}_n$ itself. Of course, as a main constraint the optimization of EP has to fulfill

$$\text{EL}_t = \sum_{n=1}^{N} \left(\hat{y}_n \cdot U_n\right). \tag{2.33}$$

Naturally, when power from RE power plants is available, these forms of power typically are more cost-efficient (and actually have to be given priority in the feed-in process according to the EEG law [28] of the Federal Republic of Germany). In order to satisfy Eq. 2.33, a precise forecast of available energy from RE power plants is necessary for the planning process. Thereby, the goal typically is to fully exploit the available RE power while minimizing the power generation (and thus OC) of fossil fuel combustion power plants.

The *reserve capacity* is the amount of power supplied as power reserve in the grid. This

reserve is not used for electrical infeed, but is supplied as balancing power. Traditionally, this form of supplied power was used to account for failing power equipment and typically was defined at a fixed percentage of the current overall power demand, or an amount of power defined by a domain expert in regulated energy markets. Nowadays with the substantial infeed of power from RE, the reserve capacity is also used to balance for the uncertain power generation of intermittent sources of RE.

As errors in the power forecasting process do occur, the reserve capacity is used to account for fluctuations in the RE power production. When having a deterministic forecast, the amount of reserve capacity has to be determined by a domain expert or as a percentage of overall RE generation prediction. When having a probabilistic forecast, on the other hand, the amount of reserve capacity can be planned according to the amount of expected uncertainty of the power forecast. For instance, the worst case (wc) minimum power generation can be estimated by evaluating the extreme quantiles of a predictive distribution in the way

$$\hat{y}_n^{(\mathrm{wc})} = \hat{P}_n^{-1}(\tau_{\mathrm{wc}}), \tag{2.34}$$

where $\hat{P}_n^{-1}$ is the inverse function of the predictive distribution of power plant with index n, and τ_{wc} is the evaluated quantile of the data, which can be chosen to be very small for grid stability relevant operations, for instance to $\tau_{\mathrm{wc}} = 0.01$ %. The parameter therein specifies the probability $P(o_n \leq \hat{y}_n^{(\mathrm{wc})}) = \tau_{\mathrm{wc}}$ that the actual "true" power generation o_n falls below $\hat{y}_n^{(\mathrm{wc})}$ (assuming the model is correctly specified). The assumption in this case is that the actual power generation from RE will exceed the worst case power forecast in 99.99 % of all cases.

The performance of optimization methodologies for unit commitment is compared in [157]. In [15], the process of unit commitment is extended to a grid considering the effect of *demand dispatch*, which is able to modify the energy demand flexibly using probabilistic wind power forecasts.

2.9.2 Economic Bidding Strategy for Wind Power Producers

This example highlights the economic bidding strategy for electricity prices and how different types of forecast can be applied. This process takes place on the electricity trading market, e.g., the European energy exchange (EEX)[4], where, i.a., energy producers and energy providers (that provide energy to households and businesses) trade electricity. If a wind farm operator wants to sell electricity, he places a bid of expected power generation $\hat{y}_{t+k|t}$ that he wants to sell in advance (e.g., 24 h in advance) to the actual generation.

The actual produced power o_p (which is only known in hindsight) typically differs from the expected power generation. As pointed out in [22, 187, 259], the overall revenue function ξ for a particular forecasting time step $t+k$ can then be defined as

$$\xi(o_{t+k}, \hat{y}_{t+k|t}) = \begin{cases} o_{t+k} \cdot \hat{q} - (\hat{y}_{t+k|t} - o_{t+k}) \cdot \hat{c}_- & , \text{if } o_{t+k} \leq \hat{y}_{t+k|t} \\ o_{t+k} \cdot \hat{q} - (o_{t+k} - \hat{y}_{t+k|t}) \cdot \hat{c}_+ & , \text{if } o_{t+k} \geq \hat{y}_{t+k|t}, \end{cases} \tag{2.35}$$

where $\hat{q}$ is the predicted electricity price, $\hat{c}_-$ and $\hat{c}_+$ are the predicted cost for an upward or downward deviation from the bidding power $\hat{y}_{t+k|t}$. The costs $\hat{c}_-$ and $\hat{c}_+$ typically are asymmetric. In general, an electricity trader optimizes his power bid subject to

$$\hat{y}_b = \underset{\hat{y}_{t+k|t}}{\operatorname{argmax}} \ \xi(o_{t+k}, \hat{y}_{t+k|t}). \tag{2.36}$$

[4] European Energy Exchange (EEX) Official Website `https://www.eex.com/`. Last accessed on 2018-01-03.

When performing a deterministic forecast, the electricity trader conducts the bidding with the estimated most likely power generation $\hat{y}_b^{(\text{det.})} = \hat{y}_{t+k|t}$ (with $\hat{y}_{t+k|t}$ being the forecast of a deterministic forecasting model). Thus, using deterministic forecasts, the optimal monetary benefit can only be reached if $\hat{y}_{t+k|t} = o_{t+k}$, which is an unrealistic assumption due to the uncertain energy generation characteristics of RE.

However, when having a probabilistic forecast of the power generation represented as a probability density function, the trading function can be optimized in closed form depending on the expected deviation costs in the form

$$\hat{y}_b^{(\text{prob.})} = \hat{P}^{-1}\left(\frac{\hat{c}_+}{\hat{c}_+ + \hat{c}_-}\right), \tag{2.37}$$

where $\hat{P}^{-1}$ is the inverse cumulative density function of the probabilistic forecast. Therefore, the actual power bid $\hat{y}_b^{(\text{prob.})}$ is chosen depending on the costs of over- or underestimation. For instance, if the cost of upward deviation $\hat{c}_-$ is three times the cost of a downward deviation $\hat{c}_+$ (i.e., $\hat{c}_- = 3 \cdot \hat{c}_+$), then the bid is placed at

$$\hat{y}_b^{(\text{prob.})} = \hat{P}^{-1}\left(\frac{\hat{c}_+}{\hat{c}_+ + 3 \cdot \hat{c}_+}\right) = \hat{P}^{-1}(0.25). \tag{2.38}$$

Thereby, when placing a bid with $\hat{y}_b = \hat{y}_b^{(\text{prob.})}$, the expected power generation is systematically under- or overestimated depending on the expected deviation costs, leading to more cost-optimal decision making on average.

2.10 Need for Research

Based on the analysis of the state of the art in this section, the following need for additional research can be identified. In the area of deterministic error measures, there are a wide variety of existing measures, however, the individual advantages of each score for a desired application may be further clarified. Also, the suitability of different error scores for the model selection of forecasting algorithms has barely been analyzed. This affects both the areas of the discrimination ability of stronger and weaker models as well as the abstraction ability of forecasting models regarding different data sets.

For probabilistic scoring rules, many scores are tailored to a specific form of representation of the forecasting uncertainty, which hinders the comparability of forecasting techniques with different representations. A common representation of probabilistic forecasts may greatly improve the comparability of forecasting algorithms.

Furthermore, the understanding of the working principles of probabilistic scoring rules is not as commonplace as it is for deterministic error scores. In particular the decomposition of scoring rules is subject of ongoing research. While it is assumed that the decomposition of scoring rules resolves the same properties among scores, a more detailed analysis may exhibit different characteristics between the score decompositions. Also the application of a scoring rule for a particular task such as forecasting model selection is often not known and may further be clarified.

Nowadays, deterministic forecasting models are relatively mature. Besides the application of deep neural networks, the most promising way of further improving the quality may be through model combination in the form of an ensemble. While there are a number of forecasting ensembles, the weighting principles of ensembles in many cases do not

exploit the full potential of the information of the base predictors. In particular regarding the (weather-)situation or lead time-dependent weighting there may still be the possibility of further performance improvement. In order to exploit the advantages of each individual model in a particular situation, a robust method for weighting of the individual models has to be used.

The eligibility of these weighting principles within an ensemble may also be analyzed for the combination of probabilistic forecasts, where the aggregation of different forecasting algorithms may improve the overall forecasting quality. Ideally, the weighting principles can be applied independently from the type of forecast, so that the weighted ensemble technique can be used only with minimal modification.

2.11 Summary of this Section

This chapter described the theoretical and methodical foundations that are the basis for the research that is conducted in this thesis. In a first step, the power forecasting nomenclature and the overall forecasting process is described, followed by a description of the numerical weather prediction process and the basic working principles of ensemble prediction systems. Based on an overview of the characteristics of wind turbines, important aspects of the area of wind power forecasting are given. Therein, the power forecasting process in the electricity market is described, followed by a distinction of the types of forecasting that can be either deterministic point forecasting or probabilistic distribution forecasting.

Afterwards, the state of the art in deterministic power forecasting models is described, which includes the areas of physical forecasting models, regressive models and forecasting models from machine learning. The state of the art section also comprised an analysis of ensemble techniques. Therein, the basic working principles of ensembles as well as the types of diversity which are required in order for an ensemble to work are shown. The main diversity types therein can be categorized in data diversity, parameter diversity, and structure diversity ensembles. In a consecutive step, ensembles for power forecasting are investigated. These ensembles typically are constructed from a set of weather forecasts and originate from ensemble prediction systems, multi-model ensembles, or using a repeated forecast of the time period in a time-lagged ensemble. Finally, the state of the art in analog ensembles is highlighted. Based on both the state of the art in deterministic forecasting and ensemble techniques, the current state of research in probabilistic forecasting techniques is described. Therein, the construction of probabilistic forecasts from single or multiple predictor models (ensembles) is detailed. Furthermore, an overview of the state of the art of quality assessment techniques for both deterministic and probabilistic forecasts is given, the section also shows a method for statistical forecast validation. This chapter also describes two application examples that explain how both deterministic and probabilistic forecasts can be used in practice. Based on the analysis of the state of the art, this section closes with an analysis of the need for research.

Chapter 3

Metrics for Model Comparison of Deterministic Forecasts

One of the major goals of research in power forecasting is to develop better performing algorithms. In order to compare the performance of forecasting algorithms, there has to be a clear definition of how the procedure of quality assessment is performed. There are a multitude of error scores which are commonly utilized, each of which is of interest to a certain participant in the energy sector, and whose names partially are the same while they are calculated differently.

Some articles describe and summarize the general assessment of forecasting errors (i.e., not domain-specific), e.g., [38, 120, 225]. Some other surveys also include sections on forecasting error scores [66, 155, 206], however, they are partially inconsistent with each other (e.g., different scores are referred to under the same name) and only mention a selection of error scores.

The main contribution of this chapter is a structured overview of existing error measures in the area of deterministic error scores. A novel categorization by the type of basic error type and normalization technique is introduced which simplifies the process of choosing the appropriate error score depending on the envisioned application. In a number of case studies, the characteristics of the presented error measures are analyzed in detail. From the insights of the case studies, advantages and limits in the application of each error score are discussed.

This section is structured in the following way: Section 3.1 first discusses desired properties of deterministic error scores. Section 3.2 summarizes deterministic error scores and categorizes them by their way of normalization in Section 3.3. An overview of these composed error scores using the normalization techniques is then shown in Section 3.4. Other relevant scores that are independent from the outlined normalization techniques are mentioned in Section 3.5. Sections 3.6 and 3.7 analyze the presented error scores in a number of case studies to show their effects. Particular attention to the aspects of score discrimination and abstraction is presented in a case study that is shown in Section 3.8. The case studies are then discussed in Section 3.9. Our key findings are summarized in Section 3.10.

3.1 Desired Error Score Properties

As mentioned in the previous section, the goal of a deterministic forecasting model f is to estimate the future location of a target predictand o_{t+k} at time $t+k$ (in power forecasting, the expected power generation) from the forecasting origin t in the most precise way possible, i.e. with a point forecast $\hat{y}_{t+k|t}$.

Naturally, a suitable error score should be able to reflect this quality given a set of forecast-observation pairs. Many basic error measures (e.g., the mean absolute error, MAE) are able to give an indication of this quality in absolute terms. Therefore, an error score should be able to *discriminate* a good forecasting model from a worse performing model.

However, error scores may also be used for other tasks such as selecting the appropriate model or parameterization of a model. When a developer creates a novel forecasting model, it is likely that he or she reports the quality of the forecasting model on a data set that is easily available to him or her. However, the quality of a forecasting model naturally depends on the data set, e.g., the seasonality, type of power plant, complexity of the terrain, available explanatory variables, and forecasting time span, besides others. These information may not be publicly available or quantifiable, therefore the comparability of forecasting models is hindered when comparing reported qualities from different authors. For instance, a forecasting model that yields low error for a solar power plant in the Australian Outback (which has generally stable conditions and very little cloud formation) may still be less capable than a model that yields a higher error for a power plant near the English coastline (which has to deal with frequent rain events and clouds). An appropriate error score for model comparison therefore ideally is able to perform an *abstraction* from such influences which may be characterized as the "difficulty" of the forecasting task. This abstraction typically is performed using a normalization technique which is applied on the error scores.

3.2 Basic Error Measures

As mentioned in Section 2.1, a forecast is typically performed using a forecasting model to transform a weather forecast into a power forecast, which in a typical form can be

$$\hat{y}_{t+k|t} = f(\mathbf{x}_{t+k|t}|\boldsymbol{\theta}), \tag{3.1}$$

where $\mathbf{x}_{t+k|t}$ are the parameters of an explanatory variable (typically an NWP) at time $t+k$ with forecast origin t, and f is the forecasting model function with model parameters $\boldsymbol{\theta}$. The *forecasting error* can be calculated after creating the forecast using

$$e_{t+k|t} = \hat{y}_{t+k|t} - o_{t+k}, \tag{3.2}$$

where o_{t+k} is the observation of the power measurement at the corresponding time $t+k$. From this simple form of forecasting error, error measures can be derived.

For quality assessment, a number of single deterministic forecasting errors are aggregated into an overall score. There exist a number of scores, which can be seen from Table 3.1. Though there are more sophisticated error scores, most of those scores are based on one of these basic scores.

Each score can either be computed for each forecasting time step k separately (if an investigation of the characteristics on a per lead time basis is aspired), or as a summarized overall score over all forecasting time steps. The formula remains the same in this case, though, naturally, all relevant points for the evaluation have to be included then[1].

[1] Optionally, the investigation of normalizations for each lead time may be of interest for special applications, such as performed in [99] which is also detailed in Section 3.9.3.

Table 3.1: Overview of basic deterministic error measures.

Error Measure Name	Formula			Purpose
Bias	Bias	=	$\frac{1}{N}\sum_{n=1}^{N} e_n$	Shows if an algorithm overestimates or underestimates a forecast (on average).
Mean Absolute Error	MAE	=	$\frac{1}{N}\sum_{n=1}^{N} \lvert e_n \rvert$	Linear absolute error measure. Proportional weighting of errors.
Mean Squared Error	MSE	=	$\frac{1}{N}\sum_{n=1}^{N} e_n^2$	Quadratic error. Smaller weighting of small errors, larger weighting of large errors.
Root Mean Squared Error	RMSE	=	$\sqrt{\frac{1}{N}\sum_{n=1}^{N} e_n^2}$	Square root of MSE has the original physical unit of the forecast.

The Bias score is just an average of all single error values

$$\text{Bias} = \frac{1}{N}\sum_{n=1}^{N} e_n. \tag{3.3}$$

In itself, this measure has the property of balancing out positive and negative errors. Therefore, it only shows whether an algorithm overestimates or underestimates a forecast on average. The bias itself is not a measure of the forecasting quality of a model, though a low bias is desirable and related to a low error.

The mean absolute error (*MAE*) is computed using

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N} |e_n|. \tag{3.4}$$

The MAE score sums up the absolute error of each forecast. It, therefore, considers the individual forecasting errors e_n in a linear fashion. If the overall minimum difference between the forecast and the power measurements is to be determined, the MAE is the appropriate score.

The mean squared error (*MSE*) is calculated using

$$\text{MSE} = \frac{1}{N}\sum_{n=1}^{N} e_n^2. \tag{3.5}$$

Unlike the MAE score, this score factors in the errors quadratically. Thus, high errors are penalized more, while low errors have lower influence on the overall score. If a forecasting model has to avoid extreme errors, the MSE score is the more appropriate error measure. However, the MSE score is a squared score, the value has little relationship with the actual differences. Therefore, this score is mostly used for optimization purposes during forecasting model training. The MSE is optimal during least-squares optimization when assuming a normally distributed error which overlays the deterministic portion of the signal [13].

The root mean squared error (*RMSE*) is computed using

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} e_n^2}. \tag{3.6}$$

The RMSE has the same qualitative meaning as the MSE score. However, as the square root of the MSE value is computed, the value is represented in the original physical unit, making it easier to relate to a forecast value.

3.3 Score Normalization Techniques

The normalization of scores is a process that aims to make score values better comparable. In the area of power forecasting, there exist a multitude of types of error normalization, each of which has a certain purpose. An overview of the various normalization techniques is given in Table 3.2.

Table 3.2: Overview of commonly used normalization techniques.

#	Normalization Technique	Normalization Variable	Purpose
1	Nominal Capacity	o_{inst}	Scale-free comparison, comparability independent of nominal capacity.
2	Current Power Generation	o_{t+k}	Examination of relative error. Errors in low generation scenarios have higher impact.
3	Deviation from Average	$\lvert o_{t+k} - \bar{o}\rvert$	Lower weighting of situations at the extreme ends of the generation spectrum.
4	Dynamic characteristics	$\lvert o_{t+k} - o_{t+k-1}\rvert$	Lower weighting of situations with high variability in the power generation.

1. The simplest way of normalization is by dividing the forecast value by the *nominal capacity* of the power plant (i.e., by o_{inst}). The error consequently is computed using

$$\frac{e_{t+k|t}}{o_{\text{inst}}}. \tag{3.7}$$

   This normalization is a constant division factor for each power plant, making it easily understandable. Using this form of normalization, a scale-free comparison of the forecasting quality for different power plants is possible. The overall installed capacity of each power plant is no longer relevant.

2. Another way of normalizing is by dividing the error through the *current power generation* of the power plant o_{t+k}, i.e., by calculating

$$\frac{e_{t+k|t}}{o_{t+k}}. \tag{3.8}$$

   This form of normalization realizes a relative error in the sense of a percentage error. This type of normalization naturally weights a certain absolute error in a low power generation scenario higher than in a high power generation scenario (as the percent-wise error is larger if the denominator is smaller).

3. The error can be normalized by factoring in the deviation of a current power generation o_{t+k} from the average power generation $\bar{o}$ in the evaluated time span, i.e.,

$$\frac{e_{t+k|t}}{\lvert o_{t+k} - \bar{o}\rvert}. \tag{3.9}$$

   This form of normalization penalizes errors near the average power generation, while errors at the extreme ends of the power spectrum have less influence on the overall error score (as the denominator is smaller when the current power generation is near the average power generation).

4. The error can be normalized with respect to the dynamic characteristics of the current power generation $\Delta o_{t+k} = |o_{t+k} - o_{t+k-1}|$, the normalized error consequently is computed using

$$\frac{e_{t+k|t}}{\Delta o_{t+k}}. \tag{3.10}$$

In general, a forecasting problem is more difficult when the dynamics of the weather situation (and thus of the power generation time series) are high. This form of normalization aims to penalize errors in situations with low dynamic variability higher, while situations with high variability are weighted lower. This way of normalization lowers the impact of difficult weather situations (i.e., weather situations that are in the process of changing).

3.4 Overview of Composed Error Scores

There are a number of additional combined error scores, which are a combination of one of the primary error scores (see Section 3.2) and a normalization technique (see Section 3.3). The effect of these derived measures consequently is a combination of the primary score and the normalization technique. The derived scores are categorized with respect to their basic score and the normalization technique in Table 3.3. The calculation of the particular scores is again shown in Table 3.4. Some authors use the same measure name for a score with a different normalization technique (see Table 3.3), therefore, the calculation formula of the precise score should always be given when reporting error scores. Some of the mentioned scores even include multiple normalization techniques.

Table 3.3: Error measures depending on their basic error measure and respective normalization technique. The computation of the scores is described in Table 3.4.

#	Basic Measure	Normalization Technique			
		o_{inst}	o_{t+k}	$\lvert o_{t+k} - \bar{o}\rvert$	$\lvert o_{t+k} - o_{t+k-1}\rvert$
1	Bias	NBias[155]	-	-	-
2	MAE	NMAE[155]	MRE[206] MAPE[38, 206]	-	RAE[38] MASE[120, 206]
3	(R)MSE	NMSE[155] NRMSE[155]	-	NMSE[38, 206] mRSE[38] KL[38]	RSE[38] mRSE[38] U2[38]

3.5 Deviation Assessment, Correlation, and Model Comparison

For deterministic forecasts, it makes sense to not only determine the average error of a forecast, but also the distribution of the errors. The standard measure for deviation assessment is the *standard deviation*. However, as it is already being dealt with errors, the term *Standard Deviation of Errors* (SDE) is introduced in [66, 155], which is nevertheless identical to the classic standard deviation computation

$$\text{SDE} = \sqrt{\frac{\sum_{n=1}^{N}(e_n - \bar{e})^2}{N-1}}, \tag{3.11}$$

Table 3.4: Combined forecasting error scores grouped by normalization technique. A categorization of these techniques is shown in Table 3.3.

	Measure		Computation
o_{inst}	Normalized Bias	NBias	$\text{NBias} = \frac{1}{N}\sum_{n=1}^{N}\frac{e_n}{o_{\text{inst}}}$
	Normalized MAE	NMAE	$\text{NMAE} = \frac{\text{MAE}}{o_{\text{inst}}}$
	Normalized MSE [155]	NMSE	$\text{NMSE} = \frac{1}{N}\sum_{n=1}^{N}\frac{e_n^2}{o_{\text{inst}}}$
o_{t+k}	Mean Root Error	MRE	$\text{MRE} = \frac{1}{N}\sum_{n=1}^{N}\lvert\frac{e_n}{o_n}\rvert$
	Mean Absolute Percentage Error	MAPE	$\text{MAPE} = \text{MRE} \times 100\%$
$\lvert o_{t+k} - \bar{o}\rvert$	Normalized MSE [38, 206]	NMSE	$\text{NMSE} = \frac{1}{N}\sum_{n=1}^{N}\frac{e_n^2}{\lvert o_n - \bar{o}\rvert}$
	Modified Root Square Error	mRSE	$\text{mRSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{e_n^2}{\Delta o_n^2 + \frac{1}{N}\sum_{n'=1}^{N}(o_{n'} - \bar{o})^2}}$
	Kullback-Leibler based	KL	$\text{KL} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{e_n^2}{\frac{1}{N}\sum_{n'=1}^{N}(o_{n'} - \bar{o})^2}}$
$\lvert o_{t+k} - o_{t+k-1}\rvert$	Root Absolute Error	RAE	$\text{RAE} = \frac{\sum_{n=1}^{N}\lvert e_n\rvert}{\sum_{n=1}^{N}\lvert\Delta o_n\rvert}$
	Mean Absolute Scaled Error	MASE	$\text{MASE} = \frac{1}{\frac{N}{N_k - 1}}\frac{\sum_{n=1}^{N}\lvert e_n\rvert}{\sum_{n=2}^{N}\lvert\Delta o_n\rvert}$
	Theil's U	RSE / U2	$\text{RSE} = \sqrt{\sum_{n=1}^{N}\frac{e_n^2}{\Delta o_n^2}}$

where $\bar{e}$ is the mean of all error values.

If a more precise assessment of the distribution is desired, the computation of higher moments of the distribution is an option. In particular, the calculation of the *skewness* and/or *kurtosis* are two reasonable options. Other possibilities to assess the errors are, e.g., using error distribution histograms.

The process of comparing forecasting models is typically done using the *skill score* [66, 155] computed by

$$\text{Imp} = \frac{e_{\text{base}} - e_{\text{eval}}}{e_{\text{base}}}, \tag{3.12}$$

where e_{eval} is the error score of an evaluated forecasting technique and e_{base} is the error of a baseline technique to compare it to. In many cases, the persistence method or a climatological forecast is used as baseline technique. The result is a factor of improvement Imp, which is positive if the evaluated technique is better than the baseline technique and negative if the baseline technique outperforms the evaluated technique. The skill score therefore often is represented as a percentage value (by multiplying it with 100). It can be applied to any measure, such as MAE, (R)MSE, or even probabilistic scores. It then represents the improvement on the respective score.

Another quality assessment technique is the *coefficient of determination* R^2, which is the

squared coefficient of correlation computed by

$$R^2 = \frac{\left(\sum_{n=1}^{N}(\hat{y}_n - \bar{\hat{y}})(o_n - \bar{o})\right)^2}{\sum_{n=1}^{N}(\hat{y}_n - \bar{\hat{y}})^2 \sum_{n=1}^{N}(o_n - \bar{o})^2}, \tag{3.13}$$

where $\bar{\hat{y}} = \frac{1}{N}\sum_{n=1}^{N}\hat{y}_n$ is the average of all forecasts. This measure shows the ability of the model to explain the variance of the data, i.e., it determines the amount of correlation between the evaluated data set and the forecasting model. However, only the amount of linear correlation is assessable, which limits its usefulness. Furthermore, the R^2 score may be high, even though the model may still be incorrect regarding bias (constant error) or scale (error regarding the amount of change in the model). The score is in the range $[0,1]$. There are a number of scores related to the R^2 score, such as Pearson's R, mutual information, or the Kullback-Leibler divergence (when evaluating distributions), which have a similar meaning. In [155], an alternative definition of the R^2 score is given with

$$R^2 = \frac{\mathrm{MSE}_{\mathrm{avg}} - \mathrm{MSE}_{\mathrm{eval}}}{\mathrm{MSE}_{\mathrm{avg}}}, \tag{3.14}$$

which aims to eliminate some of the disadvantages of the R^2 score (but is not identical in its computation). In the above formula, $\mathrm{MSE}_{\mathrm{avg}}$ is the error of the climatological sample mean model (the forecast always is $\hat{y}_n = \bar{y}$). A detailed critique of the R^2 score can be found in [155].

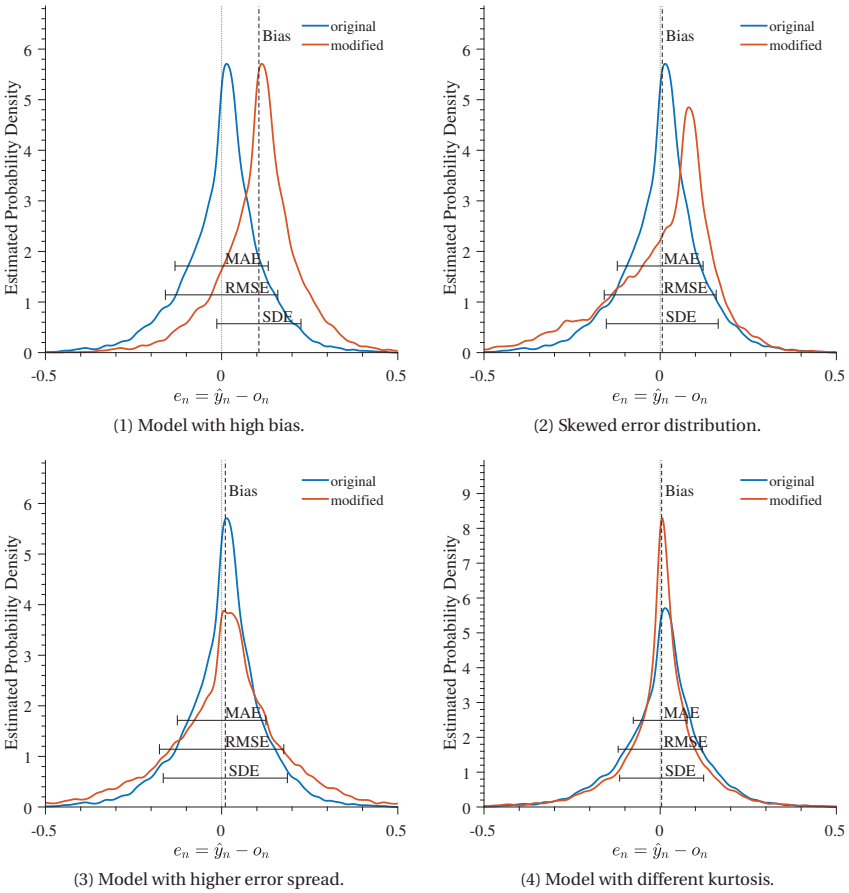
In the following, we analyze the properties of the proposed score. In Section 3.6, the behavior of error scores regarding different forms of the error distribution is analyzed. Section 3.7 analyzes the deterministic error scores regarding the properties of discrimination and abstraction as laid out in Section 3.1.

3.6 Case Study: Error Distribution Effects

In this case study, we want to investigate the behavior of different error measures given a varying form of error distribution to compute the scores on. The error distribution is taken from a real-world example from a windfarm data set (wind farm *wf4* of data set available at [76]). The error distributions are modified artificially in order to simulate possible forms of error distributions which may occur in practice. The different distributions are shown in Fig. 3.1. For the sake of better visibility, the distribution is visualized as a probability density function. We aim to include five forms of modified error distributions in the case study. Besides the unmodified original error distribution, a model with a high bias (Fig. 3.1.1), skewed error distribution (Fig. 3.1.2), higher error spread (Fig. 3.1.3), and a different kurtosis (Fig. 3.1.4) are included. The corresponding error values are displayed in Table 3.5. The table shows the percentage of change of the error values in comparison to the original (not modified) error distribution. The original absolute values and the error modification functions are given in Tables C.1 and C.2 in the appendix.

In the evaluation, we include the most popular error scores, i.e., the basis scores (see Table 3.1), and other popular and frequently used scores, such as SDE, R^2, mRSE, KL, MASE, NMSE [38], and MAPE.

All distributions except the biased distribution have no bias, as is shown by the bias score. The values of the RMSE score generally are higher than those of the MAE score due to the high weighting of elements with high distance. This effect can especially be observed if changing the kurtosis: The value of RMSE remains the same, while the MAE value decreases. Both



(1) Model with high bias.

(2) Skewed error distribution.

(3) Model with higher error spread.

(4) Model with different kurtosis.

Figure 3.1: Different forms of error distributions. The unmodified original distribution is shown by the blue line in each figure. The distribution is then modified to represent a model with a high bias (Fig. 3.1.1), a skewed error distribution (Fig. 3.1.2), a higher spread than the original distribution (Fig. 3.1.3), or a model with a different kurtosis (Fig. 3.1.4). Basic error scores are indicated for each of the error distributions. The behavior of different error scores for each of the presented error distributions is shown in Table 3.5.

scores are increased by the amount of model spread proportionally to the scaling factor used for the modification of the error distribution. The MSE is just the squared distance of the RMSE and thus shows the same characteristics as RMSE with increased magnitude.

As expected, SDE does not change when changing solely the bias. When having an (almost) unbiased model (cases 2 – 4), SDE behaves exactly like the RMSE regarding the difference percentage and magnitude. R^2 has the inverse indication of the other scores, as a higher value indicates a better result. R^2 drops when having a biased or skewed distribution and is

Table 3.5: Error scores computed from the error distributions of Fig. 3.1. This table shows the percentage of change (rounded) in relation to the unmodified original score. The original values this table is computed from can be found in Table C.2 in the appendix. The colors denote the size of the respective error, where green means low error and yellow represents a high error. The NMSE is the variant set out in [38].

#	Error Distributions	Bias	MAE	RMSE	MSE	SDE	R^2	mRSE	KL	MASE	NMSE	MAPE
0	Unmodified original	0	0	0	0	0	0	0	0	0	0	0
1	Biased distribution (+ 0.1)	1721	56	33	78	0	-27	36	33	56	-46	210
2	Skewed distribution	1	43	33	77	33	-27	34	33	43	1	141
3	More spread (* 1.5)	80	48	48	118	47	-41	48	48	48	124	50
4	Different kurtosis	-29	-10	0	0	0	0	-1	0	-10	-17	-20

especially sensitive to a higher model spread. However, a change of kurtosis has little impact on the score. The mRSE and KL measures are very close to each other and react very similar to the RMSE error and the inverse of R^2. The temporal normalization Δo_n in mRSE therefore seems to have little impact compared to the second normalization term.

MASE has a different value domain due to the normalization term. However, for the same evaluated data set (as in the present case), the percent-wise change of the MASE corresponds exactly to the MAE error. MAPE is very sensitive to changes of bias, skewness, and kurtosis. It scales linearly when increasing the spread of the error distribution. The NMSE error has a behavior which is rather hard to interpret. When adding skewness to the error distribution, the NMSE actually decreases (due to high weighting of a number of points close to the average power generation). It is rather unsensitive to changes of bias, skewness, or kurtosis. Interestingly, both MAPE and NMSE do not seem to relate to any of the other scores.

The following tables (Tables 3.6 – 3.8) show the relationship of the normalized scores to the basic measures they were constructed from. The tables were constructed by computing the difference between basic error measure and normalized score in the way $|e_{\text{base}} - e_{\text{norm}}|$, where e_{base} is the basic error measure and e_{norm} is an error score that includes normalization. The lower the number of the difference term (ideally, 0), the closer is the interpretation of the normalized score to the non-normalized score. This typically is a desirable property as the basic error measures are very well understood and have strong discrimination ability (an analysis of this effect is furthermore performed in Section 3.8). A low number therein indicates a similar discrimination ability of the normalized score.

Table 3.6 shows scores derived from the MAE score. As can be seen, the relative differences of the MASE score are exactly in accordance to the MAE score, MASE therefore has the exact same discrimination ability as MAE. MAPE, on the other hand, varies widely.

Table 3.6: Difference in change from scores based on MAE score.

#	Error Distributions	MAE	MASE	MAPE
1	Biased distribution	0.0	0.0	**154.1**
2	Skewed distribution	0.0	0.0	**97.5**
3	More spread	0.0	0.0	**1.6**
4	Different kurtosis	0.0	0.0	**10.5**

Table 3.7 shows scores which are based on the MSE, which, in this case, only is the NMSE. As can be seen, NMSE exposes widely different characteristics from the MSE.

Table 3.8 shows the change of relative difference for scores based on RMSE. As can be seen from the table, the KL score exposes the same characteristics as the RMSE with no change

Table 3.7: Difference in change from scores based on MSE score.

#	Error Distributions	MSE	NMSE
1	Biased distribution	0.0	**123.8**
2	Skewed distribution	0.0	**75.7**
3	More spread	0.0	**6.8**
4	Different kurtosis	0.0	**17.0**

in relative difference. The mRSE is very similar, and only introduces a slight deviation of the RMSE value. In accordance to expectation, SDE is the same as RMSE with the exception of the biased forecast. Somewhat surprisingly, the R^2 score also exposes characteristics that make it similar to the RMSE (for the R^2 score, the value is computed using $|e_{base} + e_{norm}|$ to account for the interpretation of R^2 where greater values are better score values).

Table 3.8: Difference in change from scores based on RMSE score.

#	Error Distributions	RMSE	SDE	mRSE	KL	R^2
1	Biased distribution	0.0	**33.4**	2.7	0.0	6.0
2	Skewed distribution	0.0	0.1	0.8	0.0	**5.9**
3	More spread	0.0	0.1	0.2	0.0	**6.1**
4	Different kurtosis	0.0	0.1	**0.3**	0.0	0.1

3.7 Case Study: Correlation of Error Scores

In this case study, the correlation of the presented error scores given a real-world data set of 45 wind farms (with data sets called *wf1* ... *wf45*) is analyzed (*EuropeWindFarm* dataset, publicly available at [76]). A normalization in the interval $[0,1]$ is performed for each input dimension. For each wind farm, as forecasting model an extreme learning machine (ELM, see [119] for details of this forecasting model) is trained using 1500 randomized hidden units and a regularization parameter $\lambda = 10^{-3}$, which are determined using the training data set so that the model does not overfit. As activation function for the ELM, a rectified linear unit function is chosen. The data set is split into training (1/3) and test data set (2/3). The complete results is given in Table D.1 in the appendix.

As can be seen from some of the results, some scores turn out to be problematic in some situations. In particular, this can be observed for the MAPE score (e.g., wf39, wf43) or the NMSE score (e.g., wf4, wf30, wf32), where the scores have massive outliers as results and high values for the standard deviation.

Based on the results of the table, the correlation of the error measures is investigated in Fig. 3.2. The figure shows the absolute value of the correlation (Pearson correlation coefficient with interval $[0,1]$). The correlation matrix is computed using the 45 wind farms for each of the 11 investigated error measures. The heat map shows the amount of correlation between the error scores. As can be seen from the figure, the elements along the diagonal line (lower left to upper right) have a correlation of 1, as the correlation of each error measure with itself is a perfect linear relationship. The bias has little correlation with the other measures. This is due to the fact that all the investigated trained models have a very small bias, in case of a model with a high bias the correlation to other error measures, such as MAE, does exist. This can, e.g., be observed in the first case study (see Table 3.5). MAE, (R)MSE, and SDE

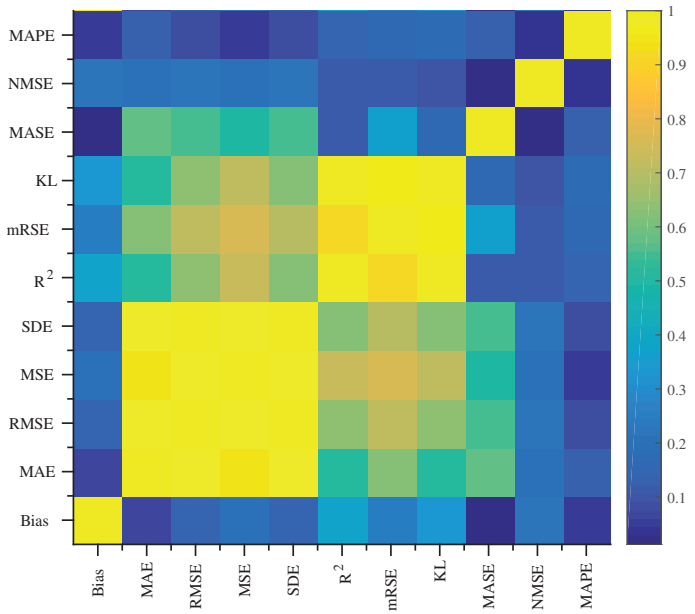


Figure 3.2: Absolute value of the Pearson correlation coefficient displayed in a matrix. Naturally, each score has perfect correlation with itself, as is displayed by the yellow diagonal line. A large block of correlated scores is formed by MAE, (R)MSE, and SDE. A second block of high correlation contains R^2, mRSE, and the KL score. MASE has moderate correlation with MAE, (R)MSE, and SDE. NMSE and MAPE have low correlation with any of the other scores.

form a central block of high correlation in the figure with correlation higher than 0.9. These measures have no form of normalization which may distort the results, therefore the scores are very understandable. Scores of this category give insights into the absolute imprecision of a forecast.

The second category of highly correlated scores are represented by R^2, mRSE, and the KL score. As can also be seen, there exists a moderate correlation to the block formed by MAE, (R)MSE, and SDE, which does make sense as the error score formulas are an extension of the basic error measures.

The MASE score forms an own category of error score. The form of normalization aims to reduce the influences from the respective location and from the observed time period. MASE still has a moderate correlation to the first block of error scores which can be expected, as it measures the absolute distance.

MAPE and NMSE have almost no correlation with any of the other error measures. Both scores employ a form of normalization which can lead to singularities in this application (as situations with very little power generation or a power generation close to the average generation are very likely to exist in the dataset). The error scores can, thus, be influenced dramatically by a very small number of individual errors. The problematic nature of these scores is further supported by the extreme range of the scores in the observed datasets which is shown in Table D.1 in the appendix. The fact that these scores have very little correlation

with MAE, (R)MSE, and SDE shows that NMSE and MAPE struggle to correctly assess the quality of a forecast correctly in this application. As NMSE and MAPE are also based on one of the basic error measures (RMSE and MAE, respectively), a certain amount of correlation *should* exist, such as is exhibited, e.g., by the MASE score.

3.8 Case Study: Discrimination and Abstraction Ability

As has been laid out in Section 3.1, error scores for model selection should ideally be able to discriminate well performing from worse performing models and at the same time they should be able to abstract from the "difficulty" of the forecasting task. Based on the results from Tables D.1 and D.2, this section aims to investigate these properties in more detail.

The basic assumption is that, given the same forecasting task that contains a nonlinear relationship of predictors and predictand, a forecasting model that is able to model nonlinear relationships (and is reasonably parameterized) should perform equally good or better than a linear model on average. An error score should therefore be able to discriminate the better performing model from the weaker model. For this case study, we therefore include a simple linear regression model in addition to the ELM model used in Section 3.7 using the same experimental setup. The results are given in Tables D.1 and D.2 in the appendix.

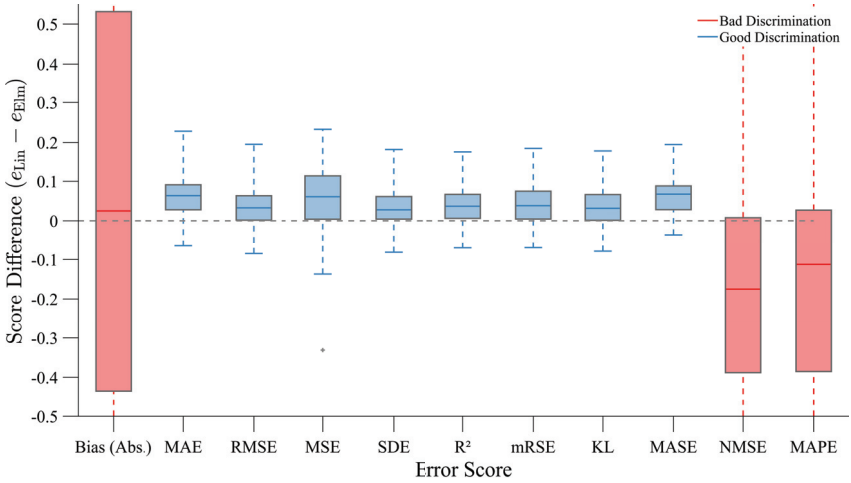


Figure 3.3: Analysis of discrimination ability of error scores. The ordinate denotes the difference of the error scores, where a higher value indicates a better discrimination. A value of 0 indicates no discrimination ability. As can be seen from the figure, MAE based scores such as MASE perform very strong regarding the discrimination ability, while the absolute bias, NMSE, and MAPE perform very weak in this regard. Red colored box plots denote the lower quartile of scores is below 0.

Fig. 3.3 investigates the discrimination ability of the two investigated models. In the figure, the difference of the two scores on the ordinate axis is given in the form $e_{\text{Lin}} - e_{\text{ELM}}$, where e_{Lin} is the error of the linear model and e_{ELM} is the error of the ELM model. All scores are normalized by dividing the individual errors by the median error value for each score. Thus,

assuming that the error of the linear model is higher, the difference should result in a value larger than 0. The box plots in the figure summarize the pairwise differences of each of the investigated wind farms for every investigated error score. Thus, a higher position of the box plot in the diagram indicates a better discrimination ability. The boxes denote the 1st and 3rd quartiles. In the figure, the error measures with lower quartiles bigger than 0 are colored in blue, whereas the other error measures have red color.

As can be seen from the figure, (Abs.) Bias, NMSE, and MAPE have a negative discrimination ability. This means these scores are not very well applicable for model selection. The other scores perform better regarding their discrimination ability. In particular, scores based on the MAE (such as MASE) show strong discrimination ability, as is, e.g., indicated by the high positions of the 1st quartile and of the median. Unlike the other scores, the R^2 score is computed using $R^2_{\mathrm{ELM}} - R^2_{\mathrm{Lin}}$ to account for the interpretation of R^2 where larger values represent better scores.

Fig. 3.4 further investigates the abstraction ability of the error scores and forecasting model. In the figure, the ordinate shows the absolute difference of each score from the median error of this score, i.e., each error score item is computed with

$$e_{\mathrm{mod}} = |e - \mathrm{median}(\boldsymbol{e})| \,, \tag{3.15}$$

with $\boldsymbol{e}$ being the tuple of error score values of every evaluated wind farm for a particular error score. Prior to that, the scores are normalized by dividing each score with the median value of $\boldsymbol{e}$ for better comparability. The box plots denote distribution of e_{mod}. The assumption regarding a good abstraction ability from the data set is that all scores are close to the median error. Good abstraction therefore is achieved if the box plots are compact in their representation and are located as close as possible to a value of 0. The box plots again show the median and the lower and upper quartiles.

As can be seen from the lower (1st) and upper (3rd) quartiles in the box plots in the figure, the non-normalized measures (A. Bias, MAE, RMSE, MSE, SDE) show a larger spread of values than the normalized scores (except for NMSE and MAPE), which meets the expectation. R^2, mRSE, KL, and MASE can reduce the average spread of errors, as is denoted by the lower position of the box plots. NMSE and in particular MAPE expose an extreme spread of values which is indicated by the box plots. As scores similar to the median error indicate a good abstraction from the data set, these scores therefore expose poor abstraction capabilities in this sense.

3.9 Discussion of Deterministic Error Scores

This section discusses the use of deterministic error scores. In Section 3.9.1, general properties of the employment of error scores are discussed. Section 3.9.2 deals with the handling of different forms of normalization. Section 3.9.3 discusses how to deal with multiple time horizons and the conditional assessment of the spread of errors.

3.9.1 General Error Score Properties

The optimal score heavily depends on the desired application and market participant. Considering the basic scores, the RMSE should always be preferred over the MSE score when presenting results, as the error units are better understandable. However, for model training the MSE score is equivalent (and faster to compute). RMSE and MAE are equally important,

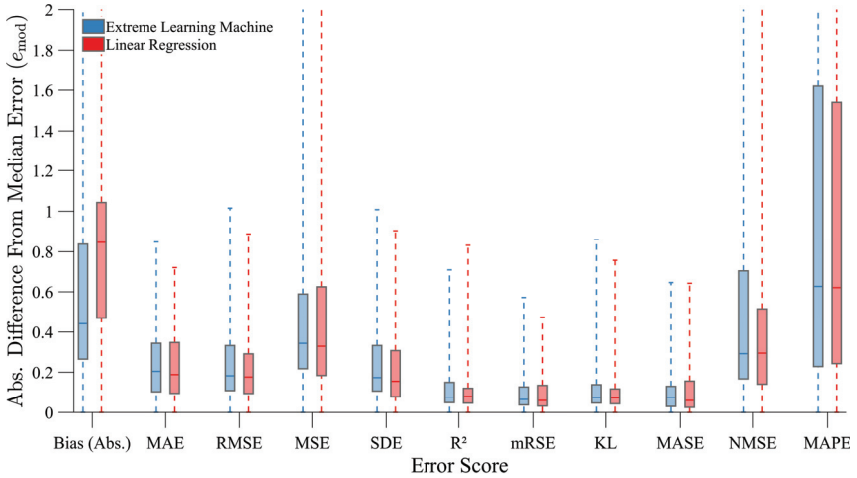


Figure 3.4: Analysis of abstraction ability of error scores. The ordinate represents the difference between the median error of the data set and each score value computed by Eq. 3.15. A value close to 0 therefore is desirable. As can be seen from the figure, the normalized error scores perform higher abstraction ability than the non-normalized basic error measures. NMSE and MAPE fail to perform an abstraction.

the more appropriate score depends on the target application (or the target audience in the industry, respectively). For electricity trading (where a deviation typically has linear costs), the MAE is preferable. The use of cost/reward functions or loss functions makes sense in conjunction with the MAE when the monetary consequences are in the center of interest. This also applies to other areas where there is a linear relationship of the forecasting error and a quantity (e.g., money) to optimize. This may, e.g., be the case when an error fee has to be paid for each *kWh* of deviation between forecast and observation. The function can be defined asymmetrically for power surplus or power deficit (e.g., the pinball function, see Section 5.6.4); furthermore it can be designed in a nonlinear fashion. For grid operators and other grid stability oriented market participants, the RMSE score is more appropriate (as the problematic nature of extreme errors is reflected more clearly).

The use of skill scores, while not investigated in detail in this section, makes sense to compare models to each other, however, they can be misleading if compared to very weak models, such as a climatological model (as a forecasting model may be presented in a too positive way). In any case, the precise computation of all used error scores should be given, as some of the score names are defined in different ways.

3.9.2 Use of Normalizations

Considering the forms of normalization, their application again depends on the desired outcome. Normalization by o_{inst} is very unproblematic and can (and should) be conducted in every case for model comparison. While the idea of the normalization with o_n is elegant for quality assessment (percent-wise error), it has the unwanted property of a singularity when having a power generation near the low end of the generation spectrum. Thus, scores

based on this normalization technique can eventually be dominated by a small number of measurements where the power generation happens to be very low (as can be observed by the MAPE score in Table D.1). Beyond unequal weighting, the division through o_n can turn out numerically problematic (as o_n may be 0). Therefore, if an error score with this type of normalization has to be used, we think error measures using this form of normalization (such as the popular MAPE score) should define some form of lower bound b for the denominator which results in a limit for maximum weighting, such as

$$o_{n,\text{limit}} = \max(o_n, b). \tag{3.16}$$

As a rule of thumb, we think the value of b should be chosen in the range of $0.01 \leq b \leq 0.2$ assuming o_n is in the range $[0,1]$ (e.g., after the normalization with o_{inst}).

The idea of the normalization with $|o_n - \bar{o}|$ is to weight extreme power generation situations lower. When computed on a data set, it aims to statistically penalize data sets who have little variation in the data. While the idea is commendable, it has the same singularity problems regarding the normalization (see NMSE [38] in Table D.1), thus should be treated again with some form of threshold limit, such as performed by Eq. 3.16.

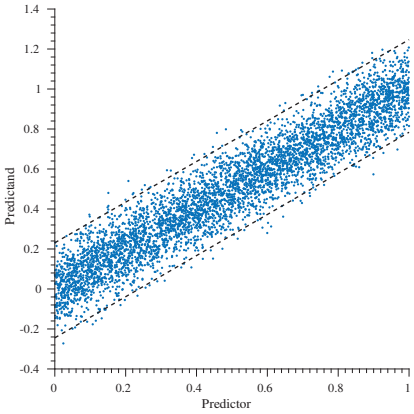
We think the normalization by Δo makes sense in filtering out the impact of highly dynamic and thus difficult weather situations and the effects of "step" errors, i.e., errors which occur when misjudging the point in time of a sudden ramp in the power generation. This can frequently happen when using numerical weather predictions. An approach for this type of normalization is discussed in [55]. In general, this form of normalization has a similar goal as the normalization technique $|o_n - \bar{o}|$, however, it seems to be more tailored towards time series (all other normalization techniques basically can be computed for standard regression problems as well). Again, in itself, this form of normalization has the same disadvantages as the two previous normalization techniques.

However, when the normalization terms are *summarized* in the denominator, such as in mRSE, KL, RAE, or MASE, the normalization lowers the impact of the *entirety* of weather situations which occurred in the data set in the evaluated period. In addition, this form of normalization eliminates the impact of singularities and therefore the need of the thresholding that is proposed in Eq. 3.16. Especially when comparing algorithms which were evaluated on different data sets, these scores can help to make the algorithms better comparable. We therefore think scores based on this form of aggregated normalization are preferable for this task.

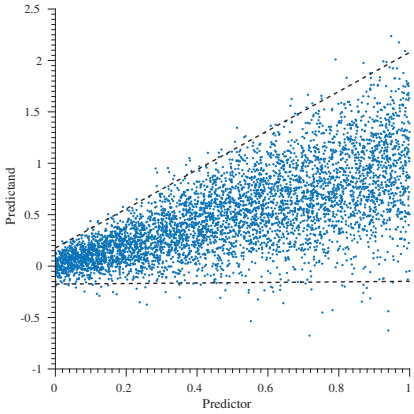
3.9.3 Multiple Forecasting Time Steps and Deviation Assessment

For some applications, it may make sense to aggregate the errors for each forecasting time step individually instead of one score for the whole forecasting time span. When an aggregation over multiple forecasting time steps is aspired, this is typically performed in the literature just by computing the score while including the errors from all time horizons. However, as errors typically increase when forecasting time steps which are further away from the forecasting origin, errors close to the forecasting horizon may dominate the overall result. Though not practiced on a regular basis, an option worth investigating is the use some form of time horizon smoothing, e.g., in the form

$$\text{RMSE}_{\text{mod}} = \frac{1}{k_{max} - k_{min} + 1} \sum_{k=k_{min}}^{k_{max}} \frac{\text{RMSE}_k}{k}, \tag{3.17}$$



(1) Homoscedastic uncertainty.



(2) Heteroscedastic uncertainty.

Figure 3.5: Different forms of uncertainty. In Fig. 3.5.1, a homoscedastic uncertainty estimation is present, the uncertainty distribution is equal throughout the input space of the predictor. Fig. 3.5.2 shows a heteroscedastic uncertainty. In this form of uncertainty, the amount of uncertainty changes depending on the predictor input space.

when aggregating forecast errors for multiple forecasting time steps, such as, e.g., performed in [99].

Even for non-probabilistic error scores we think a measure such as the SDE should be determined in the evaluation, as it gives an insight on the anticipated variation of errors. A disadvantage of this form of deviation assessment is the implicit normal distribution assumption of the error, which may not always hold. Furthermore, the error distribution is computed over the complete evaluated data set, assuming an equal error distribution over the predictors (i.e., *homoscedasticity* assumption), which is usually not the case. This form of uncertainty representation is visualized in Fig. 3.5.1. However, this form of uncertainty assessment would fail to correctly assess the uncertainty in case of the conditional uncertainty distribution of Fig. 3.5.2. Situation-dependent uncertainties can be assessed using probabilistic uncertainty assessment techniques which can be *heteroscedastic*, i.e., they are able to determine uncertainty conditionally for different areas of the input space. This form of uncertainty representation is shown in Fig. 3.5.2. While more difficult to model and assess, heteroscedastic uncertainty usually fits the actual uncertainty in a power forecasting context more precisely, as, e.g., the absolute error of a forecasting algorithm is higher in high-wind situations than in low-wind situations. The area of probabilistic forecast verification is detailed in Section 6.

3.10 Summary of this Section

This section presented some of the most frequently used deterministic error scores and gave insights on how to use them depending on the desired application. With regard to using an error score for choosing the best performing method (i.e., for model selection), we laid out two principal requirements for such error scores. As a primary requirement, it has to be able to discriminate strong performing models from weaker models. Furthermore, it should be able

to perform a best possible abstraction from the "difficulty" of the problem in order to make reported error scores more comparable. The abstraction property can be achieved using some form of normalization of the error scores. There are a number of different normalization types. A categorization of error scores by their basic error measure and a normalization techniques is introduced, which simplifies the process of choosing the appropriate error measure.

In a number of case studies we investigated the characteristics of the most frequently used error scores. Regarding the basic forms of error, scores based on absolute distances (i.e., derived from the MAE), are well suited to evaluate forecasts embedded in processes with linear overall target functions, such as estimating the monetary consequences of selling electricity at energy exchanges. In contrast to that, scores based on quadratic errors (i.e., MSE or RMSE) are well suited for power grid stability-related operations, as they better account for the problematic nature of extreme errors in comparison to linear measures. As a conclusion, a better abstraction can be achieved when using the right type of normalization. As a general recommendation, the R^2 score, mRSE, KL, and MASE are well-suited for model selection, while the use of MAPE and NMSE is discouraged. However, the scores partly also are highly correlated. Therefore, some scores may suffice to get a coherent picture of the performance of a deterministic forecasting algorithm.

Chapter 4

Coopetitive Soft Gating Ensemble: A Novel Combination Approach for Deterministic Point Forecasts

In this chapter we present a novel ensemble technique that is called *coopetitive soft gating ensemble*. The proposed ensemble technique has innovations regarding three main aspects, which we will highlight in the following.

Coopetitive Soft Gating Weighting Function

The two basic approaches for setting weights, weighting and gating, both have distinct advantages, however, they are either not dynamic (cooperation or weighting), or do not allow for cooperation (competition or gating). We present a weighting scheme which combines the advantages of both cooperation and competition in the form of a *coopetitive soft gating* technique. Coopetition, see e.g., [151], is a term originally emerging from economic research which describes the concept of competitors achieving a joint advantage by cooperating. The weighting function is designed to dynamically create gradual weights based on observable external criteria. The function is applied in the ensemble in a number of ways. Therefore, it is important that the weighting technique has as few as possible parameters to enable a fast and reliable model training. This weighting function is described and analyzed in Sections 4.1 – 4.3.

Hierarchical Ensemble Structure

As has been pointed out in Section 2.6, ensemble techniques are able to improve the forecasting performance in comparison to single predictive models. In the area of power forecasting both ensembles from machine learning (that in this context are power forecasting models, PM) *and* meteorological ensembles (EPS, MME, TLE) from multiple weather models (WM), can be used in principle. While a number of ensemble techniques have been proposed for either of the two ensemble principles (which is discussed, e.g., in [206] and [262], respectively), the use in conjunction has not yet been analyzed to the best of our knowledge. Therefore, we introduce a novel hierarchical ensemble technique that combines the principles of meteorological ensembles and PFE. The idea is that each WM is predicted using a set of multiple PM individually to create diverse base predictors. The technique is designed to then perform the overall aggregation of the individual forecasts in a post-processing step. The hierarchical

ensemble structure is described in Section 4.4.

Multi-Scheme Dynamic Weighting

Within the hierarchical ensemble, the overall weights for each ensemble member are predicted using a dynamically weighted aggregation. These weights are created dynamically based on a number of principles. The proposed technique introduces a weighting that creates the overall weights from the overall (*global*) quality, the weather situation dependent (*local*) quality, and the lead time-dependent (*temporal*) quality. This weighting is computed for each PM *and* for each WM, leading to overall six weighting factors for each ensemble member. The weighting scheme is described in Section 4.4.

Following the description of the algorithm, the proposed technique is evaluated in a number of experiments in Section 4.5. The section closes with a discussion of the applicability of the ensemble in Section 4.6 and a conclusion in Section 4.7.

4.1 The Coopetitive Soft Gating Weighting Function

The goal of the coopetitive soft gating weighting function (CSG) is to weight ensemble members according to their performance in a way that combines the properties of both principles, ensemble member weighting and gating. To recapitulate the general ensemble aggregation formula of Section 2.6.1, the ensemble is defined by

$$\bar{f}(\mathbf{x}) = \sum_{j=1}^{J} w^{(j)} \cdot f_j(\mathbf{x}), \tag{4.1}$$

where J is the number of ensemble members (or *base predictors*) indexed by j and $f_j(\mathbf{x})$ is the prediction of ensemble member j with weight $w^{(j)} \in [0,1]$, which have to comply to

$$\sum_{j=1}^{J} w^{(j)} = 1. \tag{4.2}$$

The weights could theoretically be trained *directly* using an optimization algorithm, i.e., all $w^{(j)}$ are subject to the optimization. However, this approach has a number of drawbacks. First, we end up with a $J-1$-dimensional optimization problem, which makes it more difficult to find the global optimum of weights and has high optimization run-time when having a high number of ensemble members. Furthermore, while this approach can yield accurate results, it is a "dumb" approach that is not able to adapt to dynamically changing conditions.

Therefore, we aim to define a novel weighting technique that defines a *function* that is able to assign weights for the ensemble members through an observable criterion of their performance. As the performance of a forecasting technique can be determined in different situations, the weighting technique allows for a more flexible application than a direct weight assignment. The final weights are built using a quality estimation of the different ensemble members, which typically is represented in the form of an error. For deterministic point forecasts this can be error scores, e.g., the RMSE. More details on deterministic error scores can be found in Section 3. Having a quality estimate for each ensemble member, the proposed weighting technique must fulfill the following requirements:

1. It must return a score ordered from high weights (low error) to low weights (high error), i.e., the inverse of what the error score is initially represented in.

2. It must be able to weight errors nonlinearly, as the optimal weighting scheme possibly can not be represented in a linear relationship. The "amount" of nonlinear weighting should be controllable by the user.

3. It must not be affected by the value range of the error scores, i.e., it should only factor in the relative quality differences of the contributing ensemble members.

4. It must use a low number of parameters to enable fast and robust training and application.

5. It must retain the value range of the ensemble prediction, i.e., it has to fulfill Eq. 4.2.

We fulfill criteria 1. – 4. using the weighting function $\varsigma'_\eta\colon (\mathbb{R}^+)^{N+1} \to \mathbb{R}^+$ with

$$\varsigma'_\eta(\mathbf{\Omega}, \omega) = \frac{\sum_{n=1}^{N} \mathbf{\Omega}_n}{\omega^\eta + \epsilon}, \quad \eta \in \mathbb{R}_0^+, \tag{4.3}$$

where $\mathbf{\Omega} \in \mathbb{R}^N$ is a tuple of values containing all N reference quality estimates each denoted with $\mathbf{\Omega}_n$ with $n \in 1, \ldots, N$. The variable ω is an arbitrarily evaluated point of the function, which typically is chosen to be an element of $\mathbf{\Omega}$. The real number $\eta \geq 0$ controls the amount of exponential weighting. The higher the value of η, the higher the weight of models with low errors. ϵ is a very small number that avoids division by 0 in the unlikely case that ω is 0. Assuming that $\omega \in \mathbf{\Omega}$ and that $\varsigma'_\eta(\mathbf{\Omega}, \omega)$ is computed for each element in $\mathbf{\Omega}$, we achieve criterion 5. by adjusting the weighting function by normalizing with the sum of the weights in $\mathbf{\Omega}$, i.e., using the function $\varsigma_\eta\colon (\mathbb{R}^+)^{N+1} \to [0, 1]$ with

$$\varsigma_\eta(\mathbf{\Omega}, \omega) = \frac{\varsigma'_\eta(\mathbf{\Omega}, \omega)}{\sum_{n=1}^{N} \varsigma'_\eta(\mathbf{\Omega}, \mathbf{\Omega}_n)}. \tag{4.4}$$

This form of the coopetitive soft gating formula can further be simplified to

$$\varsigma_\eta(\mathbf{\Omega}, \omega) = \frac{1}{(\omega^\eta + \epsilon) \sum_{n=1}^{N} ((\mathbf{\Omega}_n)^\eta + \epsilon)^{-1}}, \tag{4.5}$$

which yields the final coopetitive soft gating formula. The results of this function can then serve as the basis for determining the individual weights of the base predictors within the proposed ensemble technique.

An example of coopetitive soft gating with different values of the parameter η is given in Fig. 4.1 and Table 4.1. The table shows a number of reference error values $\mathbf{\Omega}$ in the range $0.2 - 0.42$ and the respective weights given different η values. As can be seen from the example, the higher the value of η, the higher the best performing model is weighted. A value of $\eta = 0$ leads to an equal weighting of all members.

An advantage of this form of weighting is that it only has a single weighting parameter η for optimization (unlike methods derived from exponential functions in the form $f(x) = e^{Ax+B}$, e.g., in [195], polynomials, sigmoid functions with $2 - 4$ parameters, or direct optimization with $J - 1$ parameters).

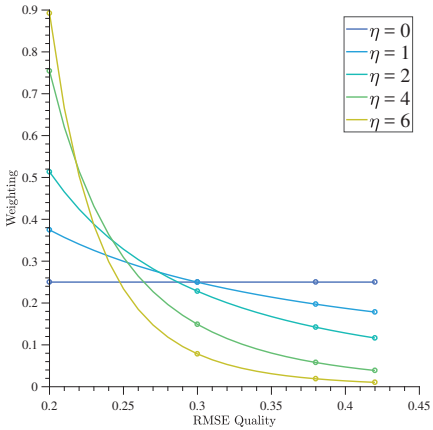


Figure 4.1: Visual representation of weighting with different values of η. Numerical values of this example are given in Table 4.1.

Table 4.1: Example of resulting weights of a set of four error values $\mathbf{\Omega}$ depending on the value of parameter η. A visual representation of the values is given in Fig. 4.1.

	Error Values Ω			
η	0.2	0.3	0.38	0.42
0	0.25	0.25	0.25	0.25
1	0.37	0.25	0.20	0.18
2	0.51	0.23	0.14	0.12
4	0.75	0.15	0.06	0.04
6	0.89	0.08	0.02	0.01

4.2 Evaluation of the Coopetitive Soft Gating Weighting Function

In an ensemble with fixed weights there are *optimal weights* which represent the optimally weighted combination of the base predictors. In order to provide sensible results, the function that assigns weights to the base predictors *has to be able* to actually create these optimal weights (or at least has to produce weights that are very close to the optimal results). These target weights are denoted with $w_t^{(j)}$ with $j = 1,\ldots,J$.

Approximation Quality

In an experiment we want to investigate the performance regarding the approximation quality of the coopetitive soft gating function in comparison to other functions which can in principle be used for the creation of weights. Goal of the experiment is to find out whether the coopetitive soft gating function is able to create a variety of target weights $w_t^{(j)}$ given some error values (i.e., values that can be used as $\mathbf{\Omega}$ in Eq. 4.5).

For the evaluation we repeatedly created J random pairs of performance criterion (error value) e_j and corresponding weight $w_t^{(j)}$ with $j = 1,\ldots,J$. The error values e_j are created in

the range $[0,1]$ while the sum of target weights $w_t^{(j)}$ conforms to Eq. 4.2. The sensible base assumption is that the lower the error value e_j, the higher the corresponding weight $w_t^{(j)}$.

The pairs are created in the following way: A number of $j = 1,\ldots,J$ error values e_j are created by drawing from a uniform distribution in the range $[0,1]$. Then, the drawn samples are sorted to be monotonically increasing represented by $\mathbf{e} = (e_1,\ldots,e_J)$ with $e_j \leq e_{j+1}$. The corresponding target weights $w_t^{(j)}$ are created by drawing iteratively from a uniform distribution in the range $[0, w_{\text{Rem.}}]$ with $w_{\text{Rem.}}$ being the remaining mass of weights which is $w_{\text{Rem.}} = 1$ in the first iteration. In the j-th iteration, $w_{\text{Rem.}}$ is set to be $w_{\text{Rem.}} = 1 - \sum_{j'=1}^{j-1} w_t^{(j')}$. In the last ($J$-th) iteration, the drawn probability is set to $w_t^{(J)} = w_{\text{Rem.}}$. The J draws are sorted to be monotonically decreasing and then represent a tuple of sorted target weights

$$\mathbf{w}_t = \left(w_t^{(1)},\ldots,w_t^{(j)},\ldots,w_t^{(J)}\right). \tag{4.6}$$

The *goal* of the experiments then is to optimize the parameters of the weighting functions (i.e., the parameter η in the case of the CSG function) in order to create the best possible approximation of $\mathbf{w}_t$ based on the values in $\mathbf{e}$ with respect to the mean absolute error (MAE, see Section 3.2). Thereby, the theoretical ability of the weighting functions to serve for the creation of weights in an ensemble is analyzed (in the practical application, the target weights are given from the training data set when using a weighted combination of base predictors that minimize the overall forecasting error).

We repeated the creation process 1000 times and evaluated the performance of the CSG technique and other approaches. Fig. 4.2 visualizes the experimental setup (with only 50 sample draws in this case for the sake of better visibility). As can be seen from the figure, the process of randomly drawing samples creates both sets of error-weight combinations which are very steep in some cases (one dominant target weight, followed by a set of lower target weights $w_t^{(j)}$), as well as more equally distributed target weights in other cases. Furthermore, the respective errors $\mathbf{e}$ vary in their position on the abscissa.

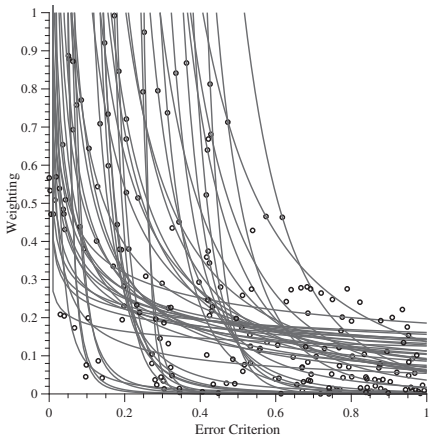


Figure 4.2: Error and optimal weight combination in 50 experiments with approximated function by CSG.

As comparison approaches, we included a simple linear regression, variants of a polynomial regression, and sigmoid variants which are used as the Platt calibration function [195]. The functional forms of the comparison approaches are detailed in Table 4.2. All algorithms are optimized using a convex interior-point optimization algorithm [31].

Table 4.2: Overview of functions used for the evaluation.

Name	**Formula**	**Nr. of Parameters**
Coopetitive Soft Gating (CSG)	see Eq. 4.5	1
Linear Regression (Deg. 1)	$w(e) = a_1 \cdot e + a_0$	2
Poly. Regression (Deg. 2)	$w(e) = a_2 \cdot e^2 + a_1 \cdot e + a_0$	3
Poly. Regression (Deg. 3)	$w(e) = a_3 \cdot e^3 + a_2 \cdot e^2 + a_1 \cdot e + a_0$	4
Sigmoid Function (2 Param.)	$w(e) = \frac{1}{1+\exp(a_1 \cdot e + a_0)}$	2
Sigmoid Function (3 Param.)	$w(e) = \frac{a_2}{1+\exp(a_1 \cdot e + a_0)}$	3
Sigmoid Function (4 Param.)	$w(e) = \frac{a_2}{1+\exp(a_1 \cdot e + a_0)} + a_3$	4

The results of the approximation experiment are given in Table 4.3 for a value of $J = 4$ and in Table 4.4 for a value of $J = 10$. The table shows MAE values between target weight $w_t^{(j)}$ and computed weight $w^{(j)}$ (computed using Eq. 4.5 in the case of the CSG function) in the form

$$\text{MAE} = \frac{1}{J} \sum_{j=1}^{J} \left| w_t^{(j)} - w^{(j)} \right|. \tag{4.7}$$

In the table, the distribution of the MAE values of the 1000 repetitions is characterized using the overall maximum error, the 90% percentile, the median, the 10% percentile, and the minimum overall error. Furthermore, the mean error and the standard deviation of the MAE over 1000 iterations are given.

Table 4.3: Statistical evaluation of the error of the function approximation experiment with $J = 4$. All measures computed on the single mean absolute error (MAE) of each experiment.

$J = 4$	CSG	Linear Regr.	Poly (Deg. 2)	Poly (Deg. 3)	Sigmoid (2 Param.)	Sigmoid (3 Param.)	Sigmoid (4 Param.)
Max	**0.220**	0.255	0.273	0.273	0.241	0.248	0.250
90% Quant.	**0.084**	0.191	0.111	0.094	0.089	0.088	0.125
Median	0.034	0.066	0.028	**0.011**	0.026	0.020	0.019
10% Quant.	0.006	0.019	0.005	**0.000**	0.005	0.002	0.002
Min	**0.000**	0.002	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Mean	0.041	0.084	0.046	**0.033**	0.039	0.036	0.043
Std. Dev.	**0.033**	0.063	0.050	0.048	0.039	0.046	0.058

As can be seen from Table 4.3 with $J = 4$, regarding the maximum and the 90% percentile, CSG performs best. Regarding the median, a number of approaches are able to yield more accurate results than the CSG technique. All approaches except the linear regression approach yield very low errors for the 10% percentile and the minimum MAE. For the mean error, the polynomial approximation with degree of 3 yields the best score. Regarding the standard deviation, however, the CSG technique performs most consistent.

When increasing the number of members to $J = 10$ as performed in the experiments for Table 4.4, the sigmoid variants perform stronger, the sigmoid function with 2 parameters

Table 4.4: Statistical evaluation of the error of the function approximation experiment with $J = 10$. All measures computed on the single mean absolute error (MAE) of each experiment.

$J = 10$	CSG	Linear Regr.	Poly (Deg. 2)	Poly (Deg. 3)	Sigmoid (2 Param.)	Sigmoid (3 Param.)	Sigmoid (4 Param.)
Max	**0.095**	0.100	0.100	0.100	**0.095**	0.100	0.100
90% Quant.	0.054	0.098	0.091	0.084	**0.043**	0.080	0.096
Median	0.024	0.083	0.057	0.039	0.016	**0.013**	0.021
10% Quand.	0.005	0.050	0.023	0.014	0.003	**0.002**	0.004
Min	**0.000**	0.010	0.007	0.002	**0.000**	**0.000**	**0.000**
Mean	0.027	0.078	0.057	0.044	**0.020**	0.024	0.035
Std. Dev.	**0.018**	0.019	0.025	0.025	**0.018**	0.028	0.033

yields the best results regarding the 90% percentile and mean error, and the sigmoid function variant with 3 parameters yields the best median value. Regarding maximum error, minimum error, and standard deviation of errors, CSG performs on par with the sigmoid approach. All polynomial variants yield worse results throughout all metrics. Interestingly, the sigmoid function variant that uses 4 parameters in many cases performs worse than variants that use less parameters. This may be due to difficulties of the gradient-based optimization algorithm to find globally optimal values in the higher-dimensional search space.

Exemplary Weighting Scenarios

In the following, we want to give more insights into the behavior of the analyzed techniques by investigating their performance visually given a number of typical scenarios which are shown in Fig. 4.3 and Table 4.5. The parameters of the weighting function are optimized to minimize the RMSE in this experiment (see Section 3.2).

Fig 4.3.1 shows a typical weighting scenario with one technique performing stronger than the others, while the other techniques still contribute to the overall ensemble. In this scenario, all techniques behave similarly. Fig 4.3.2 shows a linear weighting scenario. In this case, the linear and polynomial models perform best, however, all other approaches still yield very similar results. Fig 4.3.3 represents a scenario in which only the best performing techniques has noteworthy weight in the ensemble (a "winner takes all" scenario). In this case, CSG and the sigmoid variants perform best regarding the RMSE, the polynomial models furthermore create non-monotonically decreasing functions. Fig. 4.3.4 shows the inverse scenario: Only a single weak model has to be pruned in the ensemble. Naturally, this is an ideal weighting scenario for the sigmoid techniques with more parameters (3 and 4). All other techniques create a trade-off regarding the weights. Figs. 4.3.5 and 4.3.6 show scenarios in which the criterion does violate the monotony constraint. In Fig. 4.3.5, all approaches behave similarly by creating a functional form that minimizes the RMSE, the rightmost target weight therefore is underestimated. The higher degree polynomials do overfit the function, which, in this case, yields a lower MAE. The second figure shows the inverse situation of Fig. 4.3.1. As can be seen, all models are in principle also able to create monotonically increasing functions. When constraining the η parameter of the CSG technique to $\eta \geq 0$ (which is assumed in the present case), the form of CSG is constrained to lead to an equal weighting. This may be an advantage of the CSG function, as it is able to suppress insensible results of the weighting (but instead defaults to an equal weighting). A disadvantage of polynomial regression techniques is that they are not necessarily monotonically decreasing, which may lead to insensible results, such

as can be observed with the higher-order polynomials, e.g., in Figs. 4.3.4 and 4.3.5. While especially the sigmoid variants with three and four parameters exceeded the performance of the CSG function in this case, it should be noted that some of the cases are uncommon (e.g., Fig. 4.3.5 or Fig. 4.3.6). This is also partly the result of the many parameters that these techniques can leverage, which, however, drastically increase the search overhead.

Table 4.5: Results of the weighting scenarios of Fig. 4.3 using the RMSE score. Bold text indicates best result (Poly Deg. 3 is excluded in this case because of overfitting due to the functional structure).

	CSG	Linear Regr.	Poly (Deg. 2)	Poly (Deg. 3)	Sigmoid (2 Param.)	Sigmoid (3 Param.)	Sigmoid (4 Param.)
Fig. 4.3.1	**0.027**	0.064	0.035	0.000	0.038	0.028	**0.027**
Fig. 4.3.2	0.004	**0.000**	**0.000**	0.000	0.001	**0.000**	**0.000**
Fig. 4.3.3	0.022	0.196	0.049	0.000	0.021	0.020	**0.000**
Fig. 4.3.4	0.072	0.066	0.042	0.000	0.068	**0.000**	0.002
Fig. 4.3.5	0.209	0.118	0.081	0.000	0.075	0.062	**0.034**
Fig. 4.3.6	0.141	0.147	**0.059**	0.000	0.146	0.145	0.123

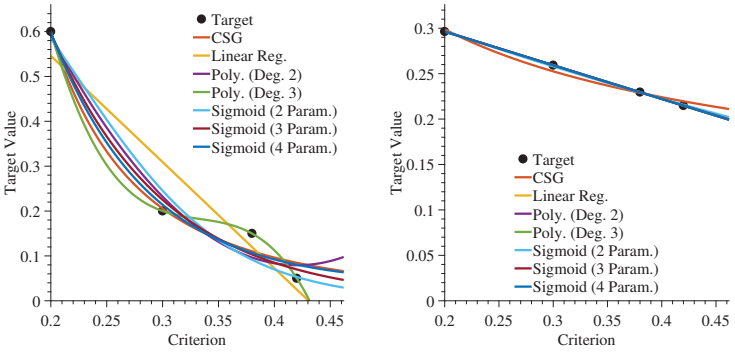
Run-Time Evaluation

Finally, we want to highlight the training times of the investigated approaches. For this experiment we additionally included the direct estimation approach for comparison purposes that is briefly mentioned in Section 4.1. In this comparison approach, the weights are trained directly which leads to a $J-1$-dimensional optimization problem (instead of the $1-4$ dimensional optimization problem using the weighting functions). The number of target weights is varied in the interval of $J \in [2, 100]$, the target weights of $w_t^{(j)}$ are sampled to form a monotonically decreasing exponential function. To enable a comparison that is closer to a real-world optimization scenario, the fitness function evaluation duration is artificially set to take 100 ms. The results are shown in Fig. 4.4. In the figure, the run-time in seconds is shown on the ordinate axis, whereas the number of ensemble members is shown on the abscissa. The values represent the average run-time of 10 repetitions.

In principle one would assume that the less parameters the weighting function has, the less iterations are required in order to train the model, which may lead to a shorter run-time. This can also be observed in the figure, as the CSG model with the lowest number of parameters takes the lowest overall run-time for all evaluated values of J. The optimization of the other models takes longer. Interestingly, the optimization of models such as the sigmoid model with 2 parameters takes longer than the optimization of some of the models with more parameters. This may be due to initialization effects or disadvantageous characteristics of the interior-point optimization algorithm [31]. The direct estimation approach (Direct Est.) does not scale, as the dimensionality of the optimization problem steadily increases.

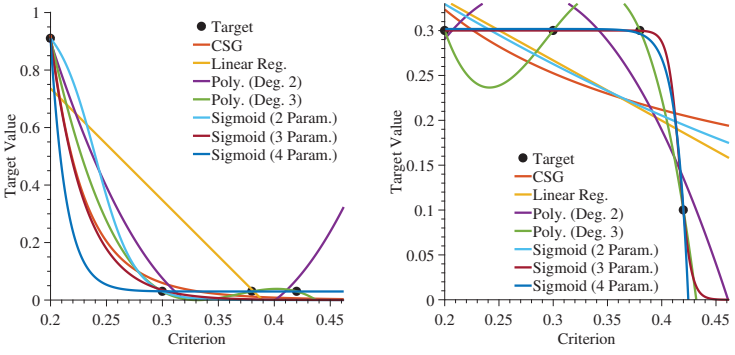
4.3 Conclusion for the Coopetitive Soft Gating Weighting Function

This section makes some concluding remarks on the coopetitive soft gating (CSG) function introduced in Section 4.1. The CSG function is specialized for the dynamic creation of



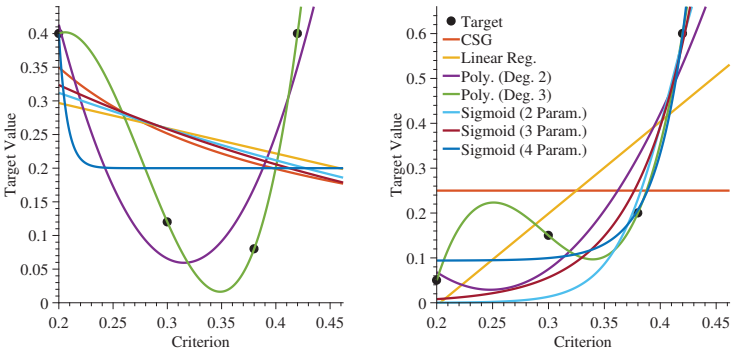
(1) Typical weighting scenario.

(2) Linear weighting scenario.



(3) Winner takes all.

(4) Pruning of weak model.



(5) Uninformative criterion.

(6) Inverse weighting.

Figure 4.3: Examples of possible weighting scenarios for the CSG technique, variants of a polynomial regression, and variants of sigmoid functions. The polynomial functions tend to create an overfitted function. The functional form of CSG and the sigmoid variants is more constrained, which is more appropriate for the given weighting task.

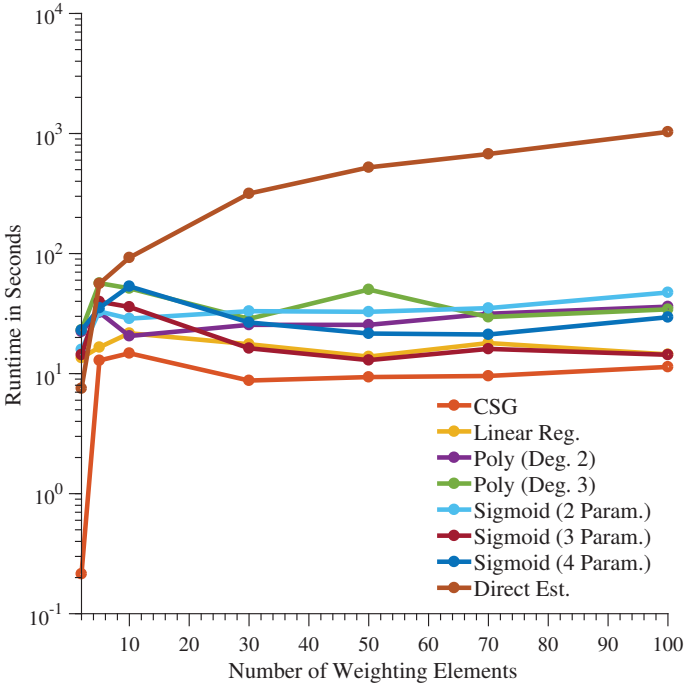


Figure 4.4: Run-time experiments for the investigated weighting functions. As can be seen from the figure, the CSG technique takes the lowest run-time, as it only has a single parameter to optimize. The other investigated approximation techniques vary in their run-time while taking up to five times as long as the CSG technique. The direct estimation approach does not scale well, as the dimensionality of the optimization problem increases with a rising number of weighting elements.

weights depending on an external (error) criterion. The base assumption is that a low error corresponds to a high weight.

As has been shown in the experimental evaluation, the CSG function is able to take various forms. Therein, it is able to compete with functions that have more degrees of freedom. As we have shown in the approximation experiments in Section 4.2, the CSG function is able to be competitively accurate for creating weights for a different number of ensemble members. Regarding the maximum error, minimum error, and the standard deviation, the CSG technique is among the best approaches throughout the experiments. It should be noted that for the weighting of ensemble members, the applicability (i.e., speed of computation, reliability) of the weighting technique may prevail the most precise determination of weights in practice. This is because a very precise (i.e., close to theoretically optimal) determination of weights "only" slightly alters the weights for the aggregation of base predictors (where each base predictor should ideally yield accurate results itself).

The main advantage of the CSG function thus is that it only has a single parameter for optimization which considerably simplifies the optimization task. Therefore, it can easier be

applied for the simultaneous optimization with not just a single, but multiple error criteria. For instance, a CSG function can be used for a set of error criteria, the results of each CSG function can then be combined to find the overall optimal weighting. Due to the lower number of parameters in comparison to other techniques, the overall optimization can be optimized faster and with higher stability, which in turn enables a reliable combination of weighting factors from more external influencing factors. It is beneficial that the meaning of the optimization parameter is very clear as a higher value leads to a higher weighting of the strongest models. When having multiple combined weighting factors, the value of the parameter η therefore also makes the importance of a particular influencing external factor very clear to the user (where a higher value indicates high dependence of the ensemble quality from the currently considered external error criterion).

4.4 The Coopetitive Soft Gating Ensemble (CSGE) Algorithm

As outlined in Section 2.6.1, ensemble models typically aim to exploit a sole principle for ensemble generation. The proposed technique aims at using multiple weighting principles. The principal structure of the proposed ensemble technique is visualized in Fig. 4.5. The weighting of the ensemble remains the same as in Eq. 4.1. However, we have a hierarchical ensemble structure: For each weather forecasting model $\psi = 1,\dots,\Psi$ (which can be an arbitrary NWP of an EPS, MME, or TLE, e.g., of an intraday or day-ahead model, for a particular time step to be forecasted), a number of power forecasting models $\varphi = 1,\dots,\Phi$ are used to forecast the target predictand for each weather forecasting model. The power forecasting models do not necessarily have to be the same for each weather forecasting model, but, for the sake of easier understanding, we will use the same type and number of power forecasting models for each weather forecasting model here. The overall number of ensemble participants (base predictors) J consequently is $J = \Psi \cdot \Phi$. The individual predictions of each power forecasting model are then aggregated and fused to an overall forecast in a post-processing step according to Eq. 4.1. The main innovation here is the way the single weights $w^{(j)}$ are constructed. In order to clarify the origin of each weighting term with respect to the weather forecasting model ψ and the power forecasting model φ, we define the weight of an ensemble member as $w^{(j)} = w^{(\psi,\varphi)}$.

The idea is as follows: For each ensemble participant we construct its weight using the coopetitive soft gating function (explained in Section 4.1) considering the following three aspects for both, power forecasting models *and* weather forecasting models, respectively (leading to 6 aspects in total for each of the J base predictors):

1. *Global Soft Gating*: The ensemble weights are determined for the respective model regarding the overall observed performance of a model during ensemble training. This is a fixed weighting term. Thereby, overall strong models have more influence than weaker models. This form of weighting is described in Section 4.4.1.
2. *Local Soft Gating*: The ensemble members are weighted depending on the model input (the NWP forecast) $\mathbf{x}^{(j)}_{t+k|t}$. This form of weighting assesses the quality of a model considering the current input, i.e., a local quality assessment is performed. The idea is that a number of models may have different strengths in a different set of weather situations (e.g., due to ensemble diversity effects). This form of weighting is described in Section 4.4.2.
3. *Lead time-dependent Soft Gating:* Models may have a lead time-dependent quality development. The goal of this form of soft gating is to weight the model depending on the lead

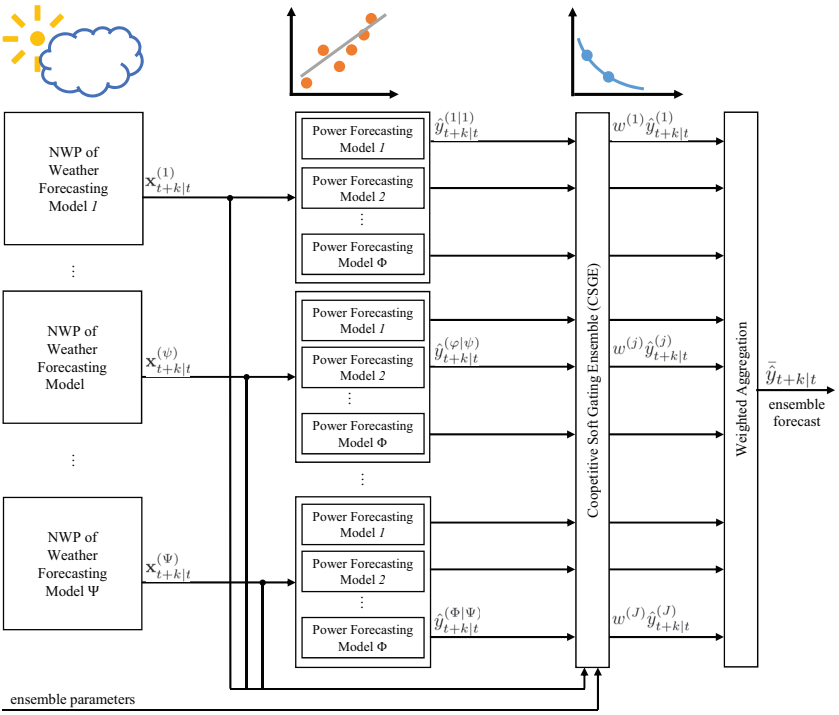


Figure 4.5: Overview of the proposed coopetitive soft gating ensemble (CSGE) model. As can be seen from the figure, the ensemble has a hierarchical structure, where a number of predictors from Ψ weather forecasting models (EPS, MME, or TLE) are used to forecast a common target predictand. Therein, Φ power forecasting models each create a power forecast for a particular weather forecast. The total number of $J = \Psi \cdot \Phi$ forecasts are combined in a post-processing stage (CSGE and weighted aggregation block in the figure) using a multi-scheme weighting and the coopetitive soft gating function.

time k. In the case of power forecasting models, methods such as the persistence method perform very well on short time horizons, while they quickly loose their quality for longer time horizons. Additionally, weather forecasting models such as intraday models typically perform very strong on short time horizons due to very recent weather measurements. This form of weighting is described in Section 4.4.3.

The overall weighting term for each ensemble member can thus be described as

$$w^{(j)} = w^{(\psi,\varphi)} = \frac{w^{(\psi)} \cdot w^{(\varphi|\psi)}}{\sum_{\psi^*=1}^{\Psi} \sum_{\varphi^*=1}^{\Phi} w^{(\psi^*)} \cdot w^{(\varphi^*|\psi^*)}}, \tag{4.8}$$

where $w^{(j)} = w^{(\psi,\varphi)}$ is the overall weight for ensemble member j computed using power forecasting model φ and weather forecasting model ψ. The weights $w^{(\psi)}$ are the *weather forecasting model* dependent weights, $w^{(\varphi|\psi)}$ are the *power forecasting model* dependent

weighting factors of power forecasting model φ computed on weather forecasting model ψ. Given the hierarchical ensemble structure, the quality of power forecasting models φ vastly influenced by the quality of the weather forecasting model ψ they are computed upon. This assumption is reflected in the process of building the product in the nominator of Eq. 4.8 where each $w^{(\varphi|\psi)}$ is weighted using the weather model weight $w^{(\psi)}$ (which makes this form of weighting more sensible in comparison to, e.g., sum-based alternatives in the basic form $w^{(\psi)} + w^{(\varphi|\psi)}$). The denominator is a normalization term which ensures $\sum_{j=1}^{J} w^{(j)} = 1$, see Eq. 4.2. The weights of the weather forecasting model can be decomposed into

$$w^{(\psi)} = w_g^{(\psi)} \cdot w_l^{(\psi)} \cdot w_k^{(\psi)}, \tag{4.9}$$

while the weights of the power forecasting model are computed with

$$w^{(\varphi|\psi)} = w_g^{(\varphi|\psi)} \cdot w_l^{(\varphi|\psi)} \cdot w_k^{(\varphi|\psi)}. \tag{4.10}$$

The indices g, l, k denote the respective weighting aspects global weighting g (Section 4.4.1), local soft gating l (Section 4.4.2), or lead time-dependent soft gating k (Section 4.4.3) for both, weather forecasting model ψ and power forecasting model φ. Using multiple weather forecasting models and multiple power forecasting models, the overall number of weights per weather forecasting model and power forecasting model combination consequently add up to six. The ensemble training process is described in Section 4.4.4. In addition to the description of the weighting factors in the different sections, Fig. 4.7 illustrates an example of the functionality of the different weighting aspects and the overall proposed technique. Fig. 4.7.1 shows the development of the global soft gating aspect over time, Fig. 4.7.2 shows the development of the local soft gating, and Fig. 4.7.3 illustrates the lead time-dependent weighting aspects. Section 4.4.5 gives an overall application example for the proposed coopetitive soft gating ensemble (CSGE) technique.

4.4.1 Global Soft Gating

The global weights $w_g^{(\psi)}, w_g^{(\varphi|\psi)}$ are fixed weights which are determined during ensemble training that takes place after the training of the individual power forecasting models, see also Section 4.4.4 . The proposed weighting technique aims at weighting the ensemble members according to their performance using coopetitive soft gating. This is performed using an error score S that compares an issued forecast to an actual observation in the form

$$S(\hat{y}_{t+k|t}, o_{t+k}), \tag{4.11}$$

where $\hat{y}_{t+k|t}$ is an issued point forecast and o_{t+k} is the corresponding observation. For the coopetitive soft gating formula, any non-negative error score can be used, e.g., the MAE error or the RMSE error, see Section 3.2 for details. For the evaluation of a data set with N forecast-observation pairs and a forecast $\hat{y}_n^{(\varphi|\psi)}$ computed on weather forecasting model ψ and power forecasting model φ, the performance can be assessed simply with

$$\bar{S}^{(\varphi|\psi)} = \frac{1}{N} \sum_{n=1}^{N} S(\hat{y}_n^{(\varphi|\psi)}, o_n). \tag{4.12}$$

The global weights $w_g^{(\varphi|\psi)}$ can then be computed for all φ with the coopetitive soft gating formula

$$w_g^{(\varphi|\psi)} = \varsigma_\eta(\bar{\boldsymbol{S}}_g^{(\psi)}, \bar{S}^{(\varphi|\psi)}), \tag{4.13}$$

$$\bar{\boldsymbol{S}}_g^{(\psi)} = \left(\bar{S}^{(1|\psi)}, \ldots, \bar{S}^{(\varphi|\psi)}, \ldots, \bar{S}^{(\Phi|\psi)}\right), \tag{4.14}$$

where ς_η is the coopetitive soft gating function of Eq. 4.5 for a power forecasting model φ computed on weather forecasting model ψ.

The global forecasting ability of a weather forecasting model ψ can be observed only indirectly as actual weather *measurements* typically are not available. As an estimate, the overall quality of a weather forecasting model can be determined using the average quality of all power forecasting models $\bar{S}_g^{(\psi)}$ for the particular weather forecasting model in comparison to the competing weather forecasting models, i.e.,

$$w_g^{(\psi)} = \varsigma_\eta(\bar{\boldsymbol{S}}_g, \bar{S}_g^{(\psi)}), \text{ where} \tag{4.15}$$

$$\bar{S}_g^{(\psi)} = \frac{1}{\Phi}\sum_{\varphi=1}^{\Phi} \bar{S}^{(\varphi|\psi)}, \tag{4.16}$$

$$\bar{\boldsymbol{S}}_g = \left(\bar{S}_g^{(1)}, \ldots, \bar{S}_g^{(\psi)}, \ldots, \bar{S}_g^{(\Psi)}\right). \tag{4.17}$$

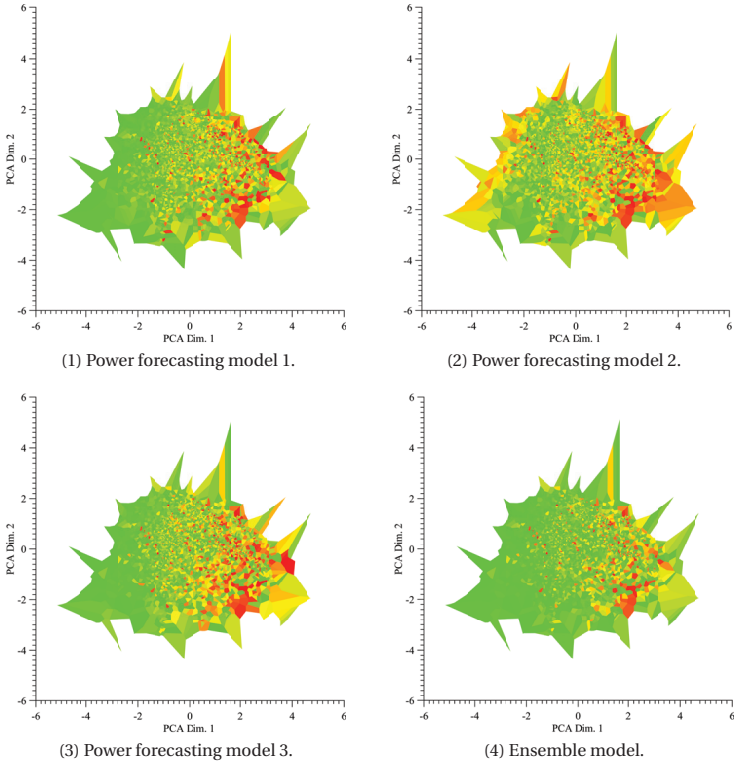
An example for the influence of the global weighting term for a number of power forecasting models is given in Section 4.4.5. Both the global weather forecasting model ad power forecasting model dependent weighting terms are computed once during an ensemble training process (see Section 4.4.4) and can then be reused for every forecast.

4.4.2 Local Soft Gating

The second weighting term depends on the values of the current NWP, i.e., the weighting is realized depending on the particular characteristics of the weather situation. The NWP forecast can be seen as a point in a feature space which characterizes the weather situation. The basic assumption is that both, weather and power forecasting models, may have strengths and weaknesses in varying areas of the feature space. This is due to the fact that different power forecasting algorithms yield different errors in certain areas of the feature space due to structure, data, or parameter diversity effects. As an assumption, for rare meteorological events in sparsely covered areas of the feature space (e.g., storms for wind turbines, or Sahara dust for photovoltaic plants), this effect may become more prominent, as a different power forecasting models used in the ensemble may fit with different precision to these rare events. Different NWP forecasts of certain weather forecasting models may also have a different precision depending on the particular situation. In particular for multi-model ensembles with different meteorological variables, more complex models may be able to resolve the above mentioned rare events more precisely. In a nutshell, we aim to exploit the advantages of each model in particular observed situations during model training using coopetitive soft gating.

In order to obtain local weights $w_l^{(\psi)}$ and $w_l^{(\varphi|\psi)}$, the neighborhood in the feature space of a weather forecast $\mathbf{x}_{t+k|t}$ has to be assessed. Similar historic weather situations are found with respect to a (historic) data set tuple $\mathbf{X}_\mathrm{H} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)$ with $\mathbf{x}_n \in \mathbb{R}^D$, which is used during ensemble training. The proposed ensemble algorithm is able to work with an arbitrary local quality assessment technique. Here we demonstrate the application with a simple nearest neighbor technique. Other techniques for assessing locality, such as multi-linear interpolation,

are investigated, e.g., in [81].



(1) Power forecasting model 1.

(2) Power forecasting model 2.

(3) Power forecasting model 3.

(4) Ensemble model.

Figure 4.6: Figs. 4.6.1–4.6.3 show the local error distribution of a number of power forecasting models on the same weather forecasting model. The NWPs are projected onto the two most important principal components to better visualize the high-dimensional forecast. The local quality is indicated by the colors, where green means low error, red indicates high error. Using the proposed local soft gating technique (Section 4.4.2), a local soft gating for each algorithm is created. Using this form of weighting, the local quality can be increased, as shown by example of the training set in Fig. 4.6.4.

A simple yet effective technique for locality assessment is a nearest neighbor algorithm, a technique that is closely related to analog ensembles which we analyzed in [85]. In order to assess the local quality of a weather forecast $\mathbf{x}_{t+k|t}^{(\psi)}$, its C nearest neighbors are searched in $\mathbf{X}_{\mathrm{H}}$ in the way

$$\boldsymbol{\alpha}^{(\psi)} = \mathrm{knn}(\mathbf{x}_{t+k|t}^{(\psi)}, \mathbf{X}_{\mathrm{H}}, C)\,, \quad \boldsymbol{\alpha}^{(\psi)} \in \mathbb{N}^{C}, \tag{4.18}$$

where $\boldsymbol{\alpha}^{(\psi)}$ is a set containing the indices of the C nearest neighbors. In the simplest case, the metric for the knn algorithm is the Euclidean distance (in the feature space of the NWP forecast), though the use of more advanced distance metrics or feature importance scaling may further improve the local quality assessment. The average local quality can be assessed

using

$$\bar{S}_l^{(\varphi|\psi)} = \frac{1}{C} \sum_{\alpha \in \boldsymbol{\alpha}^{(\psi)}} S(\hat{y}_\alpha^{(\varphi|\psi)}, o_\alpha), \tag{4.19}$$

where S is the error of the scoring rule of Eq. 4.11 of the item at index α in the historic data set using the forecast $\hat{y}_\alpha^{(\varphi|\psi)}$ of ensemble member with power forecasting model φ computed on weather forecasting model ψ and the corresponding observation o_α. From this local error score $\bar{S}_l^{(\varphi|\psi)}$, the weight $w_l^{(\varphi|\psi)}$ is computed for each power forecasting model using coopetitive soft gating of Eq. 4.5 in the form

$$w_l^{(\varphi|\psi)} = \varsigma_\eta(\bar{\boldsymbol{S}}_l^{(\psi)}, \bar{S}_l^{(\varphi|\psi)}), \tag{4.20}$$

$$\bar{\boldsymbol{S}}_l^{(\psi)} = \left(\bar{S}_l^{(1|\psi)}, \ldots, \bar{S}_l^{(\varphi|\psi)}, \ldots, \bar{S}_l^{(\Phi|\psi)}\right), \tag{4.21}$$

where each value of $\bar{S}_l^{(\varphi|\psi)}$ is computed using Eq. 4.19. In the same fashion as for the global weighting (Eq. 4.15), the relative quality for each weather forecasting model ψ is estimated indirectly using all available power forecasting models in relation to other weather forecasting models, i.e.,

$$w_l^{(\psi)} = \varsigma_\eta(\bar{\mathbf{S}}_l, \bar{S}_l^{(\psi)}), \text{ where} \tag{4.22}$$

$$\bar{S}_l^{(\psi)} = \frac{1}{\Phi} \sum_{\varphi=1}^{\Phi} \bar{S}_l^{(\varphi|\psi)}, \tag{4.23}$$

$$\bar{\mathbf{S}}_l = \left(\bar{S}_l^{(1)}, \ldots, \bar{S}_l^{(\psi)}, \ldots, \bar{S}_l^{(\Psi)}\right). \tag{4.24}$$

Figs. 4.6.1 – 4.6.3 show the local quality of a number of power forecasting models in Voronoi diagrams, where green represents areas of low error and red color indicates areas with high error. The axes are given by the two most important principal components in order to better visualize the D-dimensional NWP feature space. Using the locality assessment technique of Eq. 4.20, each model is weighted depending on the position in the feature space in a way that reduces the overall error in the ensemble. Fig. 4.6.4 shows an example of the resulting ensemble error. It should be kept in mind that in the shown case, the improvement for the training data set is displayed. An example for the development over time of the local weights is shown in Section 4.4.5. This weighting term has to be computed during ensemble application for every NWP.

An advantage of the knn technique is that no model training is required (in the basic form if no feature subspace is selected or feature weighting is applied). However, as the data set $\mathbf{X}_\mathrm{H}$ serves as basis for the locality assessment, it has to be searched in every iteration, which usually does not scale optimally if no search heuristics are being employed. The knn approach is therefore particularly useful for smaller data sets. An alternative to knn is choosing the neighbors which are within a defined maximum distance (range search). In [81], we introduced a technique for locality assessment based on multi-linear interpolation which does require a training phase. However, during model application it no longer requires the data set $\mathbf{X}_\mathrm{H}$. This technique is therefore well-suited for larger data sets. Regarding ensemble forecasting quality, both approaches behave similarly.

4.4.3 Lead Time-Dependent Soft Gating

The lead time-dependent weighting components $w_k^{(\psi)}, w_k^{(\varphi|\psi)}$ show the quality dependence of a model for each lead time k. The idea is to weight models according to their lead time-dependent performance. In the area of power forecasting, a prominent example for approaches with time step dependent performance is the persistence method, which performs well on very short lead times only.

The idea is to create a weight per lead time k by evaluating the quality differences of a number of base predictors. For the creation of this form of weighting, a training data set for a *particular* lead time k – for which a number of N_k forecast-observation pairs are created using weather forecasting model ψ and power forecasting model φ – can be denoted as

$$S_k^{(\varphi|\psi)} = \sum_{n=1}^{N_k} S(\hat{y}_n^{(\varphi|\psi)}, o_n), \tag{4.25}$$

where N_k is the number of evaluated elements for the currently evaluated lead time k. The quality of the particular forecasting time step k in relation to other forecasting time steps of the same model can then be denoted as

$$\bar{S}_k^{(\varphi|\psi)} = \frac{S_k^{(\varphi|\psi)}}{\frac{1}{k_{\max}-k_{\min}+1}\sum_{k^*=k_{\min}}^{k_{\max}} S_{k^*}^{(\varphi|\psi)}}. \tag{4.26}$$

Then, the weighting factor $w_k^{(\varphi|\psi)}$ is computed for each forecasting time step using Eq. 4.5 in relation to other members of the ensemble in the form

$$w_k^{(\varphi|\psi)} = \varsigma_\eta(\bar{\mathbf{S}}_k^{(\psi)}, \bar{S}_k^{(\varphi|\psi)}), \text{ where} \tag{4.27}$$

$$\bar{\mathbf{S}}_k^{(\psi)} = \left(\bar{S}_k^{(1|\psi)}, \ldots, \bar{S}_k^{(\varphi|\psi)}, \ldots, \bar{S}_k^{(\Phi|\psi)}\right). \tag{4.28}$$

Again, the lead time-dependent weather forecasting model qualities are estimated using the overall power forecasting models and all weather forecasting models with

$$w_k^{(\psi)} = \varsigma_\eta(\bar{\mathbf{S}}_k, \bar{S}_k^{(\psi)}), \text{ where} \tag{4.29}$$

$$\bar{S}_k^{(\psi)} = \frac{1}{\Phi}\sum_{\varphi=1}^{\Phi} \bar{S}_k^{(\varphi|\psi)}, \tag{4.30}$$

$$\bar{\mathbf{S}}_k = \left(\bar{S}_k^{(1)}, \ldots, \bar{S}_k^{(\psi)}, \ldots, \bar{S}_k^{(\Psi)}\right). \tag{4.31}$$

In case there are few data available for the training process, a smoothing over weights in neighbored lead times can be applied in order to avoid noisy weights. An example of the effect of this form of weighting is described in Section 4.4.5. This weighting term is computed once during ensemble training (see Section 4.4.4) for every lead time and can then be reused for every forecast.

4.4.4 Model Fusion and Ensemble Training

As stated above, the overall weighting $w^{(j)}$ of each ensemble member j is computed using Eq. 4.8. The main parameter of the *coopetitive soft gating ensemble* (CSGE) algorithm is the hyperparameter η of the coopetitive soft gating formula. Depending on the forecasting task and the data set, the appropriate value of η may differ. Furthermore, the value of η for each

weighting aspect (global weighting, local soft gating, lead time-dependent soft gating) may vary. The value of η therefore should be chosen independently for each weighting aspect.

Assuming the same number and types of power forecasting models for each weather forecasting model, the weighting parameter can be treated identically for each weather forecasting model. Therefore, the number of optimization parameters is six (an individual η parameter for global, local, and lead time-dependent weighting for both, weather forecasting model and power forecasting model).

The tuple of coopetitive soft gating parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_z, \ldots, \eta_Z)$ with $Z = 6$ can then be optimized using an appropriate optimization algorithm (e.g., a greedy parameter optimization using a simplex method [174] or a gradient-based interior point algorithm [31]) solving the minimization problem

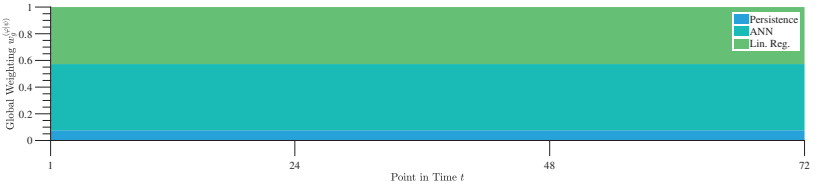
$$\begin{aligned} \underset{\boldsymbol{\eta}}{\text{minimize}} \quad & \frac{1}{N}\sum_{n=1}^{N} S(\bar{y}_n^{(\boldsymbol{\eta})}, o_n) + \zeta \cdot \sum_{z=1}^{Z} \eta_z, \\ \text{where} \quad & \bar{y}_n^{(\boldsymbol{\eta})} = \sum_{j=1}^{J} w_{\boldsymbol{\eta}}^{(j)} \cdot \hat{y}_n^{(j)}, \\ \text{subject to} \quad & \text{each } \eta_z \geq 0, \end{aligned} \tag{4.32}$$

where $\bar{y}_n^{(\boldsymbol{\eta})}$ is the forecast of the overall CSGE forecasting function, N are the overall evaluated points of a validation data set, $w_{\boldsymbol{\eta}}^{(j)}$ are the weights of a particular forecast computed using Eq. 4.8 using the hyperparameters in $\boldsymbol{\eta}$, and $\zeta \geq 0$ is a regularization parameter.

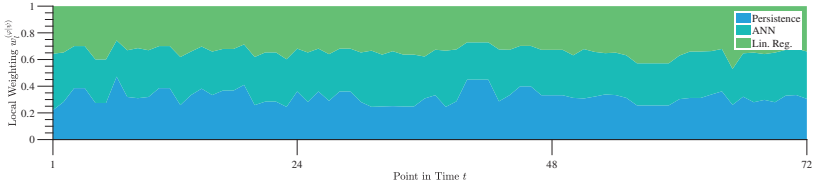
An advantage of the proposed technique is that it is a post-processing technique. Therefore, while the single forecasts of each ensemble member j are weighted differently using $w_{\boldsymbol{\eta}}^{(j)}$, the forecasts $\hat{y}_n^{(j)}$ of each of the J ensemble members remain constant, no matter what the value of $\boldsymbol{\eta}$ may be. The values $\hat{y}_n^{(j)}$ therefore only have to be computed once for the evaluated data set during ensemble training which speeds up the optimization process. Furthermore, the single weights change smoothly when varying the parameters in $\boldsymbol{\eta}$. Consequently, we end up with a smooth (continuously differentiable) optimization function.

4.4.5 Application Examples of the CSGE Technique

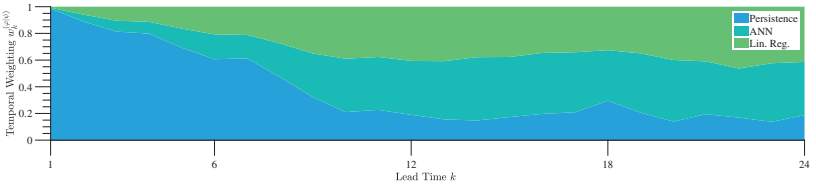
This section describes two application examples of the final CSGE algorithm. The first example shows the application of the CSGE algorithm for intraday forecasting ($k_{\text{min}} = 1$ h, $k_{\text{max}} = 24$ h, $\Delta = 1$ h) using a single weather forecasting model and an ensemble of three forecasting algorithms, namely an ANN, a linear regression, and a persistence forecast. Fig. 4.7 shows the single weighting aspects and the overall weights over time. As there exists just a single weather forecasting model, the number of weighting aspects is reduced to three. The global weights are shown in Fig. 4.7.1. The algorithm weights the single algorithms according to their expected quality (ANN best, persistence worst). These weights remain constant over time. The local weights are detailed in Fig. 4.7.2. In this particular case, all local weights are similar. The lead time-dependent weights are shown in Fig. 4.7.3. Note the different horizontal axis, which is the lead time in this case. As is to be expected, the persistence method works well on very short time horizons, but quickly loses quality in comparison to the other two approaches. The combination of the three weighting aspects is visualized in Fig. 4.7.4. As can be seen, on delivery of new NWP forecasts every 24 h, the influence of the lead time-dependent persistence technique is high. An overall forecast is shown in Fig. 4.8. It can be seen that both ANN and Lin. Reg. create different forecasts for each point in time



(1) Development of global weighting $w_g^{(\varphi|\psi)}$ over time. ANN performs best, followed by the linear regression. The global weighting of the persistence method is small.



(2) Development of local weighting $w_l^{(\varphi|\psi)}$ over time. As the particular weather situation is a fairly standard situation (and thus, it is relatively well represented in the data set), there is little observable local weighting in this particular case.



(3) Development of lead time-dependent weighting $w_k^{(\varphi|\psi)}$ over time. Note the different time-scale (24 h). Though the global performance $w_g^{(\varphi|\psi)}$ of the persistence method is low, it performs strongly on short time horizons.

(4) Combined $w^{(\varphi|\psi)}$, consisting of $w_g^{(\varphi|\psi)}$, $w_k^{(\varphi|\psi)}$, and $w_l^{(\varphi|\psi)}$, over time. As can be seen in this example, though the overall quality of the persistence method is low, it has high impact in short time horizons.

Figure 4.7: Example of the weight combination of the proposed CSGE ensemble technique. In the example, an intraday forecast is performed using a single weather forecasting model and three power forecasting models. New weather data are incorporated every 24 h. For the intraday forecast, a persistence method is combined with an ANN and a linear regression model. During ensemble training, the coopetitive soft gating parameters are optimized so that the depicted weighting emerges for an example weather situation over three days. The overall weight development $w^{(\psi,\varphi)}$ (Fig. 4.7.4) is composed of the three weighting terms $w_g^{(\varphi|\psi)}$ (Fig. 4.7.1), $w_l^{(\varphi|\psi)}$ (Fig. 4.7.2), and $w_k^{(\varphi|\psi)}$ (Fig 4.7.3). The overall forecast is shown in Fig. 4.8.

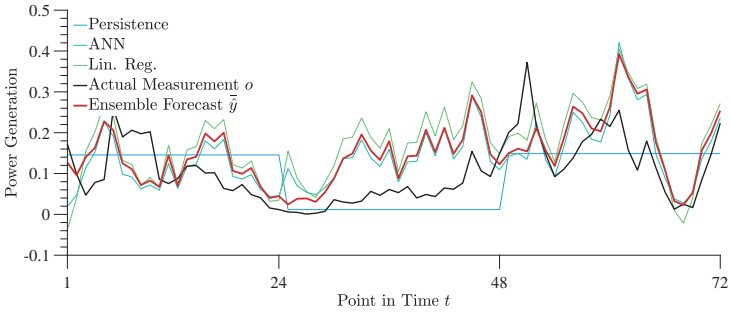


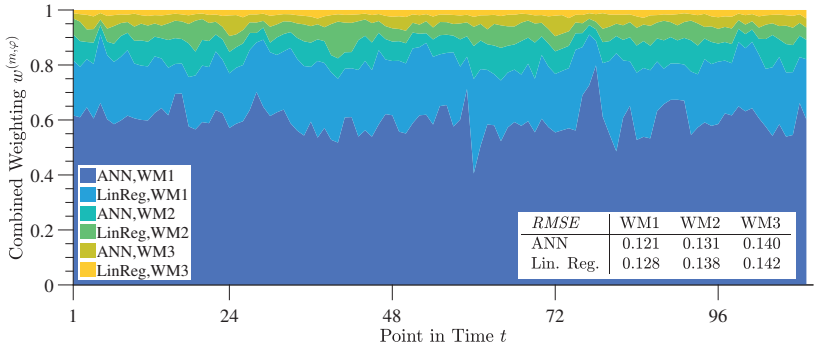
Figure 4.8: Forecasting example using combined weighting of the CSGE technique. On delivery of a new NWP each 24 h, the weighting of the persistence method is high in this example. The weights are computed as given in Fig. 4.7.

while the persistence method is updated every 24 h. As can in particular be seen for point in time $t = 25$, the ensemble method is able to improve the forecasting quality by including a persistence forecast.

The second example shows a multi-model forecast using three weather forecasting models for a day-ahead forecast ($k_{\min} = 25$ h, $k_{\max} = 48$ h, $\Delta = 1$ h), each of which is predicted using two power forecasting models (ANN and Lin. Reg.). This example is visualized in Fig. 4.9. The overall weighting over time is shown in Fig. 4.9.1. Regarding the weather forecasting models, the first weather forecasting model "WM1" (which is the ECMWF IFS model) has the highest influence, while the other two weather forecasting models have lower weights. This, again, meets the expectation, as can be seen from the table in Fig. 4.9.1, which shows the overall RMSE when performing the forecast on a single weather / power forecasting model combination. The model with the lowest RMSE error gets the highest weight. Also, the quality difference between the power forecasting models is reflected in the weighting. As there are no weather or power forecasting models which are designed for a different forecasting time period (unlike in the first example), the overall weighting differences over time are not as drastic as in the first example. The forecast which is created using the weights determined by the CSGE is shown in Fig. 4.9.2. As can be seen in the figure, a weighted ensemble forecast is created. In accordance with the combined weights $w^{(\varphi,\psi)}$, the ANN of WM1 has the highest impact, the ensemble forecast thus is very close to the forecast of this power forecasting model. Regarding the actual differences in the individual forecasts, the effects of the weather model appear more drastic than the differences induced from using different power forecasting models.

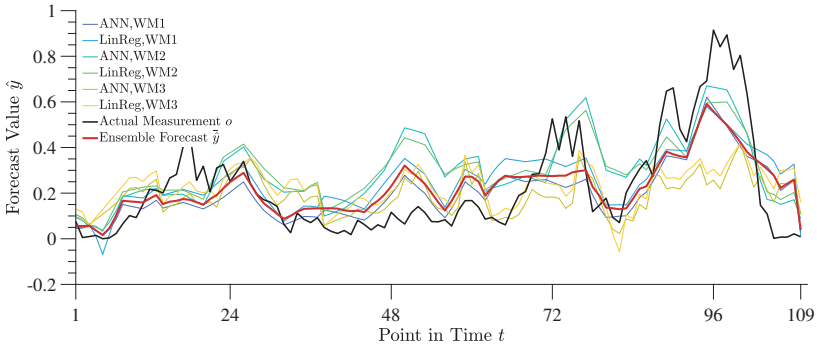
4.5 Experimental Results

This section investigates the performance of the proposed CSGE technique in comparison to a number of state of the art approaches. We evaluate the algorithms on 45 data sets which are described in Section 4.5.1. The experimental setup is described in Section 4.5.2. We examine the proposed power forecasting model using a single weather forecasting model and multiple weather forecasting models for day-ahead forecasting (Section 4.5.3), as well as for intraday forecasts (Section 4.5.4). Finally, a detailed comparison of the performance gains when using



RMSE	WM1	WM2	WM3
ANN	0.121	0.131	0.140
Lin. Reg.	0.128	0.138	0.142

(1) Multi-model CSGE consisting of 3 weather forecasting models, each predicted using 2 power forecasting models.



(2) Multi-model CSGE forecast with a total of $J = 6$ ensemble members.

Figure 4.9: Example of weighting combination for a day-ahead forecast using three weather forecasting models (multi-model ensemble) and the proposed CSGE technique. Each model has different forecasting quality (where it can be observed that WM1 has the highest quality and WM3 performs worst). The overall weighting for each weather / power forecasting model combination is shown in Fig. 4.9.1. The result is in line with the expectation when performing the prediction using a single weather and power forecasting model combination, as can be seen from the table in the figure. Values denote RMSE on normalized power values. The overall forecast is shown in Fig. 4.9.2. Depending on the particular situation, the ensemble adapts the model weights.

multiple weather forecasting models for both, day-ahead and intraday forecasts, is detailed in Section 4.5.5.

4.5.1 Data Set Used for Evaluation

The data sets used for the evaluation are derived from the publicly available data set collection of the *EuropeWindFarm* collection [76]. The 37 utilized data sets contain the weather forecasts of four weather models and power measurements of 22 consecutive months of both onshore and offshore wind farms. They and include the following meteorological variables:

- *Time Stamp* of the forecast / power measurement,
- *Lead time* from the forecasting origin,
- *Wind Speed* in 100 m height,
- *Wind Speed* in 10 m height,
- *Wind Direction (zonal)* in 100 m height,
- *Wind Direction (meridional)* in 100 m height,
- *Air Pressure* forecast,
- *Air Temperature* forecast, and
- *Power Generation* of the wind farm.

The power generation time series are normalized with respect to the nominal capacity o_{inst} of each power plant to enable a scale-free comparison in the range $[0,1]$. All weather input parameters are normalized in the interval $[0,1]$. The data has been filtered to eliminate erroneous measurements. Fig. 4.10 gives an overview of the locations of the wind farms.

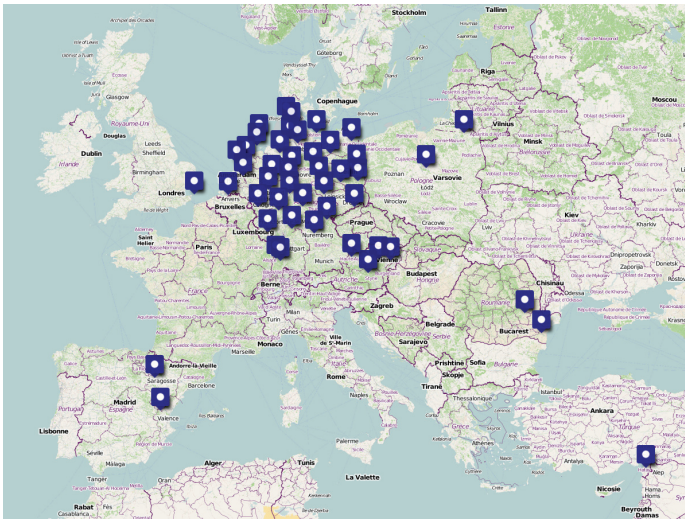


Figure 4.10: Locations of the wind farms of the *EuropeWindFarm* collection, see Section 4.5.1 for a data set description.

As the four different weather models are available for the locations of the wind power plants for the same time period, they can be used to form a multi-model ensemble. Each

weather model has the same features as the ones listed above. The weather models shown in Table 4.6 are included.

Table 4.6: Weather models in the data sets.

Weather Model Name	$k_{\min}$	$k_{\max}$	Δ
ECMWF IFS (week)	25 h	48 h	1 h
GFS (week)	25 h	48 h	1 h
COSMO EU day-ahead	25 h	48 h	1 h
COSMO DE intraday	1 h	24 h	1 h

4.5.2 Experimental Setup

In the experiments, we evaluate the power forecasting performance for the data sets (see Section 4.5.1). As laid out in the CSGE algorithm description of Section 4.4, for each weather forecasting model we perform the forecast using multiple forecasting algorithms. As optimization for the CSGE technique (see Section 4.4.4), a gradient-based interior point algorithm [31] is chosen. The regularization parameter ζ is chosen in a way that avoids overfitting of the CSGE algorithm to the validation data.

A nearest neighbor range search is used for the local weight determination with NWP features being weighted according to a feature importance weighting prior to the nearest neighbor search. This process is conducted in the following way:

- The features of each weather model contain seven features which have been selected by domain experts for wind power forecasting. All present features can thus be assumed to be relevant, consequently no feature *selection* has to be performed. However, for the nearest neighbor search, a feature *weighting* may improve the quality of the local weighting.

- A computationally inexpensive filter approach for filter weighting is chosen. As has been pointed out in [141, 220], the Relieff algorithm [140] can be used for creating feature weights for a nearest neighbor search. After computing Relieff, the resulting features weights created by Relieff (which in other applications can be used for feature ranking) are used in this case for feature scaling to achieve the effect of a Mahalanobis distance computation (assuming no covariance).

- A simple nearest neighbor range search (with Euclidean distance as distance metric performed on the scaled features) is used to find the nearest neighbors. For determining the range parameter of the range search, prior to the ensemble computation, the training data set is split into a training and test data set and the range parameter is chosen in a way that minimizes the CRPS score. The CRPS is a score that is suited to compare a distribution of values with an observation, more details are given in Section 6.1.1 and [108]. This is performed by using the identified neighbors as basis for the computation of an analog ensemble (see Section 5.6.3).

- The weather situations identified as neighbors for a certain query NWP are then used for the computation of the local weighting based on Eq. 4.18.

For the evaluation, each data set is split into a training and a test subset in a 5-fold cross-validation with a training data set that includes (4/5) of the data and a test data set (1/5).

The results presented in the case studies below thus are the results regarding 5 repetitions with each data set. When using the CSGE technique, the training data set is further split into three sets of equal size which are called *parameter set* (1/3), *optimization set* (1/3) and *validation set* (1/3) for the sake of clarity. The single power forecasting models for each weather forecasting model are trained using the parameter set. The parameter optimization of the CSGE technique is then performed on the optimization set (that serves, e.g., as historic data set for the local soft gating) and is finally optimized regarding $\boldsymbol{\eta}$ with the validation set. The parameter combination which performed best on the validation set is chosen as final model parameterization which is used to compute the final model quality on the test set.

As pointed out, e.g., in [86] and Section 3, error scores beyond the RMSE are of importance for investigating a forecast. As all evaluated models are investigated on the same data sets, the abstraction capability of error scores does not significantly increase the understanding of the behavior of the investigated forecasting models in the conducted experiments. For our evaluation, we therefore give the error distribution of the RMSE (computed using Eq. 3.6), the skill score for model comparison, and the standard deviation of the errors. To recall, the skill score describes the amount of improvement of an evaluated technique by taking an error score $\bar{S}_{\text{eval}}$ in comparison to the error score of a baseline technique $\bar{S}_{\text{base}}$. The improvement of the forecast can be computed on an arbitrary error score with

$$\text{Skill} = \frac{\bar{S}_{\text{base}} - \bar{S}_{\text{eval}}}{\bar{S}_{\text{base}}}. \tag{4.33}$$

4.5.3 Case Study: Day-Ahead Performance on Single and Multiple Weather Forecasting Models

This set of experiments aims at comparing the forecasting performance regarding the power generation of the CSGE technique in comparison to other power forecasting models (both ensembles and non-ensemble techniques). In the experiment, we use the three day-ahead weather forecasting models for all 37 wind farms. For the comparison, we included the following techniques in the experiments:

1. LR: A simple linear regression model that is computed with all features and with the single best weather forecasting model (determined using the table in Fig. 4.9.1). This model also serves as baseline for the skill score computation.
2. SVR: A support vector regression model with polynomial kernel that is used with the single best weather forecasting model (determined using the table in Fig. 4.9.1).
3. ANN: An artificial neural network model (feed forward) used with the single best weather forecasting model (determined using the table in Fig. 4.9.1).
4. LSBoost: An ensemble technique with regression trees as base predictors. The ensemble is based on the boosting principle [70] applied for least squares. The ensemble is computed with the single best weather forecasting model (determined using the table in Fig. 4.9.1).
5. Bagging: An ensemble technique with regression trees as base predictors. The bagging ensemble [19] is computed with the single best weather forecasting model (determined using the table in Fig. 4.9.1).
6. MME Eq.: The model is computed using the best non-ensemble forecasting technique (ANN model) with all weather forecasting models. The model forecasts are averaged, i.e., $w^{(j)} = \frac{1}{J}$. This technique, is, e.g., utilized in [104].

7. MME We.: The model is computed using the best non-ensemble forecasting technique (ANN model) with all weather forecasting models. The models are weighted with respect to their global quality using Eq. 4.15 with $\eta = 2$, such that the best performing weather forecasting models get the highest impact in the weighting. This technique is employed, for instance, in [200].
8. CSGE (V1): The CSGE technique is applied using *multiple* power forecasting models for the single best weather forecasting model. As there is only one weather forecasting model, the number of weighting dimensions is reduced to the three power forecasting model based weighting factors $w^{(\varphi|\psi)} = w_g^{(\varphi|\psi)} \cdot w_l^{(\varphi|\psi)} \cdot w_k^{(\varphi|\psi)}$ for each power forecasting model. The CSGE locality assessment is performed using a range search as laid out in Section 4.5.2.
9. CSGE (V2): The CSGE technique is applied using a *single* power forecasting model for each of the three weather forecasting models, the best non-ensemble forecasting technique (ANN model). As there is only one power forecasting model, the number of weighting dimensions is reduced to the three weather forecasting model based weighting factors $w^{(\psi)} = w_g^{(\psi)} \cdot w_l^{(\psi)} \cdot w_k^{(\psi)}$ for each weather forecasting model. The CSGE locality assessment is performed using a range search as laid out in Section 4.5.2.
10. CSGE (V3): The CSGE technique using *multiple* power forecasting models for each weather forecasting model is applied. The CSGE technique consequently uses all six weighting factors. The CSGE locality assessment is performed using a range search as laid out in Section 4.5.2.

Table 4.7: Performance comparison regarding the distribution of RMSE scores, the mean error, the standard deviation of errors, and the skill score of the experiments of Section 4.5.3. The color coding indicates the quality of each wind farm and power forecasting algorithm from high quality (green) to low quality (red). Bold text highlights the best achieved score for a single and multiple weather models (WM) individually.

	Single WM						**Multiple WM**			
	LR	SVR	ANN	LSBoost	Bagging	CSGE (V1)	MME Eq.	MME We.	CSGE (V2)	CSGE (V3)
Max	0.186	0.256	0.182	**0.178**	0.183	0.186	0.181	0.181	0.202	0.183
90% Quant.	0.168	0.170	**0.156**	**0.156**	0.157	**0.156**	0.156	0.154	0.154	**0.152**
Median	0.132	0.127	0.117	0.121	0.118	**0.116**	0.117	0.115	**0.114**	**0.114**
10% Quant.	0.093	0.094	**0.082**	0.086	0.084	0.084	0.083	**0.082**	**0.082**	**0.082**
Min	**0.046**	0.069	0.048	0.050	0.052	0.056	0.051	0.051	0.053	**0.042**
Mean	0.130	0.131	0.118	0.120	0.118	**0.117**	0.117	0.116	0.116	**0.114**
Std. Dev.	0.028	0.030	**0.027**	**0.027**	**0.027**	**0.027**	0.027	0.027	0.027	**0.026**
Skill	0.0%	-0.3%	9.7%	7.7%	9.2%	**10.2%**	10.2%	10.8%	11.1%	**12.6%**

The results of the day-ahead experiments are shown in Table 4.7. In total, the metrics are computed from a 5-fold cross-validation for each data set, leading to 185 experiments in total for each forecasting technique. Regarding the maximum error, the LSBoost algorithm performs best, however, many algorithms, such as CSGE and the multi-model ensembles, perform roughly on par. Surprisingly, the CSGE (V2) technique performs weak in this regard. Regarding the 90 % quantile and the median, the CSGE techniques yield the best results. The ANN technique yields strong results regarding the 90 % and the 10 % quantile as well as the minimum error. However, the multi-model techniques and in particular the CSGE (V3) technique outperform ANN regarding the median and the 10% quantile. The mean error gives similar indication as the median, where both CSGE (V1) and CSGE (V3) perform strongest.

This leads to a skill score (computed on mean error) of 10.2 % and 20.4 % in comparison to the linear regression technique that serves as the baseline technique in this experiment. Regarding the standard deviation of error scores, many of the evaluated techniques yield very similar results with a value of 0.027. Exceptions are the LR and the SVR technique, that yield higher values of the standard deviation, and the CSGE (V3) technique, that yields lower values with a value of only 0.026.

As can be seen from the performance of the examined ensemble techniques, the use of multiple weather forecasting models clearly improves the performance, e.g., in MME Eq. in comparison to a single ANN. The use of a global weighting (such as conducted in MME We.) further increases the performance. The use of local and temporal weighting in addition to the global weighting (as performed in CSGE (V2)) increases the performance, even if just one power forecasting model is used. The scores can further be improved when performing the weighting for both, weather and power forecasting models (as performed in CSGE (V3)).

The ranked performance of the algorithms among all of the 185 experiments is furthermore analyzed using the Friedman test in conjunction with the Nemenyi post-hoc test that is described in Section 2.8.3. The results are given in Fig. 4.11. The Friedman p value given in the figure indicates that the ranks are significantly different. As can be seen from the Nemenyi test, the CSGE (V3) technique has a significantly better ranked performance in comparison to the other techniques as it exceeds the difference in ranked performance of CD = 1. The CSGE (V2) technique, however, does not have a significant difference from the weighted ensemble technique. In general, all multi-model techniques except for the averaged ensemble technique have a higher ranked performance than the models that are only based on a single weather model.

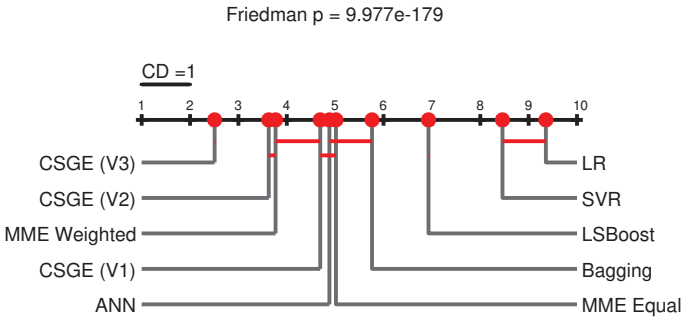


Figure 4.11: Evaluation of ranked performance using the Nemenyi post-hoc test. The algorithms have to exceed the value of the critical distance (CD) in order to be assumed to be statistically different. As can be seen from the figure, the CSGE (V3) technique has a significantly better ranked performance than the other comparison techniques. This can, however, not be stated for the CSGE (V2) technique which is not significantly different from the weighted ensemble technique.

4.5.4 Case Study: Intraday Performance on Single and Multiple Weather Forecasting Models

This set of experiment evaluates the ability of the proposed method to work as a multi-model ensemble for intraday forecasting. Therefore, an intraday weather forecasting model is added to the three existing weather forecasting models, and an *intraday* forecast is performed ($k_{\min} = 1$ h, $k_{\max} = 24$ h, $\Delta = 1$ h). The overall forecast therefore is created from one intraday forecast and three day-ahead forecasts. Bear in mind that the additional weather models are day-ahead weather models used as intraday forecasts, and, thus, typically perform worse compared to a true intraday model. However, improvements still may be achieved when using these models due to error canceling effects and adaptive weighting within the ensemble. As comparison techniques, we include the techniques used in Section 4.5.3 (but now based on the intraday weather forecasting model instead of the best day-ahead weather forecasting model) and additionally the following two techniques:

1. Pers.: The forecasting model is a persistence forecast. This is a very simple technique that just takes the most recent power measurement and assumes no change of the power generation in the future. This technique is used as baseline technique for intraday forecasting.
2. CSGE (V3+P): This technique is an extension of the CSGE (V3) technique that additionally includes the persistence technique as base predictor.

Table 4.8: Performance comparison regarding the distribution of RMSE scores, the mean error, the standard deviation of errors, and the skill score of the experiments of Section 4.5.4. The color coding indicates the quality of each wind farm and power forecasting algorithm from high quality (green) to low quality (red). Bold text highlights the best achieved score for a single and multiple weather models (WM) individually.

	Single WM							**Multiple WM**				
	Pers.	LR	SVR	ANN	LS Boost	Bagging	CSGE (V1)	MME Eq.	MME We.	CSGE (V2)	CSGE (V3)	CSGE (V3+P)
Max	0.294	0.184	0.195	**0.175**	0.176	0.179	0.183	0.181	0.180	0.191	0.179	**0.159**
90% Quant.	0.258	0.158	0.160	0.146	0.148	**0.144**	**0.144**	0.143	0.141	0.140	0.138	**0.136**
Median	0.204	0.124	0.121	0.110	0.113	**0.109**	0.110	0.105	0.105	0.105	**0.104**	**0.104**
10% Quant.	0.133	0.092	0.089	0.079	0.081	0.079	**0.078**	0.077	0.077	0.076	0.075	**0.074**
Min	0.045	0.044	0.057	0.050	0.048	0.045	**0.043**	0.051	0.052	0.050	0.041	**0.040**
Mean	0.198	0.124	0.122	0.111	0.113	**0.110**	**0.110**	0.108	0.106	0.106	0.105	**0.102**
Std. Dev.	0.050	0.027	0.027	0.026	0.026	**0.025**	**0.025**	0.025	0.025	0.025	**0.024**	**0.024**
Skill	0.0%	37.3%	38.4%	43.9%	43.0%	44.2%	**44.6%**	45.6%	46.5%	46.5%	47.2%	**48.3%**

The results of the experiment are shown in Table 4.8. As can be seen from the table, ANN is the best non-ensemble forecasting technique with the lowest error rates throughout the error distribution (except for the minimum error) and even the lowest maximum error of all power forecasting models that only use the intraday weather forecasting model. According to the expectations, the persistence method performs weakest. Within the ensemble techniques based on a single weather forecasting model, bagging and the CSGE (V1) technique perform best, where bagging exceeds the performance of CSGE (V1) with regard to the median error and CSGE (V1) performs stronger regarding the 10 % quantile and the minimum error. With respect to to the mean error and the standard deviation, bagging and CSGE (V1) again perform strongest, which leads to an improvement of 44.2 % and 44.6 % improvement over the persistence technique, respectively.

When including additional day-ahead weather forecasting models, all multi-model techniques are able to outperform all techniques based on a single weather forecasting model. Within this group of models, MME Eq. performs weakest, followed by MME We. and CSGE (V2). CSGE (V3) exceeds the performance of the other techniques in all evaluated metrics. The inclusion of the persistence method as power forecasting model in CSGE (V3+P) is further able to enhance the quality of CSGE (V3). These models therefore are able to exceed the quality of the pure persistence method from 45.6 % (MME Eq.) up to 48.3 %, CSGE (V3+P).

Fig. 4.12 again shows the ranked performance among all 185 experiments for each forecasting model. The CSGE variants therein have the best ranked performance, where the CSGE (V3+P) technique yields the best rank. The other multi-model approaches yield the next best results. Of the models that only use a single weather model, the CSGE (V1) technique performs best. The ANN is the best non-ensemble technique. The persistence method has the highest average rank. As can be seen from the figure, the Friedman p value indicates that the ranked performance of the models are different. As more models than in the experiment of Section 4.5.3 are included, the value of the critical distance also is larger. The CSGE (V3+P) technique is significantly different from the other techniques. The CSGE (V2), CSGE (V3) and the weighed multi-model ensemble form a block of ranks that is not significantly different.

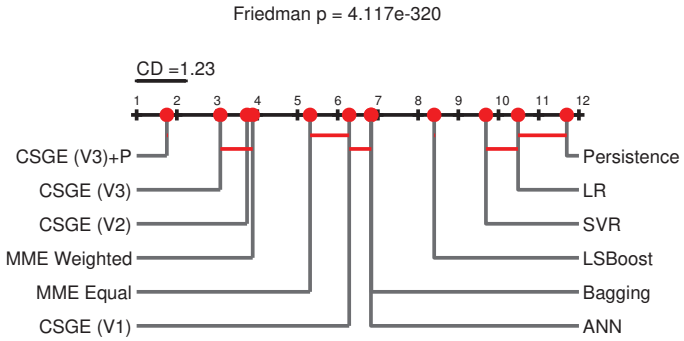


Figure 4.12: Evaluation of ranked performance using the Nemenyi post-hoc test. The algorithms have to exceed the value of the critical distance (CD) in order to be assumed to be statistically different. As can be seen from the figure, the CSGE (V3+P) technique has the best ranked performance and is significantly different from the other comparison techniques. This can, however, not be stated for the CSGE (V2) and CSGE (V3) technique which are not significantly different from the weighted multi-model ensemble.

In comparison to the day-ahead forecasts, the intraday forecasting models yield more accurate results, which is detailed in Table 4.9. An average improvement of 7.4 % regarding the mean error can be observed, with the lowest improvement can be observed for the LR technique with 4.9 % and the highest improvement for the CSGE (V3+P) technique with 10.2 %. The multi-model techniques therein are able to benefit more from the inclusion of this additional model in comparison to the models that only use a single WM.

4.5.5 Case Study: Performance Development Using a Varying Number of Weather Forecasting Models

This set of experiments gives insight into a practical problem in power forecasting: Given a number of weather forecasting models available, which of those models should be included,

Table 4.9: Performance comparison of a day-ahead and an intraday forecast regarding the improvement of the mean value of the RMSE.

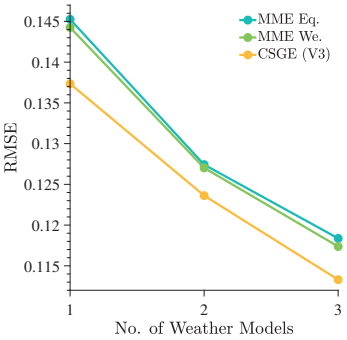
	Single WM						Multiple WM				
	LR	SVR	ANN	LS Boost	Bagging	CSGE (V1)	MME Eq.	MME We.	CSGE (V2)	CSGE (V3)	CSGE (V3+P)
Day-Ahead	0.130	0.131	0.118	0.120	0.118	0.117	0.117	0.116	0.116	0.114	0.114[7]
Intraday	0.124	0.122	0.111	0.113	0.110	0.110	0.108	0.106	0.106	0.105	0.102
Skill	4.9 %	6.9 %	5.8 %	6.3 %	6.8 %	6.4 %	8.1 %	9.0 %	8.8 %	8.3 %	10.2 %

[7] As the CSGE (V3+P) does not exist for day-ahead forecasts, the CSGE (V3) technique is used for comparison.

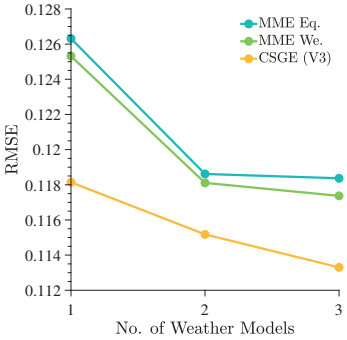
and will the overall quality eventually even be lowered when worse performing models or models with unknown performance are added to the overall ensemble. This section therefore describes the dependence of the performance of power forecasting models when adding additional weather forecasting models to the forecast. In order to evaluate the performance, the mean RMSE of the techniques MME Eq., MME We., and CSGE (V3) are computed with a varying number of weather forecasting models. Figs. 4.13.1 and 4.13.3 show the dependence of the performance when the weakest model is chosen as first model, and increasingly good weather forecasting models are added. Figs. 4.13.2 and 4.13.4, on the other hand, show the development of the performance when the best weather forecasting model is chosen first and worse models are added subsequently. For this experiment, the data is evaluated using a split validation to divide the training data set (3/4) and test data set (1/4).

For the day-ahead variants of Figs. 4.13.1, and 4.13.2, all ensemble models benefit from increasing the number of weather forecasting models, even if the additional models perform worse in comparison to the first weather forecasting model. Fig. 4.13.1 shows the RMSE error when including increasingly good performing models. As can be expected, the performance is clearly higher when adding additional weather forecasting models. As can be seen from Fig. 4.13.2, the performance improvement drops when including a third weather forecasting model and the added model performs worse than the already included models, in this case. The approaches for comparison benefit more when including a second model. However, the CSGE (V3) is able to benefit more from the inclusion of a third weather forecasting model. In particular the MME Eq. technique is not able to regulate the model influence precisely enough and therefore barely benefits from the inclusion of a third model.

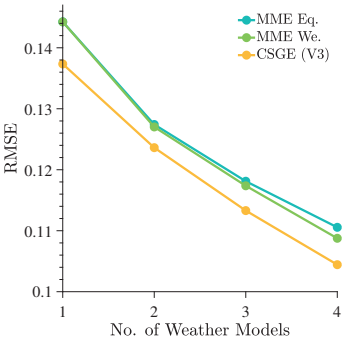
For intraday forecasts, the behavior is similar to the day-ahead forecast when adding increasingly well performing models, as is indicated in Fig. 4.13.3. When including an intraday model as fourth model in addition to the three day-ahead forecasts, all models benefit about equally from the inclusion of the fourth model. An interesting phenomenon is visible in Fig. 4.13.4, i.e., when adding additional day-ahead weather forecasting models to an intraday weather forecasting model. All models benefit from the inclusion of the best-performing day-ahead weather forecasting model. The maximum benefit of the inclusion of multiple weather forecasting models seems to be reached at this point, as no model is able to benefit from additional weather forecasting models. However, when including further weather forecasting models, the model performance of the comparison approaches *decreases*, i.e., the inclusion of additional models irritates the ensemble forecast. The CSGE (V3) technique, on the other hand, is able to recognize the weak performing weather forecasting models and is able to reduce the respective weights in a way that does not negatively affect the overall model performance.



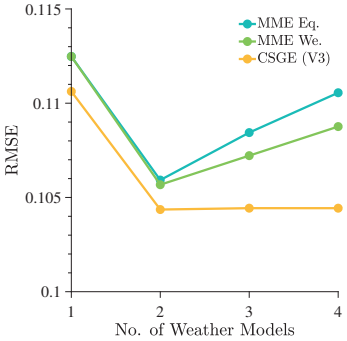
(1) Day-ahead forecast when adding increasingly well performing weather forecasting models.



(2) Day-ahead forecast when adding increasingly weak performing weather forecasting models.



(3) Intraday forecast when adding increasingly well performing weather forecasting models.



(4) Intraday forecast when adding increasingly weak performing weather forecasting models.

Figure 4.13: Average development of the RMSE quality of a number of ensemble techniques when including additional weather forecasting models to the ensemble for both, day-ahead and intraday forecasts, using the EuropeWindFarm data set collection. When adding increasingly well performing weather forecasting models, all ensemble algorithms benefit from the inclusion of additional weather forecasting models (Figs. 4.13.1 and 4.13.3). When adding increasingly weak performing models to the ensemble, the CSGE (V3) technique is able to better use this additional information. The comparison techniques' performance, on the other hand, remains about constant (see Fig. 4.13.2), or even decreases (see Fig. 4.13.4).

4.6 Properties of the CSGE Model

Beside the performance of the proposed approach, there are some other model properties worth mentioning.

Failure Mode

The ensemble weights $w^{(j)}$ are computed dynamically for each forecasting time step for which a forecast is performed. The CSGE algorithm computes the weights for the respective weighting categories using all available forecasts *at that particular time*. Thereby, the ensemble can create a forecast even if some power forecasting models fail to create a prediction, or if the NWP forecast for a particular weather forecasting model fails to be delivered. The weighting of Eq. 4.8 is then simply computed using only the forecasting models that are available at that particular point in time. The proposed technique is then able to retain the optimal weighting performance given the circumstances.

Run-time

The training process of the ensemble consists of the training of the base predictors for each combination of PM and WM and the ensemble training process. The overall run-time for training can therefore be denoted as

$$r^{(\text{all})} = r^{(\text{CSGE})} + \sum_{\psi=1}^{\Psi} \sum_{\varphi=1}^{\Phi} r^{(\varphi|\psi)}, \tag{4.34}$$

where the overall run-time for the training process $r^{(\text{all})}$ consists of the run-time for each base predictor $r^{(\varphi|\psi)}$ and the run-time for ensemble training $r^{(\text{CSGE})}$. The run-time of the ensemble training itself again depends on the number of WM and PM, the optimization algorithm, and a training step for the locality assessment (if a technique that requires a training procedure is utilized).

For model application after training, the computation time consists of the forecast creation time of the base predictors and the aggregation within the ensemble technique. The run-time complexity depends on the overall number of weather forecasting models Ψ and power forecasting models Φ. For each ensemble member the weight is computed using Eq. 4.8, which in turn is composed of Eqs. 4.9 and 4.10. For Eq. 4.9, the CSG weighting function is computed three times each with complexity $O(\Psi)$. The computation of Eq. 4.10 is similar as it evaluates the CSG weighting function three times each with complexity $O(\Phi)$. The combined complexity of Eq. 4.8 can thus be denoted as $O(\Psi \cdot \Phi)$. Furthermore, Eq. 4.8 is evaluated for each of the $J = \Psi \cdot \Phi$ ensemble members. The overall complexity of the CSGE technique therefore is

$$O\left(\Psi^2 \cdot \Phi^2\right). \tag{4.35}$$

As the denominator in Eq. 4.8 is composed of the nominators for each of the $\Psi \cdot \Phi$ ensemble members, it is an *additive* term with complexity $O(\Psi \cdot \Phi)$ that therefore does not influence the overall run-time complexity. The final weighted aggregation of Eq. 4.1 also is an additive term with complexity $O(\Psi \cdot \Phi)$.

An evaluation of the actual run-time of the CSGE weighting function can be found in Table 4.10. The table shows the run-time of the computation of the coopetitive soft gating function of Eq. 4.5 depending on the number of error measures in $\mathbf{\Omega}$ and the value of η. The

Table 4.10: Run-time of the coopetitive soft gating function.

in μs		**Value of η**					
		0	1	3	5	10	50
No. Base Predictors	1	65.76	65.24	65.60	66.23	65.53	65.69
	2	76.00	76.28	77.55	77.39	77.07	78.04
	3	76.10	76.60	77.35	77.58	77.30	78.20
	5	76.25	76.53	77.64	77.57	77.70	78.37
	10	76.11	76.48	78.41	78.47	78.52	78.94
	50	77.60	77.61	82.89	82.92	83.22	83.49
	100	78.57	78.58	88.51	88.43	88.70	**89.16**

values denote are the average values of 10^5 repetitions of a Matlab implementation with no performance optimizations. As can be seen, while there is a dependency of the run-time from both $\mathbf{\Omega}$ and η, for the typical range of the variables, the computation is computationally very inexpensive and never exceeds a value of 89.16 μs. Assuming the evaluation of all six weighting factors, the weight computation per base predictor therefore is around 540 μs at worst. The computation of the coopetitive soft gating function thereby does not add significantly to the overall forecast creation time.

Parameter Determination

The proposed ensemble technique has a low number of parameters only. Besides the power forecasting models to choose, the main parameter is the regularization parameter ζ which regulates the amount of coopetitive soft gating. While choosing an improper value of ζ negatively affects the model performance, it still will perform either as a static model averaging (too much regularization) or pure gating, possibly with overfitting (too little regularization).

An example thereof is given in Table 4.11. The table shows an exemplary trained ensemble model that uses three base predictors and only a single factor for the weighting (e.g., it only uses the weighting $w_g^{(\varphi|\psi)}$ that is computed using Eq. 4.13 given a set of $\bar{S}^{(\varphi|\psi)}$ each computed with Eq. 4.12). During ensemble training of Eq. 4.32 the value of ζ has an influence on which η is determined as optimization result. If the regularization parameter ζ is chosen too high, the weighting penalty exceeds the quality improvements of the weighting, leading to the case (1) laid out in Table 4.11 that equals an averaging of the base predictors. On the other hand, if ζ is chosen too low, a pure gating may occur, which is represented by case (4) in the table. A reasonable parameterization probably is in between the two extremes, such as is the case in (2) and (3) in the table. However, case (1) may still occur even in a model with correct choice of ζ if there is no correspondence between the values of $\bar{S}^{(\varphi|\psi)}$ and a quality improvement using a weighting of the models.

Parameter Optimization

An advantage of the proposed CSGE technique is that the ensemble training (the determination of the weights $\boldsymbol{\eta}$) is a post-processing step of the training of the ensemble members. The ensemble is trained by Eq. 4.32 which optimizes the weights rather than the single ensemble member forecasts. This, in turn, means that during ensemble training, the forecasts of the ensemble members do *not* have to be recomputed when varying $\boldsymbol{\eta}$. The evaluation of the model fitness therefore is possible with little computational effort. Furthermore, the weights

Table 4.11: Resulting Values of η depending on the amount of regularization for an example with three error criteria yielded in an ensemble.

			Error $\bar{S}^{(\varphi\|\psi)}$		
			0.5	1.1	1.3
#	**Regularization Parameter** ζ	**Resulting** η	**Resulting Weights** $w_g^{(\varphi\|\psi)}$		
1	Too much regularization	0	0.33	0.33	0.33
2	Reasonable regularization	1	0.54	0.25	0.21
3	Reasonable regularization	5	0.97	0.02	0.01
4	Too little regularization	10	1.00	0.00	0.00

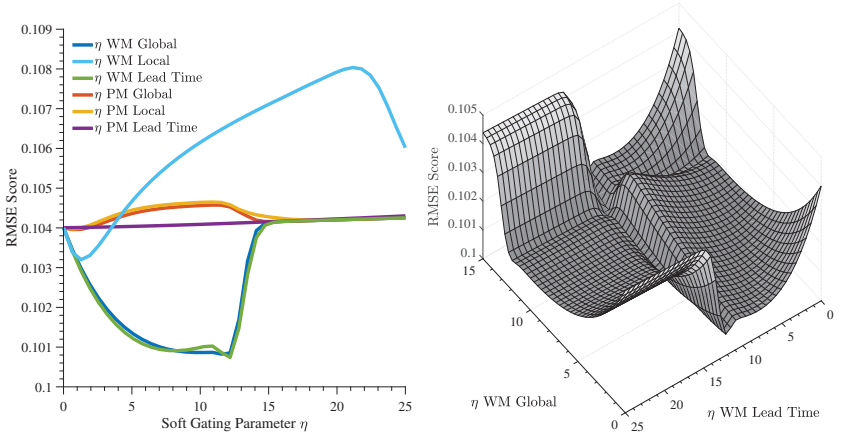
gradually change (i.e., the optimization function is continuously differentiable) when varying the coopetitive soft gating strengths $\boldsymbol{\eta}$, thus the optimization problem is smooth. One can also think of overall optimization using techniques such as simulated annealing, stochastic gradient descent, or particle swarms, possibly leading to an even better optimization result.

Two exemplary figures further detail the optimization characteristics. In Fig. 4.14.1, an example of the influence of each η parameter is given. In this example, a single η parameter is varied while all other η parameters are set to 0. As can be seen from the figure, the parameters of η that have an effect of the weather model dependent weighting are able to benefit more from an unequal weighting. Therein, the parameters that influence the global and lead time-dependent qualities are able to yield the most performance improvements. The parameter η local has an interesting characteristic in this case as it is first able to yield better results when introducing a local weighting, but then the quality decreases. For very high values it converges to the line of no improvement. In this case, the ensemble is able to yield better results when introducing global and local weighting for the PM (around a value of η of 1), however, the improvements are minor in the shown example. The overall upward trend of the error when increasing η is due to the regularization parameter ζ that penalizes high values of η.

Of course, the individual values of η may influence each other. An example of the optimization surface for the η WM global and the η WM lead time is given in Fig. 4.14.2. In the figure, the joint optimization surface for these two parameters is displayed. As can be seen, the values of the parameters both have an influence on the overall RMSE. Furthermore, the smooth nature of the optimization problem can be observed more clearly.

4.7 Conclusion of this Section

This chapter proposes a novel ensemble technique called coopetitive soft gating ensemble that combines multiple deterministic forecasts to a refined overall forecast. There are three main innovations in the design and construction of this ensemble. The innovations are (1) a hierarchical ensemble structure for the aggregation of multiple weather forecasting models *and* power forecasting models, (2) a novel weighting function with a low number of parameters, and (3) a multi-scheme weighting technique that weights the ensemble members dynamically by their overall quality, by their lead time-dependent quality, and their weather-situation dependent quality. As we have discussed, the structure of the ensemble technique is ideal for flexibly applying it in operational contexts due to its robust failure modes. In the experimental evaluation it is shown that the ensemble technique is able to perform other single models and multi-model ensemble techniques. The quality difference becomes particularly apparent if multiple weather models and the lead time-dependent quality differences are present, so in particular for intraday forecasts. It is furthermore advantageous that the ensemble is better



(1) Individual parameter optimization.

(2) Optimization Surface.

Figure 4.14: Influence of choice of individual η parameters on the overall quality of the ensemble. In Fig. 4.14.1, an example of the individual quality development is given. In the experiment, one η parameter is varied while the other η parameters are set to 0. Fig. 4.14.2 shows the optimization surface of two of the six η parameters (global η of WM and lead time-dependent η of WM).

able to retain its performance if poorly performing models are added to the ensemble. Unlike other ensemble techniques, the forecasting quality does not decrease in the investigated case.

Chapter 5

An Overview of Probabilistic Forecasting and Probabilistic Forecasting Techniques

Techniques for power forecasting on short and mid-term horizons (e.g., intraday and day-ahead forecasts) are in nearly all cases based on NWP. Weather forecasting is a stochastic process, meaning that though the current weather condition can be measured up to a certain degree, a future weather situation cannot be exactly predicted due to the chaotic behavior of the fluid dynamics of the atmosphere. This uncertainty in the weather forecasting process affects the power forecasting process. Furthermore, the uncertainty is in many cases amplified, e.g., due to the nonlinearity of the wind turbine power curve. Therefore, while the quality of deterministic point forecasts is still improved (e.g., through model combination), the performance converges towards the intrinsic uncertainty of the underlying NWP generating processes. To overcome this problem, in recent years there has been a shift from the paradigm of creating point forecasts to creating distributional (or probabilistic) forecasts [94]. Probabilistic forecasting quantifies the uncertainty of a prediction and it can be used to retain optimal decision-making performance under uncertain conditions. This is of particular interest in power forecasting for applications such as grid stability and power market operation. While probabilistic forecasting for binary events has been widely applied, e.g., for weather event forecasting (such as the probability of the event of rainfall), there is an increasing demand for probabilistic forecasting of ordinal and continuous quantities in the past decade.

While the area of deterministic forecasting aims at predicting a single value for each look-ahead time (point estimate), the area of probabilistic forecasting tries to additionally assess the uncertainty of a prediction. It can, but not necessarily has to be expressed as a probability [149]. There are a number of techniques for computing, representing, and assessing uncertainty. Though uncertainty does not necessarily have to be represented as a probability, the representation of uncertainty in the form of a probability does have a number of advantages which are described, e.g., in [13, 179], such as the straight forward inclusion in decision making processes, see, e.g., Section 2.9 for examples.

A variety of possibilities to create probabilistic forecasts of continuous quantities have emerged in recent years. These forecasting systems vary in form (e.g., continuously differentiable or stepwise constant probability distributions, intervals, or risk-indices) and the way they are computed (parametric or non-parametric uncertainty estimates, from single forecasts or ensembles). Depending on the form of uncertainty representation, different methods of performance assessment have emerged, which are in some cases specialized for a particular form of uncertainty representation.

In this chapter, we give an overview of the most relevant techniques to estimate uncertainty and highlight the most common forms of uncertainty representation. Therein, we present a holistic view of the problem of probabilistic forecasting to enable a better comparability between different forms of uncertainty representations by converting them to density functions. Having a common form of representation, the assessment of their performance is easier.

The remainder of this chapter is structured as follows: Possible prediction spaces are detailed in Section 5.1, a unified scheme for the construction of predictive distributions from quantiles is introduced in Section 5.2. General desired properties of uncertainty representation techniques are laid out in Section 5.3. Techniques for the visual investigation of these desired properties are highlighted in Section 5.4. An overview of techniques to create predictive distributions is given in Section 5.5. Sections 5.6 and 5.7 then give a more detailed description of the techniques given in Section 5.5 created from single predictive models (a single NWP) and ensembles (multiple NWP), respectively.

5.1 Prediction Spaces

Probabilistic forecasts can be created for different forms of target predictands, i.e., they can be created for a variety of prediction spaces. These predictands can be categorized into three main areas:

- *Binary predictand*: The target is a binary outcome. This form of forecast tries to predict a target event with a certain probability. Typical events in meteorological sciences can, for instance, be the probability that the amount of precipitation or wind speed exceeds a certain threshold.
- *Categorical predictand*: Categorical predictions are the generalization of a binary prediction to multiple classes. This form of prediction can be subcategorized into nominal and ordinal predictions. An ordinal prediction has a ranking of the classes attached, while nominal predictions have no rank. An example for a nominal prediction task in meteorological sciences may be the classification of cloud types. Ordinal predictions can, for instance, be discretized classes of wind intensity (e.g., on the Beaufort scale), which have to be predicted *directly*, i.e., without predicting the (continuous) wind speed first and converting the wind speed to a wind intensity class afterwards.
- *Continuous predictand*: Continuous predictions try to predict a continuous value (not a class). This can, e.g., be the wind speed (e.g., in km/h), or the power generation of a renewable energy power plant.

Here we focus on continuous predictands, as they are of particular interest for power forecasting applications. Furthermore, binary and categorical predictions can easily be constructed from a continuous forecasting task by introducing thresholds on the continuous predictand.

5.2 Representations of Predictive Distributions

The most universal representation of a predictive distribution is in the form of a probability density function (pdf) $\hat{p}(y)$, a function with property

$$\int_{-\infty}^{+\infty} \hat{p}(y) = 1, \quad \hat{p}(y) \geq 0, \tag{5.1}$$

which can be evaluated at an arbitrary value y (which typically is a power value in power forecasting applications) to get the probability density for this value. The corresponding cumulative density function (cdf) $\hat{P}(y)$ is computed in the form

$$\hat{P}(y) = \int_{-\infty}^{y} \hat{p}(y')\mathrm{d}y'. \tag{5.2}$$

Thus, a conversion of the representations is possible. To recapitulate Section 2.4.2, a predictive distribution can be constructed in different forms, such as using a

- (combination of) well-defined continuous density functions, such as parametric density functions with density estimation algorithms,
- sampling a number of deterministic forecasts from the (non-observable) underlying distribution of possible outcomes (e.g., from an EPS),
- prediction intervals, which can be represented as pdfs, or
- quantile forecasts.

The interpretation of (combinations of) continuous density functions is straight forward. The conversion from prediction intervals to density functions is explained in Section 5.6.5, while the creation of the predictive distribution from samples drawn from an underlying distribution is detailed in Section 5.7.1. As described in [92], quantiles are the basis for a number of probabilistic forecasting algorithms and provide a decision-theoretically optimal framework for predictive distributions (a decision-making process using density functions is described in Section 2.9.2). Thus, we will explain the construction and relationship from quantile forecasts to predictive distributions in the following in more detail.

In general, a quantile defines a cutting point in a set of observations. A quantile $\tau \in [0,1]$ is typically specified in the sense of a percentile, so, for instance, the $\tau = 0.1$ quantile specifies the cutting point in a set of observations where 10 % of the observed values are expected to be below the cutting point. To assess the uncertainty of a forecast, a predictive model is trained to estimate the location of a predictive *quantile forecast* $\hat{y}^{(\tau)}$ (instead of the best point estimate in the sense of a deterministic forecast) whose location is defined by

$$\hat{y}^{(\tau)} = \hat{P}^{-1}(\tau), \tag{5.3}$$

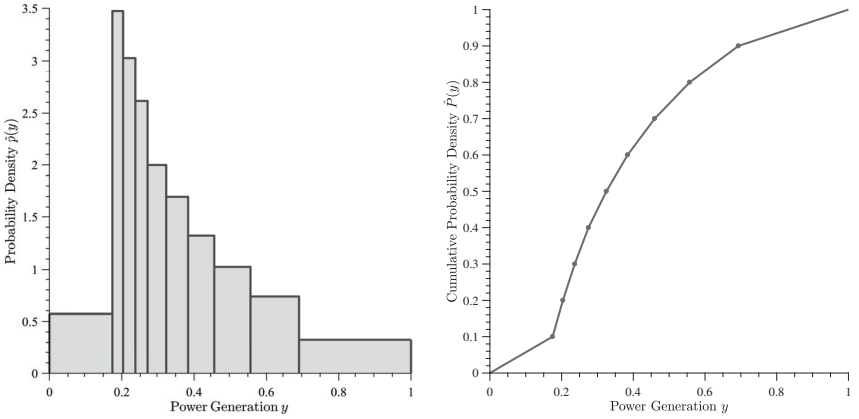
which can equivalently be expressed in the form

$$\int_{-\infty}^{\hat{y}^{(\tau)}} \hat{p}(y)\ \mathrm{d}y = \tau. \tag{5.4}$$

The locations of $\hat{y}^{(\tau)}$ with $\tau \neq 0.5$ are assumed to systematically under- or overestimate the forecast. For instance, for the location $\hat{y}^{(0.1)}$ (with $\tau = 0.1$), it is expected that the resulting function systematically underestimates the forecast and that the "true" (observed) value o is below the forecasted value only in 10 % of all cases. Thus, in a reliable forecasting system, the quantile value τ asymptotically also specifies the probability

$$P(o < \hat{y}^{(\tau)}) = \tau. \tag{5.5}$$

When creating a stepwise constant predictive distribution, an algorithm tries to estimate the quantile forecasts of a number of L defined quantiles, which systematically under- or overestimate the median point forecast. The number of quantiles depends on the desired



(1) Histogram pdf. (2) Histogram cdf.

Figure 5.1: Representation of a density function from a set of given quantile forecasts. Adjacent quantiles each have a defined probability mass, which leads to a histogram representation (Fig. 5.1.1) of the probability density function (pdf) if the probability distribution within the probability bin is assumed to be uniform (i.e., the pdf is a piecewise constant function). For the cumulative density function (cdf) construction, this leads to a piecewise linear function (Fig. 5.1.2). Alternative variants of pdf construction from ensembles are shown in Section 5.7.

precision of the density function. The entirety of quantile forecasts $\left(\hat{y}^{(\tau_1)},\ldots,\hat{y}^{(\tau_L)}\right)$ then can be used to form the overall forecast distribution. The value range between two subsequent quantile values is assumed to contain a certain probability mass p_l defined by

$$p_l = \tau_{l+1} - \tau_l, \tag{5.6}$$

which leads to the pdf shown in Fig. 5.1.1. Therefore, the narrower the interval between two predictive quantile forecasts $\hat{y}^{(\tau_l)}, \hat{y}^{(\tau_{l+1})}$, the higher the value of the probability density function. The cdf given the tuples (which help to estimate probability mass of the marginal quantiles)

$$\begin{aligned} \hat{\mathbf{y}} &= \left(0, \hat{y}^{(\tau_1)},\ldots,\hat{y}^{(\tau_L)}, o_{\text{inst}}\right), \\ \boldsymbol{\tau} &= \left(0, \tau_1,\ldots,\tau_L, 1\right), \end{aligned}$$

with installed capacity o_{inst} (or other assumed maximum value), which is the maximal possible value of the predictand, and $\hat{\mathbf{y}}, \boldsymbol{\tau} \in \mathbb{R}^{L+2}$ and $v = 1,\ldots,L+2$ can then be defined as

$$\hat{P}(y) = \frac{\boldsymbol{\tau}_{v+1} - \boldsymbol{\tau}_v}{\hat{\mathbf{y}}_{v+1} - \hat{\mathbf{y}}_v} \cdot (y - \hat{\mathbf{y}}_v) + \boldsymbol{\tau}_v, \text{ with } y \in [\hat{\mathbf{y}}_v, \hat{\mathbf{y}}_{v+1}]. \tag{5.7}$$

This form of cdf representation has clear advantages (compared, e.g., to sampling based density representations, see Section 5.7.1) when estimating the extreme quantiles, as the probability mass for these quantiles is well defined. For power forecasting, typically the value range of the predictand is bounded within the interval $[0, o_{\text{inst}}]$, therefore, the marginal

histogram bins can easily be defined (see Fig. 5.1.1 and 5.1.2). Furthermore, this form of representation is better suited for the evaluation of error scores which evaluate the pdf directly rather than the cdf, e.g., the ignorance score (see Section 6.1.2). However, this form of cdf representation assumes a uniform distribution of the probability density within each histogram bin, which may not reflect the real distribution of values. This disadvantage can be overcome partly by creating more quantile forecasts for the pdf construction, and forecasting the extreme quantiles (i.e., $\tau \to 0$ and $\tau \to 1$).

5.3 Desired Model Properties of Predictive Distributions

There are a number of properties which a probabilistic forecasting algorithm has to fulfill in order to provide a benefit over a deterministic forecasting algorithm. These predictive distributions are only beneficial if they estimate the conditional distribution of the observation values correctly. We will highlight these properties in the following (cf., e.g., [93, 193], or [259]). The principal properties of a probabilistic forecast are

- reliability,
- sharpness, and
- skill.

One central aspect of a probabilistic forecast is the *reliability*. For binary events, reliability describes whether the probability a forecasting algorithm issues matches the observed frequency of occurrence of the particular forecasted value on any given level of probability. Thus, a forecast for an event with a probability of occurrence of, e.g., 0.7 should ideally have a corresponding observed frequency of the event actually occurring of 70 %. For continuous predictands, reliability describes whether the conditional variance of a forecast is correctly assessed in each situation. Thus, a predictive distribution has to model the spread of the target variable of the observed uncertain process correctly. If the distributions of forecasted and observed values do not match, the forecasting system is unreliable. This may be due to one or multiple of the following effects:

- Bias error: The forecasting system has a systematic error regarding the predictive distributions in the sense that it creates forecasts which are too high or too low. This is an error in the sense of an offset. For forecasting binary events such as a rain event the terms *wet* or *dry* bias are also common.
- Confidence error: A confidence error means that the *spread* (i.e., the variance in the case of normal distribution) of the forecast is not correctly assessed. The created distributions are too narrow (overconfident forecasting model) or too wide (underconfident forecasting model). In the meteorological context, the spread is also referred to under the term *dispersion*, which is mostly used to describe the spread of ensemble members in a meteorological ensemble.

As reliability is a term that was popularized from the practical use of probabilistic forecasts, it is mostly described textually and visually (as, e.g., performed in [3, 94, 189, 253]). The authors of [193] proposed a scheme for the evaluation of reliability in the context of non-parametric forecasting models using tuples of quantile forecasts $\left(\hat{y}^{(\tau_1)}, \ldots, \hat{y}^{(\tau_L)}\right)$ (which can however be constructed from parametric distributions as described in Section 5.2). Given a data set with $\boldsymbol{o} = \left(o_1, \ldots, o_N\right)$ observations and corresponding quantile forecasts for each

evaluated point in time $n \in 1,\ldots,N$, the frequency of observations being below a quantile forecast can be determined using a sum of indicator functions in the way

$$\nu^{(\tau)} = \frac{1}{N}\sum_{n=1}^{N} H(\hat{y}_n^{(\tau)} - o_n) \tag{5.8}$$

with H being a Heaviside step function. The discrepancy of the observed frequency $\nu^{(\tau)}$ to the expected frequency τ can then be computed for a single quantile with

$$\nu^{(\tau)} - \tau. \tag{5.9}$$

While the authors of [193] argue that an averaging of these single discrepancies may lead to error canceling effect (if one $\nu^{(\tau_i)}$ is too low and another $\nu^{(\tau_j)}$ with $i \neq j$ is too high), we think this can easily be solved when using the absolute deviations to form an overall reliability term $\bar{\nu}$ in the form

$$\bar{\nu} = \frac{1}{L}\sum_{l=1}^{L} |\nu^{(\tau_l)} - \tau_l|. \tag{5.10}$$

We furthermore propose a generalization of this reliability evaluation which can also be used for continuous distributions such as parametric forecasts with

$$\bar{\nu} = \int_{\tau=0}^{1} |\nu^{(\tau)} - \tau| \mathrm{d}\tau, \tag{5.11}$$

where each $\hat{y}^{(\tau)}$ can be computed with $\hat{y}^{(\tau)} = \hat{P}^{-1}(\tau)$. The decision wether a forecasting model is reliable can be performed in a number of ways (e.g., using a threshold value on $\bar{\nu}$, using thresholds on each individual $|\nu^{(\tau_l)} - \tau_l|$, or using a χ^2 test such as performed in [22] and described in Section 5.4.1, which however is questioned by [193] as the assumption of uncorrelated errors may not hold). Reliability in general only contributes to a diagnostic evaluation of a forecasting model. The reliability of a forecast can also be assessed using the decomposition of scoring rules (see Section 6.1), or graphical assessment techniques, such as the probability integral transform (PIT) histogram (see Section 5.4.1). In some cases, reliability is also referred to as *calibration* directly, e.g., in [95], rather than calibration being the process of modifying a predictive distribution to achieve reliability.

Reliability is *not* a measure of forecasting precision in the sense of an accurate point estimate, but only guarantees the creation of statistically sound predictive distributions. For instance, climatological forecasts (see Fig. 5.2.1) always issue the same long-term average probability distribution independent from a particular observed weather situation. These sorts of forecast have optimal reliability, however, they do lack *sharpness*, the second central property of a probabilistic forecast. The sharpness measures the "narrowness" of a probability distribution. As for reliability, sharpness is mostly described textually and visually (e.g., in [3, 73, 94, 193, 253]). However, the authors of [193] define sharpness for non-parametric forecasting models using sets of symmetric quantiles

$$\kappa^{(\alpha)} = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{y}_n^{(\tau_{\bar{u}})} - \hat{y}_n^{(\tau_{\bar{l}})}\right), \tag{5.12}$$

where $(1-\alpha)$ is the nominal coverage probability (assumed fraction of samples within the

interval set out by the lower bound $\hat{y}^{(\tau_{\hat{l}})}$ and the upper bound $\hat{y}^{(\tau_{\hat{u}})}$ with $\alpha \in [0,1]$) with

$$\frac{\alpha}{2} = 1 - \tau_{\hat{u}} = \tau_{\hat{l}}. \tag{5.13}$$

For instance, for a nominal coverage probability of $(1-\alpha) = 90$ % (with $\alpha = 0.1$), the interval is defined with the lower bound $\hat{y}^{(0.05)}$ and upper bound $\hat{y}^{(0.95)}$. More details on the definition of intervals is given in Section 5.6.5. While not discussed in the literature, an averaging of the sharpness $\kappa^{(\alpha)}$ for an individual α can make sense to get a measure of the overall sharpness of the predictive distribution for all $\tau_1, \ldots, \tau_L$ quantiles with

$$\bar{\kappa} = \frac{1}{\lfloor \frac{L}{2} \rfloor} \sum_{l=1}^{\lfloor \frac{L}{2} \rfloor} \kappa^{(2 \cdot \tau_l)}, \tag{5.14}$$

where $\lfloor \frac{L}{2} \rfloor$ is a function that rounds $\frac{L}{2}$ down (which means that in case of an odd number of quantiles the central quantile is not evaluated). Furthermore, we propose an extension of sharpness to continuous forecasting distributions (such as parametric forecasts) which can be investigated with

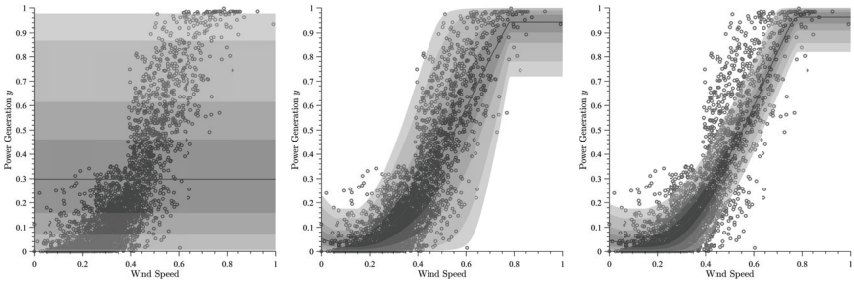
$$\bar{\kappa} = \int_{\alpha=0}^{1} \kappa^{(\alpha)} \mathrm{d}\alpha, \tag{5.15}$$

where the intervals can be computed with $\hat{y}_n^{(\tau_{\hat{l}})} = \hat{P}^{-1}\left(\frac{\alpha}{2}\right)$ and $\hat{y}_n^{(\tau_{\hat{u}})} = \hat{P}^{-1}\left(1 - \frac{\alpha}{2}\right)$, respectively.

If two reliable probabilistic forecasts are given, the one with higher sharpness is preferable (see Fig. 5.2.2 in comparison to Fig. 5.2.1). However, if a sharp forecast is not reliable, it assumes unrealistic confidences of the probability distributions, and is, therefore, not desirable (see Fig. 5.2.3). As pointed out, e.g., in [193, 259], reliability is the primary requirement of a predictive probability distribution. Other techniques for the evaluation of sharpness are diagrams that give $\kappa^{(\alpha)}$ as a function of the nominal coverage probability $1-\alpha$ which are described in Section 5.4.2.

In a nutshell, reliability refers to whether the issued predictive distributions are *correct*, while sharpness (given reliability) gives insights on the *quality* (and thereby usefulness) of the forecast. The goal of model training of a probabilistic forecasting system is to maximize sharpness (low value of $\bar{\kappa}$) while retaining reliability. This can be done using a loss function that optimizes for both reliability and sharpness simultaneously. These functions are commonly referred to as scoring rules which assess the forecasting *skill*. The skill quantifies the observable error of a probabilistic forecasting technique on a data set, it thereby incorporates errors which can be attributed to both reliability and sharpness errors. The type of error (e.g., reliability error) of a predictive distribution can be revealed using a decomposition of the error components, which has been proposed for a number of scoring rules, or using visual verification techniques, which is detailed in the following section. Scoring rules are detailed in Section 6.1.

If a forecasting system does not meet the requirements regarding forecasting skill, it can partly be optimized using post-processing techniques. This process of maximizing the forecast skill is called statistical forecast calibration (e.g., described in [93]) and is often used to achieve reliability.



(1) Not sharp, but reliable model. (2) Sharp and reliable model. (3) (Overly) sharp, unreliable model.

Figure 5.2: Visualization of the sharpness and reliability properties. The above figures show a data set (dark circles) with a trained probabilistic model with forecasted cumulative density function (cdf) indicated by the grey intervals that show the cdf values of $0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99$, and 0.999. More details on the representation of predictive distributions are given in Section 5.2. Fig 5.2.1 shows a sample climatology forecasting model. Issued forecasts of this model are not sharp, as can be seen by the very broad probability distributions that are independent from the wind speed. However, the model is reliable, as the samples are distributed in the shaded areas as the given probabilities suggest. Fig. 5.2.2 shows a reliable model with forecasts that have increased sharpness in comparison to the climatology forecast as is indicated by the narrower probability distributions. An overly sharp model that consequently creates unreliable forecasts is shown in Fig. 5.2.3. The outmost bounds indicate that only 1 ‰ of the samples are above or below the outmost bounds, respectively, which clearly is not the case in this example. If a sharp forecast is not reliable, it assumes unrealistic confidence of the probability distributions, and is, therefore, not desirable. More details on the reliability and sharpness properties can be found in Section 5.3. Details on verification techniques for predictive distributions can be found in Section 5.4.1.

5.4 Visual Verification of Predictive Distributions

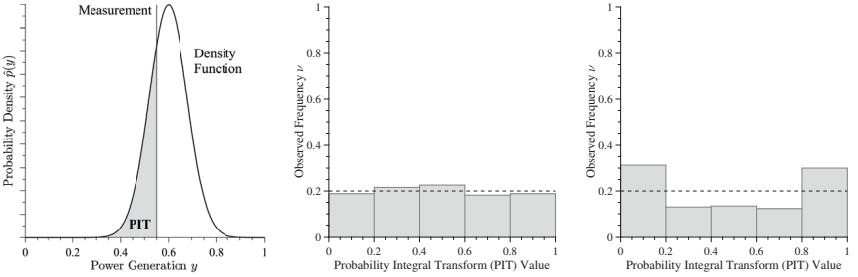
This section gives an overview of methods for the visual verification of probabilistic forecasting models regarding both reliability and sharpness. Section 5.4.1 gives an overview of methods for the visual verification of the reliability property, while Section 5.4.2 highlights a method for visual sharpness investigation.

5.4.1 Visual Verification of Reliability

While there are a number of visual verification techniques for binary events such as the probability of a rain event (examples are the reliability diagram [169], discrimination diagram [158], receiver operating characteristic (ROC) curves [107], or diagrams based on the value score for cost/loss functions [208]), they can in many cases not be applied for the visual inspection of continuous probabilistic forecasts (see Section 5.1) in a simple way.

Two techniques which can be used for visual inspection of continuous forecasts are the probability integral transform (PIT) histogram, and the quantile-quantile (QQ) plot, which are very close in their expressiveness. The PIT histogram is a simple measure for the assessment of reliability of a probabilistic forecast. It can easily be used to assess the dispersion and

bias effects of a predictive distribution. The PIT histogram is closely related to the *Talagrand diagram* (or *rank histogram*) described, e.g., in [33], which is, however, mainly used for EPS forecasts. Conceptually, this approach is also related to measuring emergence for continuous quantities [65].



(1) PIT creation. (2) PIT histogram (likely reliable). (3) PIT histogram (underdispersive).

Figure 5.3: Example of visual forecast verification techniques. A popular instrument for visual forecast verification in a continuous prediction space is created using probability integral transform (PIT), which is the integral of a predictive distribution up to the point of observation (see Fig. 5.3.1). The PIT values are analyzed in the form of a PIT histogram, which gives indication on the reliability of a forecasting system (shown in Figs. 5.3.2 and 5.3.3). Ideally, the probability bins are equally populated. The histogram gives indications on the kind of error a forecasting system makes, e.g., an underdispersive forecasting system in the case of Fig. 5.3.3.

The PIT histogram evaluates a tuple of (probabilistic) forecast-observation pairs using the PIT computed by

$$\mathrm{PIT}(\hat{p}(y), o) = \int_{-\infty}^{o} \hat{p}(y)\mathrm{d}y. \tag{5.16}$$

It represents the probability mass below the value of a measured observation o given a predictive distribution $\hat{p}(y)$. The value of the PIT therefore is in the interval $[0,1]$. A graphical representation for this process is visualized in Fig. 5.3.1. The resulting values of the PIT for each evaluated predictive density are aggregated in a histogram, where, given a reliable forecast, a number of discretized probability bins should be equally populated as visualized in Fig. 5.3.2.

For a realistic forecast with a limited number of observations, even for a reliable forecasting system it is unlikely to observe equally populated bins in the PIT histogram. A deviation from reliability can not only be investigated visually (as shown in Fig. 5.3.3), but also, e.g., using a χ^2 test [105, 248]. The χ^2 test can assess the likeliness of observing a reliable forecasting system given a PIT histogram (with null hypothesis being the system is reliable). To recall, whether this test can actually be used is part of an ongoing discussion as described in [193]. Given a relatively equal distribution of the PIT histogram bins (such as in Fig. 5.3.2), the forecasting system is likely to be reliable. If however, the model does not have enough spread, both outer bins are overpopulated (thus the forecasting system is *underdispersive*), which can be observed in Fig. 5.3.3. Other effects, such as biased or overdispersed models, can also be detected using PIT histograms.

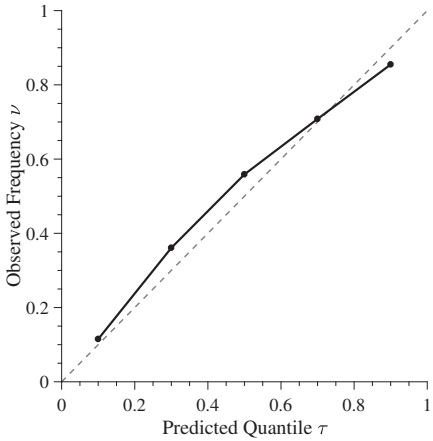


Figure 5.4: An alternative visualization technique, the quantile-quantile (QQ) plot, for the investigation of the reliability property. A deviation from the ideal (diagonal) line indicates an unrealiable forecasting system.

An alternative representation to PIT histograms are quantile-quantile (QQ) plots, a standard measure for comparing distributions [35], which are closely related to reliability diagrams (but can, unlike reliability diagrams, be used in the continuous domain). The QQ plot is especially suited when having a set with $n = 1, \ldots, N$ of the predictive quantile estimates $\hat{y}_n^{(\tau)}$, as it evaluates the frequency $\nu^{(\tau)}$ of an observation being lower than a predicted quantile forecast, i.e.,

$$\nu^{(\tau)} = \frac{1}{N} \sum_{n=1}^{N} H(\hat{y}_n^{(\tau)} - o_n), \tag{5.17}$$

with H being a Heaviside step function (cf. Eq. 5.8). The expressiveness is very close to those of PIT histograms, however, they may be easier to understand if the probability bin borders are not chosen equidistantly. QQ plots have, e.g., been utilized in [69, 177] in the area of meteorological sciences. An example of a QQ plot is shown in Fig. 5.4. In this figure, the abscissa denotes the predicted cumulative probability τ. The deviation of an investigated pair $(\tau, \nu^{(\tau)})$ from the ideal line (diagonal dashed line in the figure) indicates an unreliable forecasting system. The principle remains the same as for the PIT histogram that it is increasingly unlikely to observe a $\nu^{(\tau)}$ that deviates from the theoretical optimum given a reliable forecasting system which can be assessed using a χ^2 test.

The authors of [193] introduced an alternative version of the QQ plot that only shows the deviation and thereby is more focused on a diagnostic analysis of forecasts as shown in Fig. 5.5. We think this *reliability diagram* is the chart that enables the clearest investigation of the reliability property. An actual application example of the reliability diagram is given in Section 7.3.4.

5.4.2 Visual Verification of Sharpness

A method for the visual verification of sharpness is presented in [193]. The method is based on the evaluation of symmetric quantile forecasts that form an interval such as described in

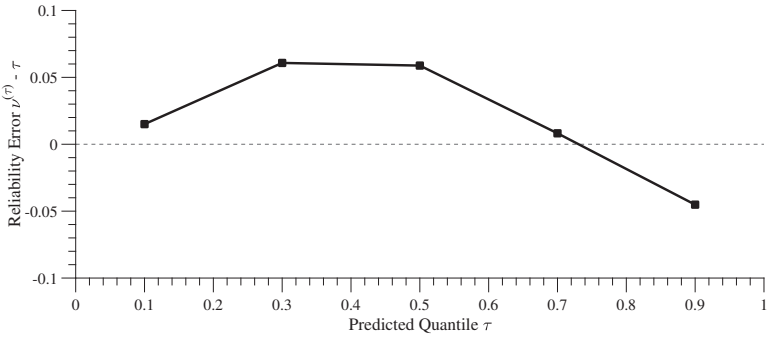


Figure 5.5: The *reliability diagram*, a version of the quantile-quantile (QQ) plot that is focused on a diagnostic analysis of reliability. In this example, the lower three quantiles have a bias (too high forecasts), while the upmost quantile creates too low forecasts.

Section 5.3 from which sharpness values $\kappa^{(\alpha)}$ given a nominal coverage probability $(1-\alpha)$ can be computed using Eq. 5.12. The diagram then computes the value of $\kappa^{(\alpha)}$ depending on the nominal coverage probability.

An example of this diagram is given in Figs. 5.6 and 5.7. Fig. 5.6 shows four probability distributions of various types, namely two normal distributions (truncated in the interval $[0,1]$), a uniform distribution, and a Beta distribution. The corresponding sharpness diagram is shown in Fig. 5.7. For the creation of this diagram, first, the quantile positions $\hat{y}^{(0.05)}, \hat{y}^{(0.1)}, \ldots, \hat{y}^{(0.95)}$ are determined using $\hat{P}^{-1}(\tau)$ (in the case of a non-parametric forecasting model, the quantile forecasts are given directly). Each sharpness value is then computed given a nominal coverage probability $(1-\alpha)$ using Eq. 5.12. The diagram gives insights about the individual sharpness values and thus of the shape of the probability distribution. As can be seen in the present case, the normal distribution with $\sigma = 0.1$ is sharper than the uniform distribution up to a nominal coverage of $(1-\alpha) = 0.85$. The beta distribution and the Gaussian distribution with $\sigma = 0.2$ are both less sharp than the other investigated distributions. The corresponding overall sharpness values $\bar{\kappa}$ of the individual probability distributions are given in Table 5.1 which condenses the insights of Fig. 5.7. Again, it should be mentioned that sharpness is only a metric of the forecasting quality if the forecasting model is reliable (as, for instance, the actual corresponding observations o are not used for the creation of this diagram). Sharpness values therefore should not be used as a metric of the forecasting quality but rather as a diagnostic evaluation tool.

Table 5.1: Overall sharpness values $\bar{\kappa}$ of the distributions examined in Fig. 5.6 computed using Eq. 5.14.

Probability Distribution	$\bar{\kappa}$
Normal Distribution ($\mu = 0.5, \sigma = 0.1$)	0.154
Normal Distribution ($\mu = 0.5, \sigma = 0.2$)	0.301
Uniform Distribution $[0.33, 0.67]$	0.170
Beta Distribution ($a = 0.4, b = 0.4$)	0.689

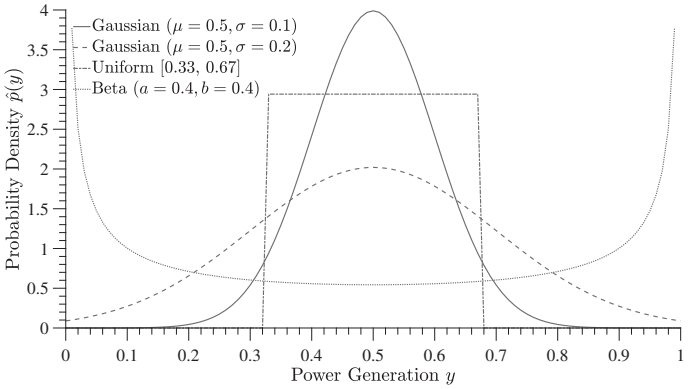


Figure 5.6: Example of sharpness assessment of four probability density functions (pdf). The figure shows two normal distributions with varying variance, a uniform probability distribution, and a Beta distribution. Sharpness diagrams of these pdfs are given in Fig. 5.7.

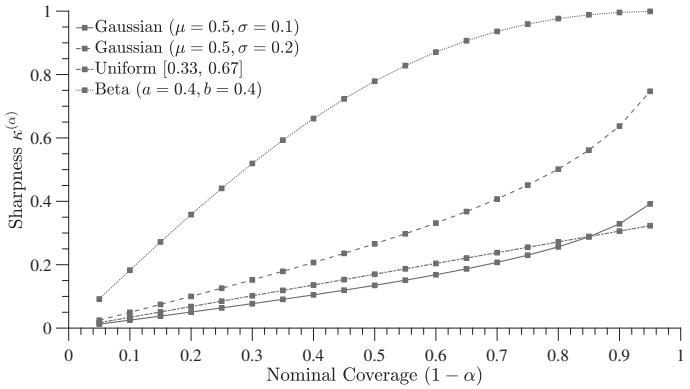


Figure 5.7: Example of a sharpness diagram. The diagram shows the sharpness values given different nominal coverage values $(1-\alpha)$. As can be seen from the figure, the pdfs have varying sharpness given different nominal coverage values. The overall sharpness values $\bar{\kappa}$ are given in Table 5.1.

5.5 Overview of Forms of Predictive Distribution Construction

This section gives an overview of methods to create predictive distributions. The most important methods are shown in Fig. 5.8. We introduce a novel categorization of methods by their construction from single predictive models or ensembles. In the context of power forecasting, those predictive models can be single weather models or multiple weather models in an ensemble. Therein, predictive distributions from single predictor models can be created using parametric density functions, similarity-based forecasts, or direct training of machine learning models using modified cost functions. Probabilistic ensemble methods, on the other hand, can be constructed using direct pdf construction from sampling, processed ensemble member density functions, or estimating the uncertainty using skill categories.

More details on the construction of predictive distributions from single predictive models are given in Section 5.6, while the creation of predictive distributions from ensembles is detailed in Section 5.7.

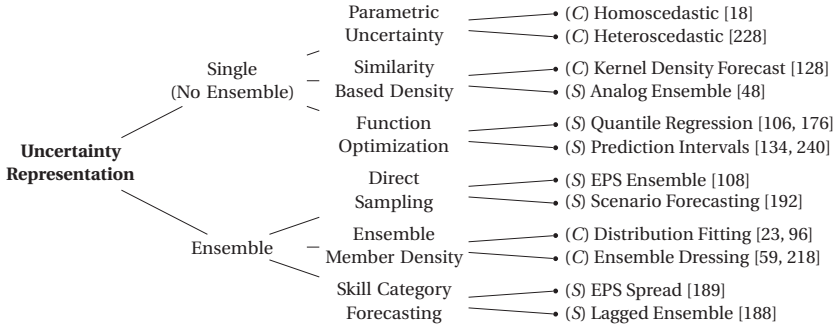


Figure 5.8: Categorization of existing representations of forecast uncertainty by their origin from a single weather model or an ensemble method. (*C*) and (*S*) indicate a continuous or stepwise constant (non-continuously differentiable) probability distribution, respectively. More details on uncertainty representations from single predictive models are given in Section 5.6, while uncertainty representation from ensembles is described in Section 5.7. EPS is the abbreviation for *ensemble prediction system*, which is also detailed in Section 5.7.

5.6 Predictive Distribution Construction from Single NWP Predictor Models

This section describes the predictive distribution construction from a single NWP model, i.e., without using a meteorological ensemble (such as an EPS or multi-model ensemble).

5.6.1 Parametric Density Functions

Parametric density functions create a predictive density function by estimating the parameters of a density function with predefined basic shape (i.e., functional form). In the *homoscedastic*

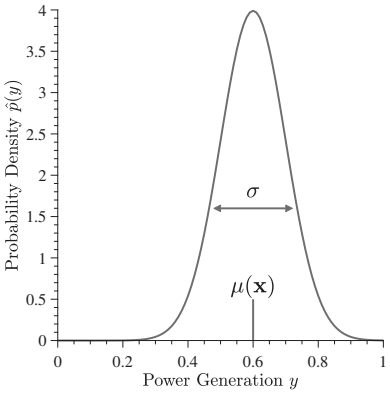


Figure 5.9: A homoscedastic parametric Gaussian function detailed in Section 5.6.1. The expectation value of the density function is at the location of the deterministic forecast.

case (i.e., the width of the density function does not change over the predictand space y), the predictive distribution $\hat{p}(y)$ can be created using, for instance, a normal distribution $\mathcal{N}$. The value of the expectation can be given by a deterministic forecast $\hat{y} = \mu(\mathbf{x})$ based on NWP $\mathbf{x}$, while the standard deviation σ is estimated during model training using a validation set with

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(o_n - \mu(\mathbf{x}_n)\right)^2} \tag{5.18}$$

and

$$\hat{p}(y|\mathbf{x}) = \mathcal{N}(y|\mu(\mathbf{x}),\sigma). \tag{5.19}$$

A simple homoscedastic approach is, for instance, proposed in [18]. An example of a parametric density function is shown in Fig. 5.9. The figure shows a parametric density function with predefined width which is positioned so that the expectation value of the distribution is positioned at the location of the deterministic forecast. The use of other parametric distributions (e.g., beta or gamma distributions) is discussed in [259]. In the *heteroscedastic* case, the distribution is extended to model the standard deviation as a function $\sigma(\mathbf{x})$ which is also functionally dependent on the input value of $\mathbf{x}$ in the form

$$\hat{p}(y|\mathbf{x}) = \mathcal{N}(y|\mu(\mathbf{x}),\sigma(\mathbf{x})). \tag{5.20}$$

A popular parametric heteroscedastic approach is homogeneous (non-)linear regression, which is proposed, e.g., in [228]. In the statistical domain, heteroscedastic autoregressive models are also frequently used (e.g., in [14]).

5.6.2 Kernel Density Estimation

A pdf can be represented from a set of kernels using kernel density estimation (KDE) techniques, an overview of KDE methods for power forecasting can be found in [259]. A KDE

performs a non-parametric density estimation and can be expressed as

$$\hat{p}(y) = \frac{1}{N} \sum_{n=1}^{N} K\Big(\frac{y - o_n}{h}\Big), \tag{5.21}$$

with observations o_n where h is a kernel width parameter and N is the number of data points used for the KDE. K is the kernel function which is normalized so that

$$\int_{-\infty}^{+\infty} K\Big(\frac{y - o_n}{h}\Big) \mathrm{d}y = 1. \tag{5.22}$$

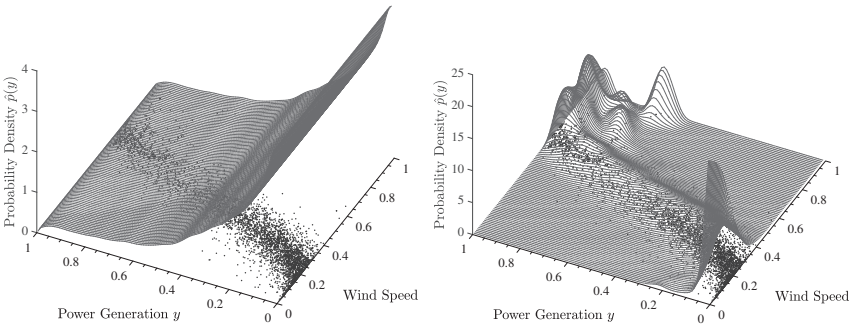
Depending on the type of bounds of the predictand (for a typical power forecast, the bounds are $[0, o_{\text{inst}}]$), different kernel functions can be applied, e.g., Normal, Beta, or Gamma kernels as described in [259]. Kernel density methods have the advantage of not making any assumptions on the functional form of the overall probability distribution, and, therefore, they are very well suited for modeling multi-modal or skewed distributions. KDE methods, on the other hand, require many samples in order to perform accurate kernel density estimation. For most kernels, the result is a continuously differentiable function. When performing a KDE on all available predictor-predictand pairs during training, the result can be considered as the unconditional probability density of the data set (the sample climatology in the case of meteorology) as described, e.g., in [124, 259]. Though practitioned on a regular basis, this is a reinterpretation of the result of the KDE to the predictive distribution that is not accurate from a theoretical point of view. This forecasting system is shown in Fig. 5.10.1. As can be seen, the forecast does not depend on the actual value of the wind speed but is constant, leading to a forecast with only little sharpness.

In practice, however, when performing a KDE, a *conditional* subset of observations (conditioned, e.g., by the current NWP forecast for which a probabilistic power forecast will be created) should be chosen in order to create a sharper predictive distribution. An example of these conditioned KDE forecasts is visualized in Fig. 5.10.2. In the example, the conditional probability distributions (grey lines) are created from the observations o (the black dots). When performing a forecast with KDE, the probability distribution with corresponding wind speed of the forecast is used.

This process is related to the analog ensemble, see Section 5.6.3. An analysis of KDE methods for long-term forecasting horizons is performed in [196]. A multivariate variant of KDE for wind distribution modeling is detailed in [256]. In this variant, a joint KDE of the multivariate predictor variables and predictand is created. Then, this joint KDE is conditioned (e.g., by the predictor) to create a predictive distribution with increased sharpness (similar to the case laid out in Fig. 5.10.2).

5.6.3 Analog Ensemble

An analog ensemble is a nearest neighbor search technique which creates a forecast from historically similar weather situations (details are given in Section 2.6.4). Having a historic set of weather situation and power generation pairs, the basic idea is that a specified number of J similar weather situations in the set (where each situation is represented by the NWP forecast $\mathbf{x}$) are found in a historic data set using an appropriate distance metric (e.g., the Euclidean distance). The sorted tuple (from low to high) of corresponding power generation measurements $\hat{\mathbf{y}}^{(\text{AE})} = \left(\hat{y}_1, \ldots, \hat{y}_J\right)$ that form the ensemble, which consists of the power measurements of the J NWP forecasts with smallest distance regarding $\mathbf{x}$, can then be used for a



(1) Representation of an unconditional KDE.

(2) Representation of a conditional KDE.

Figure 5.10: Representation of conditional kernel density estimation (KDE) with normal distributions truncated in the interval $[0,1]$. The black dots are observations o of a training data set. The grey lines represent the conditional probability distribution of the power generation depending on the wind speed. While the unconditional KDE of Fig. 5.10.1 is likely reliable, the quality is poor due to the unconditional forecast. The conditional KDE better represents the dependency of wind speed and power generation.

weighted deterministic forecast (e.g., performed in [85]), or for the assessment of forecasting uncertainty [48]. The respective quantile forecast $\hat{y}^{(\tau)}$ is then created simply by counting the sorted forecasts up to element j in the form

$$\hat{y}^{(\tau)} = \hat{\mathbf{y}}_j^{(\mathrm{AE})}, \text{ with } j = \inf\{j' : j' > \tau \cdot J\}, \; j,\; j' \in \mathbb{N}, \tag{5.23}$$

yielding a stepwise constant probability representation which can be transformed to a pdf using Eq. 5.7. While this technique constructs a stepwise constant predictive distribution, it has similarities with kernel density estimation techniques regarding similarity-based construction of the forecasting distribution, such as the property of creating asymptotically unbiased forecasts (i.e., the climatological forecast when using all available data points for the pdf computation), which is described in [12].

5.6.4 Quantile Regression

The idea of quantile regression (QR) is to optimize a modified cost function during model training of the forecasting model to create a forecasting model which creates a quantile forecast $\hat{y}^{(\tau)} = f_\tau(\mathbf{x})$. Rather than using an MAE or MSE function as error criterion, a modified form of the MAE called quantile score (often also called *pinball* or *check* function) is chosen. Therein, the parameter $\tau \in [0,1]$ is chosen which represents the respective quantile value τ.

This error function with $e = o - \hat{y}^{(\tau)}$ and observation o is defined as

$$\rho_\tau(e) = \begin{cases} \tau \;\; |e| & ,\text{if } \; e \geq 0, \\ (1-\tau) \; |e| & ,\text{if } \; e < 0, \end{cases} \tag{5.24}$$

as shown in Fig. 5.11. A parameter of $\tau = 0.5$ is equivalent to the conventional MAE function. When using Eq. 5.24 as loss function during model training, the model minimizes the loss with respect to the utilized loss function (the pinball function reaches its minimum value

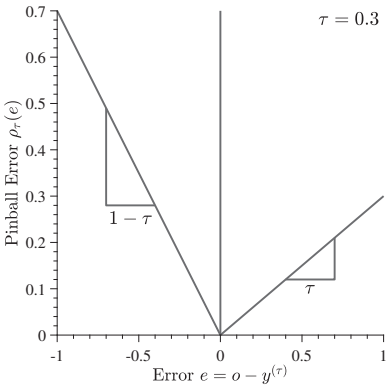


Figure 5.11: Visualization of the pinball function that is used for quantile regression. For quantile regression, an adapted form of the MAE is utilized. It computes a skewed error depending on the chosen value of τ.

in the τ quantile of a distribution of values which we have shown in the proof denoted in Appendix E). For the quantile score, this is achieved when the forecasting model creates the quantile forecast $\hat{y}^{(\tau)}$ of the distribution of uncertain observations (instead of, e.g., the median forecast). After model training, the power forecasting model is then optimized to create a quantile forecast, i.e., the value of $\hat{y}^{(\tau)}$. If this process is repeated for a number of L forecasting models, each trained using a different value of τ in the error function ρ_τ, the entirety of trained models then can create a tuple of quantile forecasts $\left(\hat{y}^{(\tau_1)}, \ldots, \hat{y}^{(\tau_L)}\right)$. A cdf can be created from the predicted quantile forecasts using Eq. 5.7.

QR is relatively intuitive, as models for deterministic forecasts can be used directly with the modified cost function. However, QR techniques typically use regularization techniques (that penalize too close positions of adjacent quantile forecasts) to avoid "line crossing" of neighboring quantile functions (as, per definition, the quantile forecasts are not allowed to cross). The phenomenon of line crossing is visualized in Fig. 5.12. The figure shows a set of quantile regression models for different τ values that form an underconfident predictive distribution (too much spread). Line crossing effects are visible in the upper right corner.

The idea of quantile regression is laid out in [137], while [176] extends the approach to nonlinear quantile regression functions with application to probabilistic wind power forecasting. In [106], quantile regression for wind power forecasting is used in conjunction with a fuzzy ARTMAP network trained on decomposed Wavelet features.

5.6.5 Prediction Interval Forecasting

Prediction interval (PI) forecasting is an uncertainty estimation technique which creates intervals to estimate the probability distribution. An overview of the area of interval forecasting techniques can be found in [134, 240]. In this section, the main principle of PI forecasting is explained, as well as the relation of PIs to the representation using quantiles. In PI forecasting, an algorithm tries to forecast a conditional interval I_α given a nominal confidence $(1-\alpha)$ with $\alpha \in [0,1]$ which specifies the desired probability of an observation occurring inside the

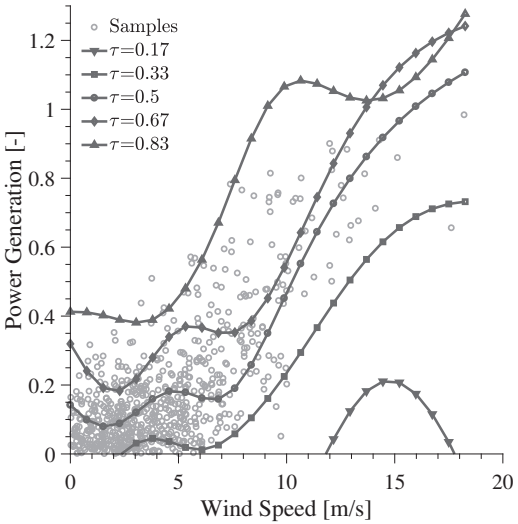


Figure 5.12: Set of five quantile regression models that lead to an underconfident predictive distribution (too much spread). Line crossing effects are visible in the upper right corner. Per definition, quantiles are not allowed to cross. If regularization techniques are included in the training process, this phenomenon can be suppressed.

interval

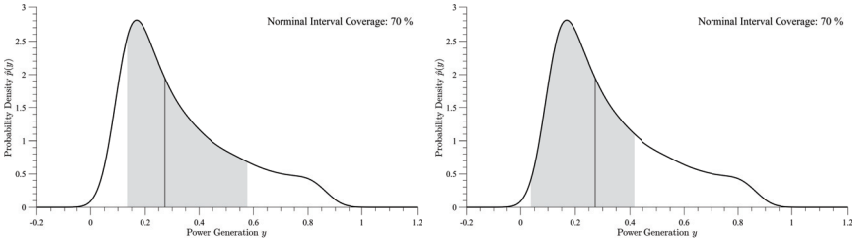
$$I_\alpha = [\hat{l}, \hat{u}], \tag{5.25}$$

where $\hat{l}$ is the lower and $\hat{u}$ is the upper bound of the respective prediction interval. The narrowness of a PI can then be denoted as $\hat{u} - \hat{l}$. For interval training, an objective function, the *interval score* (IS) [240], is defined. It has the basic form

$$\mathrm{IS}_\alpha = (\hat{u} - \hat{l}) + \frac{2}{\alpha} \cdot (o - \hat{u}) \cdot H(o - \hat{u}) + \frac{2}{\alpha} \cdot (\hat{l} - o) \cdot H(\hat{l} - o), \tag{5.26}$$

where H is a Heaviside step function. The first term rewards narrowness of the PI, while the second and third term penalize out-of-interval observation occurrences. During model training, the overall forecasting technique is then optimized on a data set to minimize the IS.

PIs do not necessarily have to be centered (regarding the probability mass) around the median of the underlying probability distribution of the predictand. While two PIs may have the same nominal confidence $(1 - \alpha)$, they can nevertheless be positioned in different ways within the pdf. This phenomenon is visualized in Fig. 5.13. The figure shows two PIs with same nominal coverage probability ($\hat{P}(\hat{u}) - \hat{P}(\hat{l})$, 70 % in the present case), where Fig. 5.13.1 shows a centered PI and Fig. 5.13.2 displays an uncentered PI. While both PIs clearly give an indication on the expected number of observations within the PI, they do not specify the position of the PI within the underlying pdf. Thus, the directional fraction of samples *outside* the PI can only be derived from a centered PI. However, we have shown (see the proof conducted in Appendix G) that a PI forecasting technique can typically yield centered prediction intervals when it is optimized on the IS (see Eq. 5.26) as the score reaches the minimum value at the



(1) Centered prediction interval.

(2) Uncentered prediction interval.

Figure 5.13: Example of two prediction intervals (PI) with nominal coverage probability $\hat{P}(\hat{u}) - \hat{P}(\hat{l})$ of 70 %. The black continuous line represents the overall predictive pdf $\hat{p}(y)$, while the shaded area represents the coverage area of the PI. The vertical grey line denotes the median of the pdf, i.e., the location of $\hat{y}^{(0.5)} = \hat{P}^{-1}(0.5)$. The example shows the difficulties with uncentered PIs. While both example PIs have the same nominal coverage probability, they are positioned differently within the overall pdf. Using an uncentered PI, while the fraction of observations within the PI is known, the quantity and direction (under- or overestimation) of the elements outside the PI are unknown in the interval representation. Having a centered PI, the interval can be represented as quantile forecasts, which do not suffer from this effect.

location of the centered PI given a nominal confidence $(1-\alpha)$.

In the literature, intervals are often seen as a form of uncertainty representation which is independent from the representation as a predictive distribution. Thus, they have a different form of quality assessment (e.g., PI nominal confidence, PI coverage probability, or the interval score [95, 259]) and are, therefore, not directly comparable to other forms of uncertainty representation. As pointed out in [191], predictive distributions can be created from PIs (and vice versa). The interval bounds $\hat{l}, \hat{u}$ can be interpreted as quantiles, which leads to the definition of an interval $I^{(\alpha)}$ of

$$I_\alpha = [\hat{y}^{(\tau_{\hat{l}})}, \hat{y}^{(\tau_{\hat{u}})}], \text{ with} \tag{5.27}$$

$$\hat{y}^{(\tau_{\hat{l}})} = \hat{P}^{-1}(\tau_{\hat{l}}) = \hat{l}, \tag{5.28}$$

$$\hat{y}^{(\tau_{\hat{u}})} = \hat{P}^{-1}(\tau_{\hat{u}}) = \hat{u}, \tag{5.29}$$

$$(1-\alpha) = \tau_{\hat{u}} - \tau_{\hat{l}}. \tag{5.30}$$

Assuming centered prediction intervals, i.e., the center regarding the probability mass of the intervals is the median forecast which can be achieved by model training using the IS (as also discussed in [190]), the relationship between the nominal coverage rate and quantile position is

$$\tau_{\hat{l}} = 1 - \tau_{\hat{u}} = \frac{\alpha}{2}, \tag{5.31}$$

which is the only way to directly estimate the quantile values without knowing the complete probability distribution in the first place (the black continuous line in Fig. 5.13). In fact, the IS only is a version of two pinball functions (Eq. 5.24) scaled with $\frac{\alpha}{2}$ with relationship

$$\frac{\alpha}{2} \cdot \mathrm{IS}_\alpha(o) = \rho_{\tau_{\hat{l}}}(o - \hat{y}^{(\tau_{\hat{l}})}) + \rho_{\tau_{\hat{u}}}(o - \hat{y}^{(\tau_{\hat{u}})}). \tag{5.32}$$

We conducted the proof of this relationship in Appendix F. The resulting centered PI can then

be interpreted as a predictive cdf in the same way as quantile forecasts using Eq. 5.7, where the quantile forecasts in turn are constructed using Eqs. 5.28 and 5.29. The repeated computation of centered PIs for different coverage rates can lead to more detailed forms of the cdf which may be needed for some decision-making problems as described in [191]. Uncentered PIs, however, are not well suited for the construction of a predictive density function, as the PI may be positioned differently within the pdf for each value of the nominal confidence. In the literature, the interval bounds $\hat{l}, \hat{u}$ are estimated using autoregressive models [103], other approaches perform interval forecasting using artificial neural networks [198], or extreme learning machines [239], and optimize the model parameters using optimization algorithms that are not based on gradients (as IS is not continuously differentiable), e.g., using particle swarm optimization (PSO). As can be seen from the types of algorithms used (e.g., PSO), PI algorithms often use computationally expensive methods with long training times.

5.7 Predictive Distribution Construction from Ensemble Predictors

Many probabilistic forecasting algorithms are based on ensemble forecasts, which is an umbrella term for the aggregation of multiple forecasts to an overall forecast. Where Section 5.6 described the creation of predictive distributions from single NWP models, this section details the creation of predictive distributions from multiple NWP models.

Ensembles can be formed in a number of ways and we will highlight the most popular forms in the following. While technically every ensemble principle (such as EPS, MME, TLE, or PFE, see Section 2.6.3 for details) can be used for the different forms of predictive distribution creation, it often only makes sense to use a single type of ensemble for a particular technique.

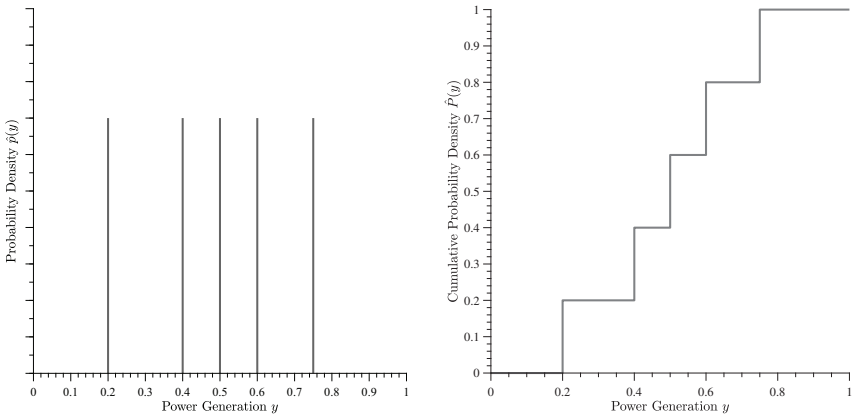
5.7.1 Density Function Sampling (EPS Ensemble)

As previously mentioned, an NWP can be regarded as a sample from the underlying (non-observable) probability distribution of possible weather situations for a certain point in time. Some forecasting systems draw multiple samples from the underlying probability distribution which are typically ensemble prediction systems (EPS) [108] or scenario forecasting systems [192]. Given a number of J sampled NWP forecasts, a deterministic power forecast for each NWP can be created, forming a set of power forecasting values $\hat{y}_1, \ldots, \hat{y}_J$. These power forecasts in turn are samples of an underlying predictive distribution in the power generation space, they are assumed to be independently and identically distributed (i.i.d.).

Unlike quantile forecasts as used in the construction of the predictive distributions (see Section 5.2) these forecasts do not provide information regarding their position within the density function as they are random samples of the underlying probability distribution. Thus, they also do not have a defined probability mass between the power forecasting values in the ensemble. Instead, they are equally important realizations of the stochastic process. Therefore, a density function can then be estimated from a set of power forecasting values in the form of a number of δ functions which form the overall pdf in the form

$$\hat{p}(y) = \frac{1}{J} \sum_{j=1}^{J} \delta(y - \hat{y}_j). \tag{5.33}$$

This form of density function is sketched in Fig. 5.14.1. The corresponding cdf consequently



(1) Raw ensemble members. (2) Step cdf from EPS.

Figure 5.14: Cdf construction using the EPS sampling technique. Fig. 5.14.1 shows the raw distribution of forecasts of the ensemble members. Using a direct sampling strategy from an EPS, the cdf can be constructed in the form of a step function (Fig. 5.14.2). This representation is often used for the uncertainty representation of ensemble prediction systems or scenarios.

is defined as

$$\hat{P}(y) = \frac{1}{J} \sum_{j=1}^{J} H(y - \hat{y}_j), \tag{5.34}$$

where H is a Heaviside step function and the single power forecasts serve as increments for the overall cdf. An example for this form of cdf is also shown in Fig. 5.14.2.

This form of cdf representation is popular in the area of meteorological sciences, in particular for a probabilistic representation of ensemble prediction systems (EPS). While the cdf is simple in its structure and has a clear motivation when using samples from an EPS, it has some weaknesses in its marginal quantiles. For instance, an EPS assumes that the probability of an observation lying below the lowest EPS member is 0, which, clearly, is not the case in practice. Therefore, step cdfs are sometimes post-processed to include an "out-of-sample" probability.

5.7.2 Distribution Fitting / Model Output Statistics

A frequently applied technique for the creation of a predictive distribution from an ensemble is to fit a parametric distribution to the ensemble forecasts $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_J)$. Some authors frame this technique as the correction of the model output of a predictive distribution which may be subject to model biases and dispersion errors (rather than the construction of a predictive distribution). These types of errors may be due to (seasonal) sampling effects, long-term non-stationary (i.e., the systematics of the distribution of the data changes over time, e.g., due to changing climate), underdispersive, or biased EPS forecasts. This effect can be corrected using some form of statistical post-processing (also called model calibration) in order to correct for reliability. The main principle of model calibration is to exploit the structure of past forecast-observation pairs in order to correct systematic errors in the output of the model. This process

is often also called (ensemble) model output statistics, or (E)MOS. Conventional MOS enable a bias correction of a deterministic model. Other related approaches for probability calibration are sigmoid calibration [195], or isotonic regression techniques [255].

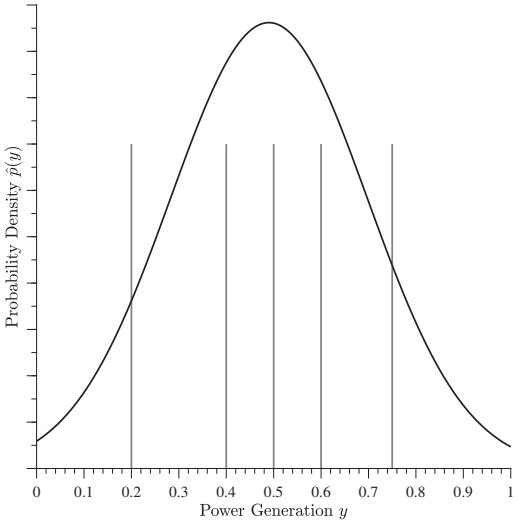


Figure 5.15: Distribution fitting fits a parametric density function over the entirety of ensemble members using linear correction terms for distribution mean and spread.

EMOS, on the other hand, create a probabilistic density forecast. This process makes the assumption that the ensemble members are independent samples from the same probability distribution of possible outcomes. The distribution fit is then constructed in the way

$$\hat{p}(y) = K\Big(\frac{y-\mu}{\sigma}\Big), \tag{5.35}$$

where μ is the EMOS center, and σ is a spread parameter. K is a normalized kernel function that conforms to Eq. 5.22 with $\int_{y=-\infty}^{+\infty} K(y)\mathrm{d}y = 1$. The parameters are then defined using

$$\mu = a_1 + a_2 \cdot \bar{\hat{y}}, \tag{5.36}$$
$$\sigma = b_1 + b_2 \cdot \mathrm{Std}(\hat{\mathbf{y}}), \tag{5.37}$$

with $a_1, a_2, b_1, b_2 \in \mathbb{R}$ and $\sigma > 0$. Therein, $\bar{\hat{y}}$ is the ensemble mean and $\mathrm{Std}(\hat{\mathbf{y}})$ is the empirical standard deviation of the ensemble. The parameters a_1, a_2 define a linear relationship between ensemble mean and optimal density function center, whereas the parameters b_1, b_2 model a linear relationship between ensemble spread and the spread of the fitted density function. A visualization of the resulting probability distribution created using EMOS is visualized in Fig. 5.15.

A simple technique with fixed a_1=0, a_2=1, b_1=0, b_2=1 (and, thus, no model training) is proposed in [249]. The authors of [96] introduced a variant for the optimization with respect to the CRPS score (see Section 6.1.1). In [23], the performance of distribution fitting approaches is compared to ensemble dressing approaches (see Section 5.7.3).

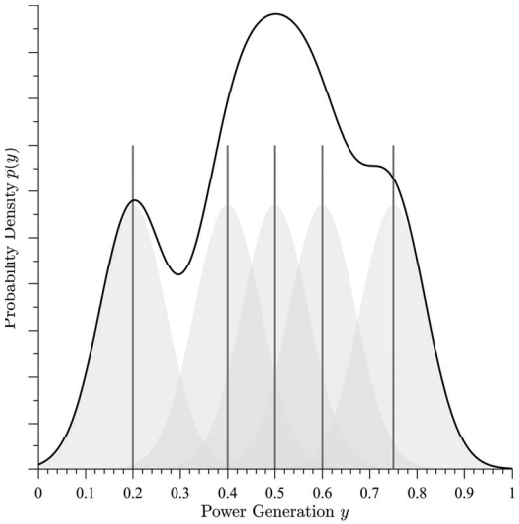


Figure 5.16: Ensemble dressing approaches represent the overall density function as a combination of components that are defined by a set of individual (deterministic) forecasts.

5.7.3 Ensemble / Kernel Dressing

Ensemble or kernel dressing is a technique which is related to kernel density estimation (KDE). However, a number of ensemble forecasts are used as basis for kernel dressing rather than a set of historic power measurements. The working principle is illustrated in Fig. 5.16. In some contexts, this process is also referred to as statistical post-processing. The basic idea is to construct a predictive density function using

$$\hat{p}(y) = \sum_{j=1}^{J} \left(\pi_j \cdot K\Big(\frac{y-\mu_j}{\sigma}\Big)\right), \tag{5.38}$$

from J ensemble members with kernel K, center μ_j, bandwidth parameter σ, and weighting coefficient π_j that also ensures conformity with Eq. 5.1 with $\sum_{j=1}^{J} \pi_j = 1, \pi_j \geq 0$. While it is on a principle level also possible to adjust the parameter σ for each of the j base predictors individually, this is typically not performed in the literature but only estimated as a function of the ensemble members as laid out below.

A framework for the categorization of Kernel dressing methods is laid out in in [23] with the categories

- (Standard) Kernel Dressing (SKD) [58],
- Bayesian Model Averaging (BMA) [199], and
- Affine Kernel Dressing (AKD) [23].

The methods differ in the way the parameters μ_j and σ are constructed. SKD determines the

parameters using the equations

$$\mu_j = \hat{y}_j + a_1, \tag{5.39}$$
$$\sigma = b_2 \cdot \mathrm{Std}(\boldsymbol{\mu}), \tag{5.40}$$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)$, where a_1, b_2 are parameters to be optimized. SKD thereby enables a bias correction of the ensemble members by modifying a_1, and is able to vary the bandwidth parameters σ as a function of the spread of the ensemble members and parameter b_2. In SKD, a common choice for the weighting coefficients is to weight them equally. BMA extends the correction term to

$$\mu_j = a_2 \cdot \hat{y}_j + a_1, \tag{5.41}$$
$$\sigma = b_1. \tag{5.42}$$

where the parameter σ is trained directly. Therein, a_1, a_2 can be optimized using linear regression in a first step. Afterwards, the parameter σ and the individual weights π_j are subsequently trained using an expectation maximization algorithm as explained, e.g., in [199].

Besides this correction term, BMA has a different perspective with respect to the weighting coefficients π_j by an application in a Bayesian framework: The BMA prediction is the average of individual forecasts weighted with the likelihood that an individual model is correct given a set of ensemble point forecasts.

BMA is considered to be superior to SKD if the expected qualities of the models differ, e.g., for multi-model ensembles. Such an extension of the combination of a multi-model ensemble into a BMA framework is described in [150]. In [59], a multi-model ensemble for hydrologic predictions is created using BMA to optimally account for the dispersion of the forecast. Multivariate tempo-spatially consistent calibration techniques are often based on Gaussian copulas [217], which include BMA as a prerequisite for copula creation.

AKD, on the other hand, interprets ensembles as sources of information rather than a set of possible outcomes, where one outcome is assumed to be true (such as is the assumption in BMA approaches). It once more extends the correction term to

$$\mu_j = a_3 \cdot \bar{\hat{y}} + a_2 \cdot \hat{y}_j + a_1, \tag{5.43}$$
$$\sigma = b_1 + b_2 \cdot \mathrm{Std}(\boldsymbol{\mu}), \tag{5.44}$$

and thereby combines the elements of SKD and BMA while including a term based on the ensemble mean $\bar{\hat{y}}$ as described in [23]. A case study for the prediction of regional climate change using ensemble dressing extended with temporal autocovariance is conducted in [218] and includes multivariate dressing functions and AKD. A discussion on the differences of ensemble dressing and KDE is conducted in [12]. As described in [259], for bounded variables, such as power generation or wind speed, it may be advantageous to use Beta or Gamma kernels instead of a Gaussian.

5.7.4 Forecasting the Skill Category from Risk Indices

Skill category forecasting is a technique for uncertainty assessment. It creates categories of estimated skill (estimated quality of the prediction) that are determined using a measure of risk of a prediction. This risk index can be computed a priori, i.e., the observation (true power measurement) is not needed for the risk index computation. The two most popular forms of risk indices are the *normalized risk prediction index* (NRPI) and the *meteo-risk index* (MRI).

The NRPI computes the risk based on the spread of an ensemble forecast [189], while the MRI directly evaluates the NWP forecast regarding wind speed differences on lagged time horizons [188].

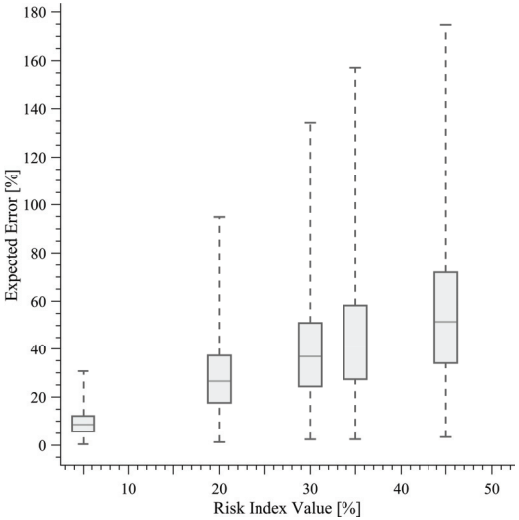


Figure 5.17: Skill Score - Error Quantiles diagram, which shows the expected error per skill category. The skill category depends on the expected risk of a forecast.

The risk index values are often discretized to represent a number of easily interpretable *skill categories*, e.g., 5 categories from 1 (very certain) to 5 (very uncertain), or color codes (green, yellow, and red). Having a historic data set with N data points from which the corresponding tuples of forecasting errors $\mathbf{e}$ and risk indices $\mathbf{r}$ with $\mathbf{e}, \mathbf{r} \in \mathbb{R}^N$ have been computed, the risk categories can be created in the following way. The risk index values are sorted and p skill categories are defined on the risk index values (using skill category borders between the skill categories) in a way that the categories are equally populated (each with $\approx \frac{N}{p}$ samples).

Then, a distribution of the errors is created from the subset of corresponding forecasting errors which exist in each of the p risk categories. An example for a number of five skill categories from risk indices with corresponding error estimates is shown in Fig. 5.17. The horizontal axis denotes the value of the risk index, which is binned to achieve five equally populated skill categories in the shown case. Then, on the ordinate the expected errors (MAE) for each skill category are computed from a quantile estimation as visualized by the box plots.

After this training process, when operationally performing a forecast, the risk index is computed a-priori for the current prediction, then the conditional distribution of the errors of the matching risk category is used as uncertainty estimate. When the resulting error distribution is assumed as a symmetrical interval, the forecasting pdf can be created, e.g., using Eq. 5.7.

5.8 Summary of this Section

This chapter summarizes the most widely used forms of representation of probabilistic forecasts, shows required properties of probabilistic forecasting models, and highlights the techniques that are used to create these representations in the first place. The main categorization used to classify techniques for the creation of probabilistic forecasts is whether the probabilistic forecast is created using a single predictor model (a single weather model for power forecasting) or using an ensemble as basis for the creation of the probability distribution. The main approaches that use single predictor models are parametric density functions, kernel density forecasts, analog ensembles, quantile regression, and prediction interval techniques. Techniques that are created from ensembles are either created using sampling approaches (e.g., when using an ensemble prediction system), distribution fitting or ensemble dressing techniques, or the use of skill categories. It may be noted that, while utilized on a regular basis in the literature, the use of kernel methods for ensemble forecasting is somewhat sloppily defined, as kernel functions typically measure the similarity between data points and do not make guarantees about the probability mass. Many authors in the literature furthermore use techniques from conventional distribution estimation algorithms (such as the maximum likelihood algorithm), it may therefore make sense to define ensemble dressing techniques using entirely the nomenclature of these techniques.

Besides a structured overview of existing techniques, as main contribution a scheme for a common representation of probabilistic forecasts from different forms of representations is proposed. This in particular is relevant for representations that are not naturally represented as a probability distribution, such as regression quantiles or prediction intervals. The idea of the common representation is to convert each form of representation to a probability distribution. Having a common representation of probabilistic forecasts, the evaluation of each is score is better comparable, as the same score can be used for the evaluation of all representations of probabilistic forecasts.

With respect to the overall goals of this thesis, the introduced common scheme for the representation of probabilistic forecasts can be used to enable an easy combination of multiple forecasts to an overall refined forecast in an ensemble that is independent from the particular probabilistic forecasting technique. This is further exploited in Chapter 7, where a novel ensemble technique for probabilistic forecasts is presented that dynamically aggregates probabilistic forecasts. The ensemble technique is based on the assumption that all base predictors are represented in the form of a probability distribution.

Chapter 6

Evaluation Metrics for Probabilistic Forecasts

Error scores for the performance assessment of probabilistic forecasts are frequently referred to as *scoring rules*. In comparison to deterministic error measures, many probabilistic scoring rules lack intuition. Where deterministic forecasts need to be close to observations, probabilistic forecasts need to concentrate the probability mass close to the observations (they have to be sharp) *and* they have to correctly assess the spread of the probability distribution (commonly referred to as reliability) depending on the amount of uncertainty [94] regarding the power generation as has been detailed in Section 5.3. Intuition rises by understanding how different types of errors affect scoring rules. In order to compare the performance of probabilistic forecasting techniques, there has to be a clear definition of how the process of quality assessment is performed. The most general approach to evaluate probabilistic forecasts is using scoring rules which compare a predictive distribution to an actual observation. Sections on probabilistic forecast evaluation for wind power forecasting are included in [94, 129, 259].

The main contribution of this chapter is an investigation of uncertainty assessment techniques for probabilistic forecasts including an analysis of the decomposition properties. Decomposition of scoring rules refers to the partitioning of a score into its integral parts, which are identified to be a reliability term, a part that accounts for the uncertainty of a forecast, and a third term that is called resolution. In a number of case studies, the characteristics of the presented scoring rules are analyzed in detail. This process is defined as *metaverification* in [168], which describes the evaluation of performance measures and lays out desirable properties of scoring rules such as the characteristic of being *proper* [24] and the robustness to *hedging* [126], both of which are further detailed in the following sections. From the insights of the case studies, advantages and limitations in the application of each error score are discussed.

First, Section 6.1 highlights ways to numerically assess the quality of probabilistic forecasts using scoring rules for probabilistic forecasts. In Section 6.2, the properties of the presented uncertainty assessment techniques are investigated in a number of experiments. Finally, our insights are discussed and summarized in Section 6.3.

6.1 Scoring Rules and Score Decomposition

This section introduces the concept of numerically evaluating probabilistic forecast-observation pairs. General aspects of scoring rules and of the decomposition of scoring rules are discussed

in Section 6.1. In Sections 6.1.1–6.1.3, the most widely used scoring rules for which a decomposition has been proposed are portrayed.

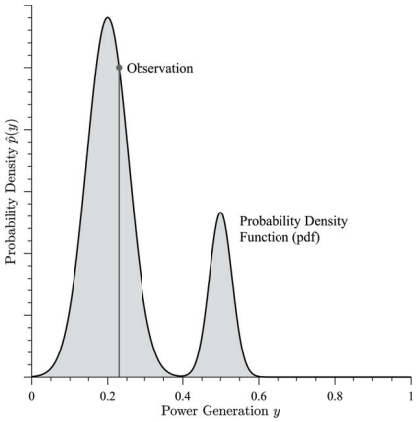
For the assessment of the quality of a forecasting algorithm, the forecasted predictive distribution is compared to the corresponding "true" observation which the predictive distribution tried to predict. In probabilistic time series forecasting, the main difference to the conventional comparison of distributions is that the predicted distribution typically is time-variant, and thus, differs for each evaluated time step. The error scores suited for this type of comparison are called *scoring rules*. A scoring rule $S(\hat{p}, o)$ scores a predictive distribution $\hat{p}$ with respect to a corresponding observation o.

Table 6.1: Overview of the most common scoring rules for continuous variables. In the interval column, the underlined value indicates the optimum value of the score.

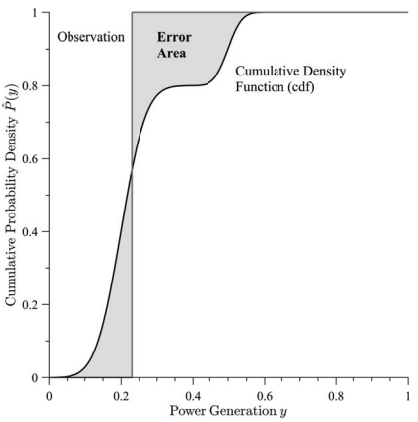
#	Scoring Rule Name	Computation		Local	Proper	Interval	Remark
1	Linear Score	LS	$= \hat{p}(o)$	✓	✗	$[0, \underline{+\infty}]$	intuitive, but not proper
2	Quadratic Score	QuadS	$= 2\hat{p}(o) - \int_{-\infty}^{+\infty} \hat{p}(y)^2 \mathrm{d}y$	✗	✓	$[0, \underline{+\infty}]$	varying definitions in [94], [95]
3	Spherical Score	SphS	$= \hat{p}(o) / (\int_{-\infty}^{+\infty} \hat{p}(y)^2 \mathrm{d}y)^{\frac{1}{2}}$	✗	✓	$[0, \underline{+\infty}]$	realization of pseudospherical sc.
4	Ignorance Score	IGN	$= -\ln \hat{p}(o)$	✓	✓	$[\underline{-\infty}, +\infty]$	see Sec. 6.1.2
5	Continuous Ranked Ignorance Sc.	CRIGN	$= -\int_{-\infty}^{+\infty} \ln\lvert(\hat{P}(y) - \big(1 - H(o-y)\big)\rvert \mathrm{d}y$	✗	✓	$[\underline{-\infty}, +\infty]$	see Sec. 6.1.2
6	Continuous Ranked Probability Sc.	CRPS	$= \int_{-\infty}^{+\infty} (\hat{P}(y) - H(y-o))^2 \mathrm{d}y$	✗	✓	$[\underline{0}, +\infty]$	see Sec. 6.1.1
7	Quantile Score	QS	$= \rho_\tau(\hat{P}^{-1}(\tau) - o)$	✗	✓	$[\underline{0}, +\infty]$	see Sec. 6.1.3
8	Interval Score	IS	$= (\hat{u} - \hat{l}) + \frac{2}{\alpha}(o - \hat{u})H(o - \hat{u}) + \frac{2}{\alpha}(\hat{l} - o)H(\hat{l} - o)$	✗	✓	$[\underline{0}, +\infty]$	see Sec. 5.6.5
9	Dawid-Sebastiani Score	DSS	$= \frac{(o - \mathbb{E}[\hat{p}(y)])^2}{\mathrm{Std}[\hat{p}(y)]^2} + 2\ln(\mathrm{Std}[\hat{p}(y)])$	✗	✓	$[\underline{0}, +\infty]$	moment-based repr. of $\hat{p}(y)$
10	Predictive Model Choice Criterion	PMCC	$= -(o - \mathbb{E}[\hat{p}(y)])^2 - \mathrm{Std}[\hat{p}(y)]^2$	✗	✗	$[-\infty, \underline{0}]$	moment-based repr. of $\hat{p}(y)$
11	Hyvärinen Score	HS	$= 2\frac{\hat{p}''(o)}{\hat{p}(o)} - \big(\frac{\hat{p}'(o)}{\hat{p}(o)}\big)^2$	✓	✓	$[\underline{-\infty}, +\infty]$	$\hat{p}'(o)$, $\hat{p}''(o)$ indicate derivations

Table 6.1 gives an overview of the most common scoring rules, which are partly detailed in [94, 95, 127]. In this table, the computation of the scores is given, as well as an indication whether the score is a proper score, and if it is local. One of the most important criterions of a scoring function for uncertainty assessment is that it is *proper*. Being (strictly) proper means that a score reaches its optimum value if (and only if) a predicted distribution exactly matches the "true" (non-observable) distribution to compare it to. Proper scores are nontrivial to *hedge*, which means that a predictive distribution cannot be created in a way that exploits a systematic weakness in the score definition to achieve a better score (as described, e.g., in [126]). As an example for hedging a score we use the PMCC score (#10 in Table 6.1). The score consists of two terms, one element that penalizes the squared difference of the expectation value to the actual observation and one term that penalizes the standard deviation of the predictive distribution (which in theory should encourage sharper predictions). The closer the score is to a value of zero, the better. If we wanted to hedge this score, we could simply always issue a overly sharp predictive distribution (where $\mathrm{Std}[\hat{p}(y)] \to 0$), which would yield a better score without any side effects as the expectation value $\mathbb{E}[\hat{p}(y)]$ remains unchanged. Therefore, the score does *not* yield its minimum value if the predicted distribution matches the "true" distribution, thus it is not proper. For more details on propriety, we refer to [24, 193].

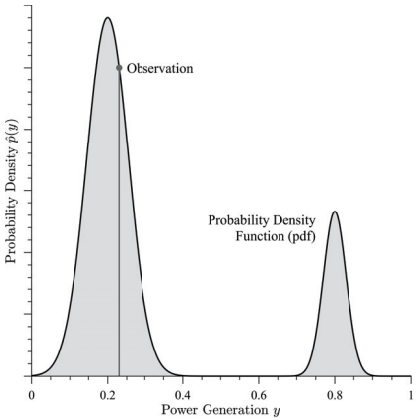
While propriety is universally considered a desirable property, there is a broad discussion whether scoring functions should be *local*. A local score only evaluates the probability density at the location of the observation o. This property is visualized in Fig. 6.1. The figure shows two pdfs with corresponding cdfs, which are given by mixture models of two Gaussians. The left Gaussians of the two mixture models are identical, while the right component is shifted in one of the pdfs. A local score, e.g., the IGN score, is agnostic of the form of the distribution, as it only evaluates the probability density at the location of the observation



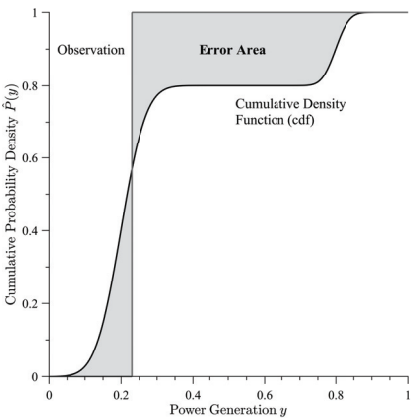
(1) Pdf of two close Gaussians.



(2) Cdf of the pdf of Fig. 6.1.1.



(3) Pdf of two distant Gaussians.



(4) Cdf of the pdf of Fig. 6.1.3.

Figure 6.1: Illustration of the locality principle. Figs. 6.1.1 and 6.1.3 show two predictive pdfs with corresponding observation. The two pdfs are constructed from a mixture model of two Gaussians, where the left component is identical in both pdfs. The right component is shifted by a different amount in the two pdfs. A local score (such as the IGN score) *only* evaluates the probability density at the location of the observation (dark grey line). Therefore, the error score of IGN is (virtually) identical in both pdfs. However, using a non-local score (e.g., the CRPS), the corresponding cdfs are evaluated instead of the pdfs. The shaded areas in Figs 6.1.2 and 6.1.4 indicate the error area of the CRPS which correspond to the two pdfs. As can be seen from the figures, non-local scores evaluating this error area yield different errors for the two variants of the distribution.

(such as shown in Figs. 6.1.1 and 6.1.3). In contrast, non-local scores also consider the concentration of probability mass around the location of the observation. Non-local scores have the intuitive appeal that a (correct) narrower forecasting distribution should be rewarded, as shown in the example of Fig. 6.1. However, in the literature it is often argued that, given two predictive distributions, a non-local score can in principle lead to a worse score for the predictive distribution that has a higher probability density for the target observation. Further arguments for and against local scores are given, e.g., in [8, 24, 160]. As can be seen from Table 6.1, scoring rules have varying propriety and locality properties. Furthermore, they differ in terms of the evaluated function (pdf, cdf), whether they evaluate the overall density function (e.g., the scores 1 – 6 in Table 6.1), or sampled cdf values (e.g., QS and IS), and whether they only evaluate moments of the pdf (e.g., DSS, PMCC). The energy score (not shown in the table) is a multivariate generalization of the CRPS score [95]. It is therefore often used in the evaluation of scenarios [98, 186] which is detailed in Section 2.2. The authors of [94] relate deterministic error scores (see Section 3) to the framework of (probabilistic) scoring rules using the term *consistent scoring functions*, which form a special case of (probabilistic) scoring rules for point forecasts.

For the evaluation of a dataset with $n = 1, \ldots, N$ forecast-observation pairs, for all scoring rules the score can be performed using an averaging in the form

$$\bar{S} = \frac{1}{N} \sum_{n=1}^{N} S(\hat{p}_n(y), o_n). \tag{6.1}$$

For quality estimation, scoring functions often are compared to a reference forecast which serves as baseline (such as a climatological forecast) using the *skill score* in the way

$$\mathrm{SS} = 1 - \bar{S} / \bar{S}_{\mathrm{ref}}. \tag{6.2}$$

These skills have a range of $[-\infty, 1]$, a score of 1 would be a perfect result, while a score less than zero can be considered "unskillful", as it performs worse than the baseline technique.

Scoring functions assess the overall skill of a forecasting system. However, if a more detailed analysis regarding reliability and sharpness (see Section 5.3 for an explanation) is desired, some scoring rules have the ability to be decomposed into the three components *reliability*, *resolution*, and *uncertainty* in the form

$$\bar{S} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}. \tag{6.3}$$

The reliability property is detailed in Section 5.3. Resolution describes the ability of a power forecasting model to issue a forecast different from the mean observation of the evaluated time period. It thereby heavily depends on the uncertainty, the average (unconditional) spread of the observations in the dataset of the evaluated time period. In some cases, the term UNC – RES is also defined as the *potential* score, e.g., in [108, 159]. The potential score is "resolution accounted for uncertainty", it thereby is more closely related to the sharpness property. In a reliable forecasting system, an increase in resolution also leads to an increase of sharpness [93]. Using a decomposition of error scores, the *reasons* for an error are better visible (i.e., whether it is due to an unreliable system or a lack of sharpness). In the following, we will highlight the

- continuous ranked probability score (CRPS) [108, 161],
- ignorance score (IGN) [229, 246], and

- quantile score (QS) [10]

in Sections 6.1.1 – 6.1.3, which are the most popular scoring rules for which a decomposition has been proposed in the literature.

6.1.1 Continuous Ranked Probability Score (CRPS)

The continuous ranked probability score (CRPS) probably is the best-known measure to assess the quality of a predictive distribution. It was first introduced in [161]. It is the continuous ranked version of the binary Brier score (BS). The idea of the CRPS is to assess the difference of the area between two cumulative density functions (cdf). The CRPS is computed by

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} \left(\hat{P}(y) - H(y-o)\right)^2 \mathrm{d}y, \tag{6.4}$$

where $\hat{P}(y)$ is the cdf of the predictive distribution at a particular point in time and $H(y-o)$ is the corresponding cdf of the observation, which is a Heaviside step function that has its step at the observation location. The value of CRPS is minimal if the uncertainty of the predictive distribution matches the spread of the observations on average. If a deterministic forecast is evaluated with the CRPS, the error function becomes equivalent to the mean absolute error (MAE) as the cdf of the deterministic forecast is a step function in this case. The CRPS can be interpreted as the integral of Brier scores for all possible threshold values, which also leads to its decomposition as described in [229]. To sketch this idea, we need to define the (binary) Brier score (BS) with probability $c^{(n)}$ of an event occurring, observation $z^{(n)} \in \{0,1\}$ of whether the event occurred and N forecast observation pairs with

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}\left(c^{(n)} - z^{(n)}\right)^2. \tag{6.5}$$

If we want to apply this to the continuous domain, we need to introduce a threshold a that binarizes the forecast to get the probability $c^{(n)} = P(1-\hat{P}_n(a))$ with $\hat{P}_n(a)$ being the cdf of the continuous forecast evaluated at threshold a and $z^{(n)} = H(o_n - a)$ to get the Brier score $\mathrm{BS}(a)$. In (continuous) power forecasting, this can be interpreted as the question what the probability is that the power generation is larger than a value a. This is then compared to whether the event $o_n > a$ actually occurred.

As has been described, e.g., in [250], the BS can be decomposed as

$$\mathrm{BS}(a) = \underbrace{\sum_i P(c_i)[c_i - \bar{z}_i]^2}_{\mathrm{REL}} - \underbrace{\sum_i P(c_i)[\bar{z}_i - \bar{z}]^2}_{\mathrm{RES}} + \underbrace{\bar{z}(1-\bar{z})}_{\mathrm{UNC}}, \tag{6.6}$$

where i is the index of I possible values with $\{c_1, \ldots, c_I\}$ that is composed of the unique values in the tuple $\left(c^{(1)}, \ldots, c^{(N)}\right)$. The value $P(c_i)$ denotes the frequency with which each category has been forecasted, $\bar{z} = P(z=1)$ is the *unconditional* (climatological) probability of the event occurring. The value $\bar{z}_i$ is the *conditional* probability of occurrence after having issued a specific forecast c_i with $\bar{z}_i = P(z=1|c_i)$. The first term (REL) is the average conditional bias (conditioned by each forecasted category c_i), RES is the average distance to the unconditional probability of occurrence (where a large difference generally is preferable). Finally, UNC describes the variance within the observations. This is also the score value that is achieved if the climatological forecast is created.

As has been defined in [108], the CRPS can be expressed as the integral of the BS over all thresholds a in the form

$$\text{CPRS} = \int \text{BS}(a)\, \mathrm{d}a. \tag{6.7}$$

The methodology for decomposition presented in [229] suggests to create a joint *ordered* set of thresholds $\{a_b\}$ which consists of issued forecasts $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N\}$ (shown here as a set of quantile forecasts $\hat{\mathbf{y}}_n = \{\hat{y}_n^{(\tau_1)}, \ldots, \hat{y}_n^{(\tau_L)}\}$ with L quantiles for each point in time $n \in 1, \ldots, N$) and observations $\mathbf{o} \in \{o_1, \ldots, o_N\}$ for each of the N evaluated data points with

$$\{a_b\}_{b=1,\ldots,B}, \quad a_b < a_{b+1}, \quad a_b \in \{\hat{\mathbf{Y}}, \mathbf{o}\}. \tag{6.8}$$

Though the problem domain is continuous, the set from which $\{a_b\}$ is constructed may contain duplicate elements (e.g., due to rounding effects) so that

$$B \leq (\underbrace{N \cdot L}_{\text{no. of elements in } \hat{\mathbf{Y}}} + \underbrace{N}_{\text{no. of elements in } \mathbf{o}}). \tag{6.9}$$

Having the ordered set $\{a_b\}$, we can create sections with *constant* Brier score (BS) that can be denoted as

$$\text{CRPS} = \sum_{b=1}^{B-1} (a_{b+1} - a_b) \cdot \text{BS}(\bar{a}_b) \tag{6.10}$$

with $\bar{a}_b = a_b + \frac{a_{b+1} - a_b}{2}$ being the mean of a_{b+1} and a_b. Using Eq. 6.6, we can then compute the decomposition elements of the CRPS with

$$\text{CRPS} = \sum_{b=1}^{B-1} (a_{b+1} - a_b) \cdot \big(\text{REL}(\bar{a}_b) - \text{RES}(\bar{a}_b) + \text{UNC}(\bar{a}_b)\big). \tag{6.11}$$

An alternative form of decomposition for the CRPS for ensemble prediction systems is shown in [108]. There also exist versions of the CRPS for categorical forecasting tasks (RPS) [229]. A variant of the CRPS for multi-model ensembles is discussed in [244], the authors furthermore introduce a variant of the CRPS, the CRPSS_D, which aims to eliminate some of the bias effects of the CRPS for EPS sampled distributions (Eq. 5.34) with small ensemble sizes. In [97], quantile weighted variants of the CRPS are used to put emphasis on certain parts of the predictive distribution.

6.1.2 Ignorance Score (IGN / CRIGN)

There are a number of scoring functions which emerged from information theoretic scores. "Classic" measures such as the Kullback-Leibler (KL) divergence can not directly be applied since the predictive distribution $\hat{p}(y)$ varies in each time step and can only be compared to a single respective observation o at that time. A popular scoring rule therefore is the ignorance score (IGN), which is a modified version of the KL divergence based on Shannon entropy [12]. IGN is defined by

$$\text{IGN} = -\ln \hat{p}(o), \tag{6.12}$$

where $\hat{p}(o)$ is the probability density value of the predictive distribution at the location of the observation. In contrast to CRPS, the ignorance score is a *local* score, i.e., it only evaluates the probability density $\hat{p}(o)$ at the location of the observation o and not the entire predictive

distribution. The ignorance score is referred to using a number of names, such as logarithmic score [95, 259], divergence score [245], or predictive deviance [136]. For binary predictions, such as the precipitation probability forecast, the cross-entropy score [181, 245] is highly related to ignorance. The decomposition of the ignorance score was performed in [229, 246]. For continuous forecasts, the decomposition of IGN is only possible in its continuous ranked form (CRIGN) computed by

$$\text{CRIGN} = -\int_{-\infty}^{\infty} \ln|\hat{P}(y) - \big(1 - H(y-o)\big)|\,\mathrm{d}y, \tag{6.13}$$

which loses its locality property, since the whole predictive distribution is considered instead of the local evaluation at observation o. The decomposition is nevertheless based on the IGN score, which for binary events is defined as

$$\text{IGN}(a) = \frac{1}{N}\sum_{n=1}^{N} \ln|c^{(n)} - (1 - z^{(n)})|, \tag{6.14}$$

again with $c^{(n)} = P(1 - \hat{P}_n(a))$ with $\hat{P}_n(a)$ being the cdf of the continuous forecast evaluated at threshold a and $z^{(n)} = H(o_n - a)$. As described in [229], it can be decomposed to

$$\text{IGN}(a) = \text{REL} - \text{RES} + \text{UNC}, \tag{6.15}$$

$$= \underbrace{\sum_i P(c_i)\left[\bar{z}_i \ln\frac{\bar{z}_i}{c_i} + (1-\bar{z}_i)\ln\frac{1-\bar{z}_i}{1-c_i}\right]}_{\text{REL}} \tag{6.16}$$

$$- \underbrace{\sum_i P(c_i)\left[\bar{z}_i \ln\frac{\bar{z}_i}{\bar{z}} + (1-\bar{z}_i)\ln\frac{1-\bar{z}_i}{1-\bar{z}}\right]}_{\text{RES}} \tag{6.17}$$

$$+ \underbrace{-\bar{z}\ln\bar{z} - (1-\bar{z})\ln(1-\bar{z})}_{\text{UNC}}, \tag{6.18}$$

where i is the set of I possible values with $\{c_1, \ldots, c_I\}$ that is composed of the unique values in the tuple $\left(c^{(1)}, \ldots, c^{(N)}\right)$. The value $P(c_i)$ denotes the frequency with which each category has been forecasted, $\bar{z} = P(z = 1)$ is the *unconditional* (climatological) probability of the event occurring. The value $\bar{z}_i$ is the *conditional* probability of occurrence after having issued a specific forecast c_i with $\bar{z}_i = P(z = 1|c_i)$.

Similar to the CRPS, the decomposition of the CRIGN is performed by integrating over all thresholds of the ignorance score IGN as shown in [229]. This can be written as

$$\text{CRIGN} = \int \text{IGN}(a)\,\mathrm{d}a \tag{6.19}$$

and can be decomposed in the same way as described in Eqs. 6.7 – 6.11.

6.1.3 Quantile Score (QS)

The quantile score (QS) or pinball loss proposed in [137] is a rather novel non-local scoring rule directly derived from the pinball function ρ_τ used for quantile regression (see Eq. 5.24). In the economics literature (e.g., in [97]), the quantile score function is also referred to as *tick* or *check* function. The quantile score was originally designed for the use with quantile regression

as presented in Section 5.6.4, however, it can be applied to any predictive distribution if the inverse cdf can be computed. The QS for a single quantile τ is defined as

$$\mathrm{QS}_\tau = \rho_\tau(\hat{P}^{-1}(\tau) - o), \tag{6.20}$$

where $\hat{P}^{-1}(\tau) = \hat{y}^{(\tau)}$ is the quantile forecast of the τ quantile. As is described in [116], the pinball loss function is a proper scoring rule closely related to the CRPS (see Section 6.1.1), but it is said to be easier to implement than CRPS. The quantile score is frequently used in practical applications and data science competitions, e.g., in [115]. The pinball loss function can be decomposed in a similar fashion to the CRPS [10] using a discretization of the forecasts. The decomposition is performed by introducing A equally populated sets I_a with $a \in 1,\dots,A$ that are created from the overall *sorted* set of all forecasts $\{\hat{y}_1^{(\tau)},\dots,\hat{y}_N^{(\tau)}\}$ of all N created quantile forecasts $\hat{y}_n^{(\tau)}$ with $n \in 1,\dots,N$ and $\hat{y}_n^{(\tau)} = \hat{P}_n^{-1}(\tau)$. The borders of $\{\hat{y}_1^{(\tau)},\dots,\hat{y}_N^{(\tau)}\}$ that divide the set to form each I_a therefore are the $\frac{1}{A}$ percentiles of the set. As denoted in [10], the composition is then given by

$$\begin{aligned}
\mathrm{QS}_\tau &= \sum_{n=1}^{N} \rho_\tau(\hat{P}_n^{-1}(\tau) - o_n), && (6.21)\\
&= \underbrace{\frac{1}{N}\sum_{a=1}^{A}\sum_{n\in I_a}\left[\rho_\tau(o_n - \bar{\hat{y}}^{(\tau,a)}) - \rho_\tau(o_n - \bar{o}^{(\tau,a)})\right]}_{\mathrm{REL}} && (6.22)\\
&\quad - \underbrace{\frac{1}{N}\sum_{a=1}^{A}\sum_{n\in I_a}\left[\rho_\tau(o_n - \bar{o}^{(\tau)}) - \rho_\tau(o_n - \bar{o}^{(\tau,a)})\right]}_{\mathrm{RES}} && (6.23)\\
&\quad + \underbrace{\frac{1}{N}\sum_{n\in I_a}\rho_\tau(o_n - \bar{o}^{(\tau,a)})}_{\mathrm{UNC}}, && (6.24)\\
& && (6.25)
\end{aligned}$$

where $\bar{o}^{(\tau)}$ is the *unconditional* τ quantile of all N observations, $\bar{o}^{(\tau,a)}$ is the *conditional* τ quantile of the observations in the set I_a, and $\bar{\hat{y}}^{(\tau,a)}$ is the discretized quantile forecast that is created with

$$\bar{\hat{y}}^{(\tau,a)} = \frac{1}{N_a}\sum_{n=1}^{N_a} y_n^{(\tau)} \tag{6.26}$$

from all $N_a = \frac{N}{A}$ forecasts that correspond to the observations in the set I_a.

The quantile score (QS) evaluates a single quantile τ, *not* an entire predictive distribution. In operational practice and for model selection, in many cases the overall form of the predictive distribution is of interest, not just the location of a single quantile (for example, an extreme quantile of a distribution). One could argue that a simple averaging of quantile scores over all L forecasted quantiles in the form

$$\overline{\mathrm{QS}} = \frac{1}{L}\sum_{l=1}^{L}\mathrm{QS}_{\tau_l} \tag{6.27}$$

enables the assessment of an entire predictive distribution. However, this clearly is not unproblematic, as we will show in the following section.

6.2 Experimental Evaluation

For the investigation of the score characteristics (the scores mentioned in Table 6.1) the data of the wind farms of the *EuropeWindFarm* data sets (publicly available at [76]) are used, which contain measurements and weather forecasts of 37 wind farms in Europe over a two-year time period. If not otherwise specified, the wind farm "wf3" of EuropeWindFarm is taken for the investigation of certain error effects. In the experiments, the similarity and discrimination abilities (of well performing probabilistic forecasting models to weaker performing forecasting models) of scoring rules are investigated in Section 6.2.1. In Section 6.2.2, an analysis of the quantile score decomposition is performed. Furthermore, influences of bias (Section 6.2.3), dispersion (Section 6.2.4), number of quantiles (Section 6.2.5) and the effects of parameter variation (Section 6.2.6) are investigated.

6.2.1 Score Discrimination Ability

This experiment aims at investigating the discrimination ability of the most popular scoring rules which are shown in Table 6.1. The idea is that scoring rules with high discrimination ability are able to show the differences of probabilistic forecasting models with different capability more clearly which in general is a desirable property.

In order to evaluate the similarities and discrimination ability of different scores, four probabilistic forecasting models were individually trained and evaluated for each of the 37 wind farms using a 4-fold cross-validation. The power forecasting models were selected with the goal of providing a diverse set of probabilistic forecasts with different qualities and characteristics. The most simple, but still reliable probabilistic model (Clim) is the sample climatology that forecasts the empirical marginal distribution of power measurements of the training data sets. This model does not take time-dependent weather information into account. The homoscedastic linear regression (LR) model creates the forecast based on a weighted combination of the weather parameters to predict the expectation value in the form of a normal distribution. The parameter estimation of LR is based on maximum likelihood which is optimal with respect to the ignorance score. As examples of more advanced models, the analog ensemble (AE) and quantile regression extreme learning machines (QR) consider nonlinearity and heteroscedacity. AE quantiles are estimated by the empirical quantiles of 40 training sample power measurement analogs where the corresponding normalized weather parameter vector has the lowest Euclidean distance to the currently queried sample. The number of 40 analogs was determined empirically by choosing the minimum CRPS of the validation set. Apart from this, no cost function is optimized for model training. The quantile regression (QR), which minimizes the quantile score per quantile, is built on top of nonlinear features which were derived with a combination of an extreme learning machine with 1000 randomly generated nonlinear features and a principle component analysis (PCA) where only 90 components are kept. The derived representation allows to model nonlinear dependencies by linearly weighting these features to predict single quantiles. The linear weights can then be estimated with conventional QR.

The results of the different power forecasting models evaluated individually with the scores given in Table 6.1 are shown in Fig. 6.2 (except the Hyvärinen score because it is incompatible with the pdf construction from quantiles). Each subplot shows an individual scoring rule for the four power forecasting models and is displayed such that a *lower* position within the plot indicates a *better* result. The box plots indicate the lowest and highest achieved score value (lower and upper whiskers) among all evaluated data sets, as well as the lower and upper

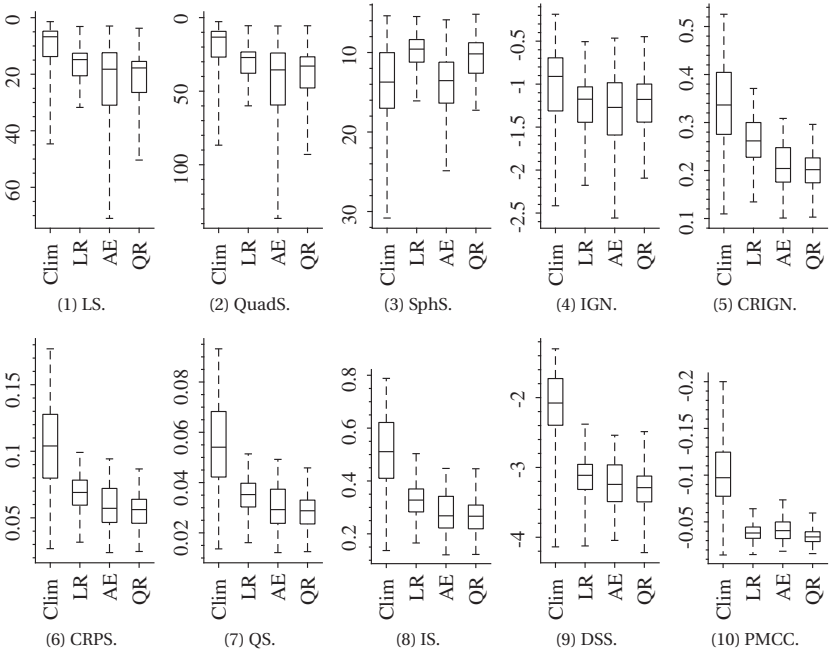


Figure 6.2: Analysis of scores given in Table 6.1 (except Hyvärinen Score) based on four probabilistic power forecasts: climatological forecast (Clim), homoscedastic linear regression (LR), analog ensemble (AE), and quantile regression (QR) of 37 wind farms. Due to the complexity of the models it can be assumed that Clim performs worst, LR yields better results, and AE and QR perform strongest. The score axis is chosen so that better forecasts are in a lower plot position. In accordance to each other, the simplest model (Clim) shows the worst value of the score median. This meets the expectation except for SphS which fails to show a similarly defined order of power forecasting models as given by the other scores. The models AE and QR can be expected to perform better than LR (which is linear and homoscedastic). CRIGN, CRPS, QS, IS, and DSS excel at the task of ordering the models by quality, while LS, QuadS, IGN, and PMCC deliver intermediate results. In this application with a stepwise constant pdf from quantiles, SphS is not very well applicable.

quartiles of the distribution of errors, and the median error. Without the use of the available input weather data, the Clim power forecast model is identified as the worst forecasting algorithm by all of the researched scores except for SphS. Most of the investigated scores also indicate a better performance of AE and QR in comparison to LR. This is not true for PMCC, which is non-proper as it prefers overly sharp models with a potentially low deterministic error of the forecast expectation over less sharp but reliable forecasts with similar deterministic quality. The quartiles and the lower quality boundaries are close to the expected order of the power forecasting model qualities for CRIGN, CRPS, QS, and IS. This indicates that these scores can discriminate the qualities of different models particularly well. DSS gives a less clear indication, but is nevertheless in a similar range. One aspect could be here, that DSS simplifies the distribution too much by making the assumption of an unskewed normally distributed forecast without kurtosis. Although LS is non-proper, the median matches the expected order, similar to QuadS and IGN. Nonetheless, the overall discriminative quality of this score is lower. In addition, LS will prefer overly sharp predicted densities over a reliable probability prediction.

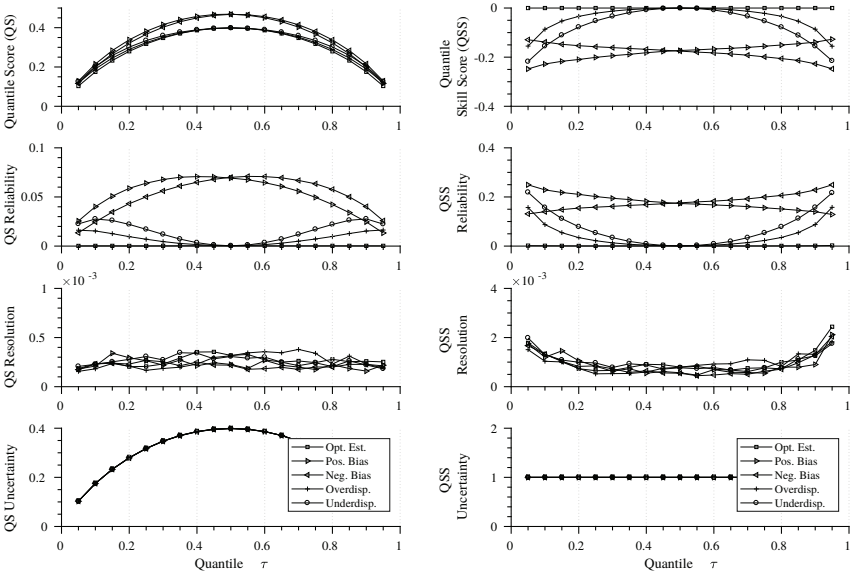


Figure 6.3: Quantile score (QS, left) and quantile skill score (QSS, right) decomposition on synthetic data, for details see Section 6.2.2. The quantile forecasts are modified to create systematic bias and dispersion errors. As can be seen from the lowermost left figure, the uncertainty component for QS depends on the evaluated quantile τ (unlike for the decomposition of other scoring rules as can easily be seen the terms of Eqs. 6.6 and 6.18 that describe the uncertainty). The QSS removes this systematic variation of the uncertainty component, which leads to a much clearer interpretation of the predictive distribution, as the error effects (of, e.g., the reliability component) are much more visible and better comparable for the different quantile levels.

6.2.2 Evaluation of Distributions Using the Quantile Score and Decomposition

Goal of this experiment is to show the properties of the quantile score and its decompositions when it is used for the evaluation of an entire predictive distribution instead of only a single quantile position (for which it was originally designed for). In a synthetic experiment, we analyze the behavior of the QS and its decomposition (proposed in [10]) on a set of 1000 samples drawn from a normal distribution $\mathcal{N}(x|\mu{=}0,\sigma{=}1)$. The quantile forecasts $\hat{y}^{(\tau)}$ are computed for the 19 quantiles in the set $\boldsymbol{\tau} = \{0.05, 0.1, \ldots, 0.95\}$ from the distribution directly, i.e., the quantile positions are optimal with respect to the random sample generating process. The quantile positions are then modified artificially to create a positive bias, a negative bias, and an underdispersive and overdispersive distribution (by addition or multiplication of a defined value to the quantile positions, respectively). Fig. 6.3 shows the results of the experiment.

The left-hand side of the figure shows the QS with its decomposition components. As can be seen from the graphs, the QS values, even for the optimally estimated variant, vary depending on the chosen quantile τ_l. According to expectation, for the optimal estimation (Opt. Est.) of the quantiles, the reliability (error) is 0 for each evaluated quantile. The resolution component is also close to 0 (as indicated in the figures, the scaling of the resolution component is 10^{-3}), as in the present experiment there are no explanatory variables (predictors), which means the optimal estimation can be seen as a climatological forecast (which has 0 resolution). The change in the overall QS value can therefore mostly be attributed to the uncertainty component (UNC), which also is the result of the decomposition. For other forms of decomposition (e.g., CRPS, CRIGN), UNC depends exclusively on the values of the tuple of observations $\mathbf{o} = (o_1, \ldots, o_N)$. For the QS, the interpretation of UNC is *different* as it also depends on the value of the quantile τ_l, as shown in the lowermost left figure, contributing in different quantities to the overall QS. QS values are therefore *not* comparable if they have been computed on different quantiles. Furthermore, the different error contributions are hardly visible in the summary QS in the topmost left graph in Fig. 6.3. The decomposition is therefore necessary for precisely determining types of errors.

The quantile skill score (QSS) with climatology forecast as baseline technique can be used to enable a better comparability. For each quantile τ_l, the QSS can be computed using

$$\mathrm{QSS} = \frac{\mathrm{RES} - \mathrm{REL}}{\mathrm{UNC}} \tag{6.28}$$

when assuming the climatology forecast as baseline as described, e.g., in [229]. This skill score computation is equivalent to the one proposed in Eq. 6.2 (and is possible for every decomposable skill score, for that matter). This yields an equal QSS for each of the quantile forecasts of "Opt. Est.", as visible in the graphs on the right-hand side of Fig. 6.3. The synthetically introduced errors of the experiment can also be observed more clearly, even when only examining the QSS summary score (top right). There, the bias effects are clearly visible (errors on all quantile levels, which can be attributed to reliability as shown in the second figure on the right side). Also dispersion errors are clearly visible, as the errors for the central quantiles are low, and increase towards the outer quantiles. However, both QS and QSS are not able to give an indication on the direction of the dispersion error. As can be seen, UNC now is normalized. Consequently, quantile skill score values for different quantile levels are better comparable.

Furthermore, the QS decomposition proposed in [10] contains a subsampling parameter

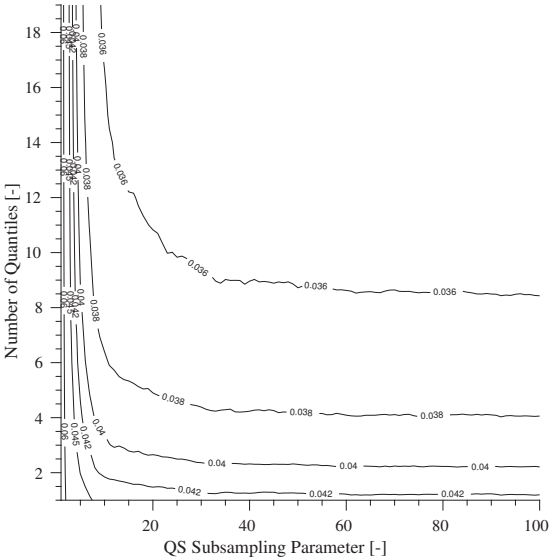


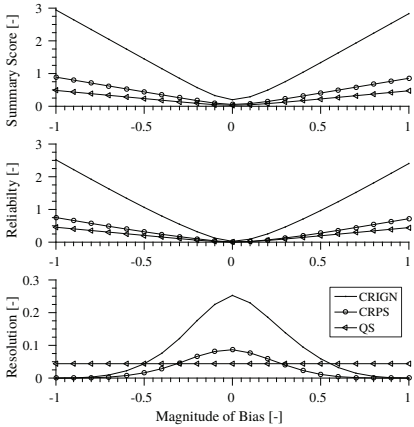
Figure 6.4: Development of the quantile score (QS) over the number of evaluated quantiles and the QS decomposition subsampling parameter (A in Eq. 6.21). As can be seen, the score differences converge for higher values of the subsampling parameter or number of quantiles (upper right corner).

(A in Eq. 6.21) for probability binning, which typically is an unwanted property of a scoring rule. The development of the QS value depending on the number of quantiles and the value of the subsampling parameter is shown in Fig. 6.4. As can be seen, when choosing very small values for the number of quantiles or the subsampling parameters, the QS does change significantly. However, the difference in the score converges for higher values of the two elements (upper right corner). For model comparison, the same number of evaluated quantiles and the same value for the QS subsampling parameter should still be chosen for achieving comparable results of the QS decomposition.

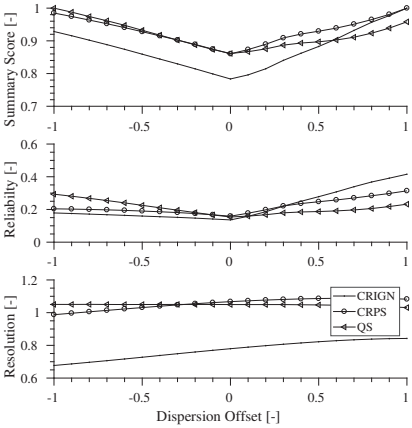
In summary, using the QSS, the error values for each quantile level are on the same magnitude and thus are comparable (unlike for the conventional QS). Thus, the QSS error is more suited for an overall assessment of the form of the probability distribution using Eq. 6.27 than the conventional QS, as discussed in [16] (without a dedicated discussion of the QS decomposition).

6.2.3 Influence of Bias on Decomposed Scores

One type of typical error appears with biased forecasts. The following example tries to examine some aspects of the score behavior under “laboratory conditions” when adding different bias values to the forecasting distribution of an actual wind farm. Negative forecasts or forecast values which exceed the nominal capacity are explicitly allowed and not “clipped” in order to provide a pure biased forecast with the same amount of information. As forecasting algorithm, an analog ensemble is used. The behavior of the three decomposable scoring rules CRIGN,



(1) Forecast system with artificially modified bias.



(2) Forecast system with artificially modified dispersion.

Figure 6.5: Experimental investigation of bias (Fig. 6.5.1) and dispersion effects (Fig. 6.5.2) on decomposed scoring rules. A real world wind farm data set is artificially modified to induce bias and dispersion errors. As can be seen from the figures, all investigated scores are able to correctly identify the unmodified forecasting system. However, the scores react with different sensitivities to certain error effects. Furthermore, differences in the resolution components are visible, giving insights on systematic differences in the informative value of the different forms of resolution, in particular with respect to the QS, which can be interpreted in the sense of a "theoretical potential" of the forecasting algorithm rather than an assessment of model spread. In Fig. 6.5.2, the error values are each normalized with respect to the highest summary score value for better visibility within the figure. The uncertainty components are not shown as they are either constant (CRPS, CRIGN) or in the shape of the lower left graph in Fig. 6.3 in case of the QS.

CRPS and QS regarding this artificial bias is displayed in Fig. 6.5.1. In this experiment, we investigate the scoring rules in their original form and not in any form that modifies their expressiveness (such as using skill scores).

The first row shows the behavior of the summary scores, which are not decomposed. The scores show similar behavior, since they all reach their minimum when there is no artificially added bias. With increasing absolute value of the bias, the quality of the forecast decreases which is visible by the higher values of the error scores. As can be seen from the figure, the scores differ in their absolute value. Where QS and CRPS are relatively similar, the CRIGN exceeds the CRPS value roughly by a factor of three. The CRPS may be seen as preferable in the sense of interpretability, since the magnitude of the bias can be observed almost one to one in the score result (for instance, a bias of +1 leads to a CRPS value of 1). The second row shows the reliability part of the examined scoring rules in the experiment. Since increasingly adding a bias will lead to more and more observations outside areas with high predicted probability, the reliability decreases. This behavior is similar to the behavior of the summary score. The most surprising result can be seen in the third row which shows the resolution part

of the scoring rules. On one hand, CRIGN and CRPS again show similar behavior, as the shape is similar, and the ratio between the scores is equal to the ratio in the summary score. On the other hand, the resolution part of the QS is *constant* and therefore apparently independent from the magnitude of the bias. In contrast to CRIGN and CRPS, the QS resolution is able to show the "theoretical capability" of a biased forecast rather than the resolution of the current state of predicted forecasts. The theoretical capability indeed does remains constant for a biased forecast. Using a simple calibration technique (i.e., a bias correction), the forecast can then easily be corrected to achieve reliability (and thus, a low summary score).

6.2.4 Influence of Dispersion on Decomposed Scores

The second basic type of systematic error is when the dispersion (or spread) of a forecast is not estimated correctly. Goal of this experiment is to show how errors regarding the spread influence the values of scoring rules. In order to show the effects on the scores, a second experiment of real wind farm data manipulated artificially is conducted using a quantile regression algorithm. To create a pure dispersion error, the actually trained forecasted distribution is manipulated to systematically create forecasts that are over- or underconfident while keeping the median forecast constant. The result of the experiment is shown in Fig. 6.5.2. In the figure, the dispersion is modified with the function 2^x, so that a dispersion value of -1 represents an overconfident forecast (50 % width of true distribution), a value of 0 represents the original distribution, and a value of 1 represents 200 % width, which means the forecast is underconfident. The score values are normalized with the maximum value of each summary score for better visualization within the same figure.

As can be seen from the summary scores, all scores have their minimum at the point of correct estimation of the distribution (which confirms the propriety assumption of the scores). The CRPS penalizes confidence errors roughly equally in both directions, while CRIGN penalizes underconfidence more. QS, on the other hand, penalizes overconfidence more drastically. In accordance with the expectation, this behavior is reflected in the reliability component of each score, as the induced errors can be attributed to reliability errors. However, for CRIGN and CRPS the resolution component does steadily increase with wider forecasted distributions. The QS, on the other hand, remains constant for all evaluated situations. This confirms the assumption that the actual interpretation of the resolution component of QS is quite different to the interpretation of CRPS and CRIGN. While resolution for CRPS and CRIGN can be interpreted in the sense of average spread of the forecasting distribution, resolution of the QS again describes the theoretical capability of a forecasting system to yield accurate forecasts after calibration (e.g., by adjusting the spread of a parametric forecasting model).

6.2.5 Influence of Number of Quantiles

In this experiment, the behavior of scoring rules is analyzed when increasing the precision of the predictive distribution on a real-world data set (wf3). Goal of this experiment is to examine how different scoring rules behave when varying the amount of detail of a predictive distribution. In order to modify the precision of the distribution, the overall predictive distribution is constructed using a different number of quantiles. The quantiles $\tau_1, \ldots, \tau_L$ are defined to be equidistant in the range $[0, 1]$. As probabilistic forecasting algorithm, an analog ensemble with 40 neighbors is utilized. The results are shown in Fig. 6.6. The QS for the entire distribution is computed with Eq. 6.27. Each score is normalized with respect to its highest value that it achieved in the data set for better visibility within the same figure.

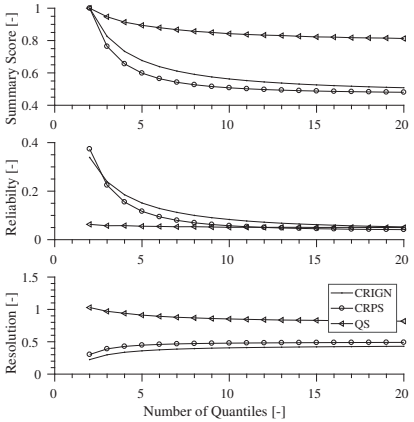


Figure 6.6: Analysis of the effects of an increasingly detailed form of the predictive distribution by using a higher number of equidistant forecasted quantiles on the scoring rules scores. As can be seen, a more detailed form of the predictive distribution yields better score values regarding the summary score. The independence of the QS from the number of evaluated quantiles regarding reliability can be observed.

As can be seen from the figure, all scores yield better results with increasing precision of the predictive distribution. Therein, CRPS and CRIGN benefit more clearly from an increase of the precision of the predictive distribution. This decrease of the error is attributed to both an increase in resolution and a lower reliability error. All scores converge to a lower bound of the score value when increasing the number of evaluated quantiles.

The QS barely benefits from more evaluated quantiles, which is completely in line with the expectation, as it evaluates the quantile positions only, and not the entire predictive distribution. A major difference in the characteristics of the QS is visible in the reliability part of the QS. Apart from minor sampling effects, the reliability is independent from the number of evaluated quantiles, as each individual quantile from which the overall QS is computed using Eq. 6.27 is reliable. So, when summing up a set of quantiles which can be assumed to have approximately equal reliability, the overall mean reliability remains the same independently from the actual number of quantiles that have been utilized for the computation. The decrease of the score value, however, has to be attributed to the change in uncertainty when summarizing the QS on multiple quantiles, as has been shown previously in Fig. 6.3. This experiment therefore again shows the difficulties when using the QS (unlike the QSS) to evaluate the overall predictive distribution, as described in more detail in Section 6.2.2. The higher score of the CRPS when having only few quantiles can be expected as laid out in [244]. The authors furthermore propose a modified version of the CRPS to account for distributions built from EPS forecasts (using Eq. 5.34) with small ensemble size. EPS forecasts, however, are not used in this experiment. The proposed modified CRPS targets a practical problem in forecasting (the assessment of forecasts built upon small EPS ensembles) rather than improving upon a systematic weakness of the CRPS.

6.2.6 Varying Parameter Characteristics of Probabilistic Forecasting Techniques

In this experiment, the behavior of the scoring rules regarding a practical aspect in forecasting is examined. Some forecasting algorithms (e.g., analog ensembles, kernel density methods) have the property of being able to gradually perform forecasts from high-resolution, but likely unreliable, forecasts, to very smooth and reliable, but broad and unspecific forecasts. A forecasting model can be optimized more towards one of the above goals. In this experiment, we therefore evaluate the scoring rules given an analog ensemble algorithm while performing a "sweep" from a small number of analog situations up to all available analogs, which corresponds to sample climatology. Therein, we aim to analyze the summary score, reliability, and resolution aspects of the scoring rules.

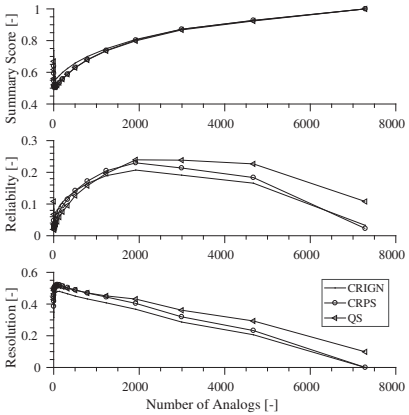


Figure 6.7: The experiment investigates the behavior of scoring rules for forecasting algorithms with sweep characteristics from point forecasts to climatological long-term average, in the evaluated case using an analog ensemble. As can be seen from the figure, the optimum number of analogs is clearly visible. Somewhat unexpected, the reliability error does increase after this optimum before again approaching the low reliability error of the climatological forecast.

The results of the experiment are shown in Fig. 6.7. The evaluated scoring rules are normalized by the maximum value of each summary score for better visibility in a single figure. As can be seen from the figure, the examined scores agree upon the optimal number of analogs. The scores values remain low in the area between 20 and 150 analogs. This behavior is in line with the expectation laid out above. The resolution component is on a high level for a small number of analogs and steadily decreases with an increasing number of analogs, which also matches the expectation. The QS, however, shows the unexpected behavior of having nonzero resolution at the maximum number of analogs, though the predictive distribution remains constant over time for this type of forecast. It could be expected that the reliability error of the analog ensemble decreases quickly up to a certain point, and then only improves slowly towards the climatological forecast with optimum reliability. While a rapid decrease of the error for low numbers of analogs can be observed, the reliability error *increases* after this optimum. In contrast to expectation, more analogs therefore do not necessarily lead to a more

reliable forecast. An explanation is the typically occurring dominance of low power situations in a data set that exhibits this problem (as, even when the current NWP is high, a probability for a low power generation is forecasted because all nearest neighbors in the higher power generation scenarios have already been utilized for the analog ensemble creation, thus for the remaining analogs, insensible situations are chosen). The predictive distribution therefore is dominated by observations that are not similar, which hinders the ability of the technique to accurately predict high power generation and other rare events.

6.3 Discussion and Conclusion of Probabilistic Error Scores

This section discusses our key findings which are supported by the experiments regarding the probabilistic error scores in Section 6.3.1 and the decomposition characteristics in Section 6.3.2.

6.3.1 Score Applicability

In the experimental evaluation we have investigated the behavior of a number of commonly used scoring rules. One of the major results is that the quadratic score and the spherical score are not well suited for the evaluation of quantile based pdf representations. However, it remains to be seen whether these scoring rules are more capable on continuously differentiable density functions. The Hyvärinen score is a local score, however, it is incompatible with the pdf construction from quantiles as the derivations of step-pdfs reduce to δ (Dirac) functions ($\hat{p}'$), or 0 ($\hat{p}''$), respectively. As a universal measure for all possible forms of distributions it is therefore not well suited. The centered IS is just a scaled version of the QS for both interval borders. Therefore, for IS the same properties apply as for the QS. Considering the divide and conquer principle, the QS remains the more flexible, easier to optimize scoring rule that does not bear the risk of uncentered prediction intervals. Therefore, we argue that QS is preferable to IS. The QS is the only scoring rule investigated in this context that is able to evaluate defined probability levels within the predictive distribution. Therefore, it is well suited for the investigation of extreme quantiles. In particular in the context of power forecasting, this may be useful for the planning of balancing power in the power grid.

Scores such as DSS and PMCC create moment-based statistics of the predictive distribution. While they are suited to assess the most important characteristics of a predictive distribution (the model spread and bias effects), they may turn out problematic when resolving higher-moment properties, such as with heavily skewed distributions which are frequent in power forecasting. However, they are easily understandable, which may be an advantage. When using density forecasts from quantiles, local scores may struggle with their evaluation, e.g., if adjacent quantile forecasts collapse in the same power value (leading to a density value of infinity at the particular point of the quantile forecast). This may be overcome using regularization techniques. While the SphS and QuadS are nonlocal scores, they nevertheless exhibit properties of local scores (as they both evaluate the density at the location of the observation $\hat{p}(o)$), and thus, suffer from the same phenomenon.

Propriety is argued to be a necessary property for scoring rules (e.g., to avoid hedging). When using scoring rules during model training for optimization, using a proper scoring rule is indispensable. We have found the use of non-proper scoring rules to not necessarily be a problem when using the scoring rule for evaluation only. However, their use still should be discouraged due to the availability of similarly simple, but proper scoring rules. When performing gradient-based optimization of model parameters, IGN is a very natural choice

since it is related to maximum likelihood estimation, while QS is particularly well suited for quantile regression parameter training. One score not evaluated in this context is the energy score, which is the multivariate generalization of the CRPS. It is tailored to the evaluation of scenario forecasts. As result of the observations of the experiments, we can in summary give a recommendation for the scores CRPS, CRIGN, and QS as starting point for the evaluation of probabilistic forecasts when not aspiring the investigating of very specific characteristics of the predictive distribution.

6.3.2 Decomposition Characteristics

The decomposition of scoring rules can make sense for the determination of the sources of errors of forecasting algorithms. Computing decomposed scores, however, in many cases is a lengthy and not intuitive process with non-compact representation as result (three components). The use of alternative forms of decomposition may also be advantageous for certain task. For instance, the form of decomposition of [108] which uses the decomposition into a "potential" component rather than uncertainty and resolution is easier explicable, as the error is equal to the one of the reliability component in the sense that it is defined in $[0, +\infty]$, where lower is better.

The QS does have some attractive decomposition properties which are advantageous to the form of decomposition of CRIGN and CRPS in some situations. For instance, the resolution component expresses more a "theoretical potential" in the sense of correlation rather than a pure description of the model spread. This can, for example, be utilized for model selection: Having computed a number of models with corresponding decomposed scores, the model with the highest resolution can be chosen as best model rather than the model with minimum summary score. Thereby, the model with the highest resolution may yield an even lower summary score *after* calibration. To the best of our knowledge, this systematic difference in the representation of the resolution component has not yet been pointed out in the literature.

The computation of the decomposition typically is performed in the general form suitable for CRPS and CRIGN alike which is laid out, e.g., in [229]. This form of decomposition breaks down the continuous decomposition into a series of binary problems for every threshold. While this approach is appealing in theory, the computation is cumbersome in practice, as its complexity is $O(N^2)$ with N being the number of evaluated samples. As simplification, a binning of the quantile values into a number of defined discretized forecast values of the quantile forecasts can be achieved without sacrificing accuracy in a noteworthy way. The computational complexity is then reduced to $O(N)$. However, this binning of power values again introduces a hyperparameter, which typically is an unwanted property for a scoring rule. The rather "visual" form of decomposition in [108] does not have these disadvantages, however, has yet only been proposed for the CRPS. Likewise, the QS decomposition does have a binning parameter which is suboptimal. However, in the experiments the differences have been shown to converge for higher values of the parameter. Nevertheless, the value of the parameter should always be mentioned when using the QS decomposition.

Chapter 7

Probabilistic Coopetitive Soft Gating Ensemble: A Novel Combination Approach for Distribution Forecasts

This section extends the methodology of the coopetitive soft gating ensemble technique of Section 4 to probabilistic distribution forecasts. The idea is to use a scheme for the combination of probabilistic forecasts and modify the CSGE technique by optimizing the ensemble with respect to probabilistic scoring rules. The idea is to retain the main innovations of the CSGE technique, namely the hierarchical ensemble structure, the innovative coopetitive soft gating weighting function, and the multi-scheme weighting, all of which are described in more detail in Section 4.

An overview of the combination of probabilistic forecasts is given in Section 7.1. The novel technique that is called *probabilistic coopetitive soft gating ensemble* (PCSGE) is described in Section 7.2. In Section 7.3 the performance of the PCSGE technique is analyzed in comparison to other state of the art probabilistic forecasting techniques and probabilistic ensembles for both, a day-ahead and an intraday forecast. The investigation also includes an analysis of the reliability and sharpness properties. The findings are summarized in Section 7.4.

7.1 Combination of Probabilistic Forecasts

As described in Section 2.4.2, a distribution forecast that uses a probability density function (pdf) can be denoted with

$$\hat{p}_{t+k|t}(y) = f(\mathbf{x}_{t+k|t}|\boldsymbol{\theta}), \tag{7.1}$$

where $\hat{p}_{t+k|t}(y)$ is the pdf of a target quantity y (i.e., the power generation y for power forecasting applications) created for the point in time $t+k$ with lead time k from the forecasting origin t. For many relevant lead times, the pdf is mainly based on an NWP forecast $\mathbf{x}_{t+k|t}$ and a function f with governing function parameters $\boldsymbol{\theta}$ that performs the conversion of the NWP forecast to a power forecast.

In the case of a weighted aggregation of probabilistic forecasts in an ensemble, the probabilistic ensemble distribution forecast $\tilde{\hat{p}}_{t+k|t}(y)$ can be written as

$$\tilde{\hat{p}}_{t+k|t}(y) = \sum_{j=1}^{J} w^{(j)} \cdot \hat{p}^{(j)}_{t+k|t}(y), \tag{7.2}$$

where $j = 1,\dots,J$ are the ensemble members and each $\hat{p}^{(j)}_{t+k|t}(y)$ is a single probabilistic forecast of forecasting model j that participates in the ensemble. The single weights have to fulfill $\sum_{j=1}^{J} w^{(j)} = 1$ and $w^{(j)} \leq 0$ in order to comply to

$$\int_{-\infty}^{+\infty} \tilde{\hat{p}}_{t+k|t}(y) = 1. \tag{7.3}$$

The main innovation again is the way the single ensemble members are weighted using the weighting factors $w^{(j)}$ depending on the current weather situation and lead time.

7.2 The Probabilistic Coopetitive Soft Gating Ensemble Technique (PCSGE)

This section describes the proposed probabilistic coopetitive soft gating ensemble (PCSGE). An overview of the technique is given in Fig. 7.1. The weighting of the ensemble remains the same as the general ensemble aggregation formula of Eq. 7.2. However, we have a hierarchical ensemble structure: For each weather forecasting model $\psi = 1,\dots,\Psi$ (which can be an arbitrary NWP of an EPS, MME, or TLE, e.g., of an intraday or day-ahead model, for a particular time step to be forecasted), a number of power forecasting models $\varphi = 1,\dots,\Phi$ are used to forecast the target predictand for each weather forecasting model. The power forecasting models do not necessarily have to be the same for each weather forecasting model, but, for the sake of easier understanding, we will use the same type and number of power forecasting models for each weather forecasting model here. The overall number of ensemble members J consequently is $J = \Psi \cdot \Phi$. The individual predictions of each power forecasting model are then aggregated and fused to an overall forecast in a post-processing step according to Eq. 7.2. The main innovation here is the way the single weights $w^{(j)}$ are constructed.

The methodology of the (deterministic) CSGE algorithm of Section 4.4 can be used straight-forward using the three weighting aspects of global weighting (Section 4.4.1), weather situation-dependent weighting (Section 4.4.2), and lead time-dependent soft gating (Section 4.4.3). The overall PCSGE forecast is created using Eq. 7.2. The single weights $w^{(j)}$ in Eq. 7.2 are computed using Eq. 4.8.

In the deterministic case, the computation of the weights is based on deterministic error scores of Eq. 4.11. For probabilistic forecasts, the performance can be assessed using a *scoring rule S* that replaces the deterministic error score of Eq. 4.11 for the PCSGE technique. Similar to deterministic error scores, all scoring rules evaluate the performance using a predictive distribution $\hat{p}_{t+k|t}(y)$ and a corresponding observation o_{t+k} in the form

$$S(\hat{p}_{t+k|t}(y), o_{t+k}). \tag{7.4}$$

For the coopetitive soft gating formula, any non-negative scoring rule can be used, e.g., the CRPS [108]. The overall performance assessment formula of Eq. 4.12 consequently is also slightly altered, leading to the evaluation of a data set with N forecast-observation pairs and a probabilistic forecast $\hat{p}^{(\varphi|\psi)}_n(y)$ computed on weather forecasting model ψ and power forecasting model φ with

$$\bar{S}^{(\varphi|\psi)} = \frac{1}{N} \sum_{n=1}^{N} S(\hat{p}^{(\varphi|\psi)}_n(y), o_n). \tag{7.5}$$

Apart from the two modifications of Eq. 7.4 and Eq. 7.5 laid out above, the methodology of the

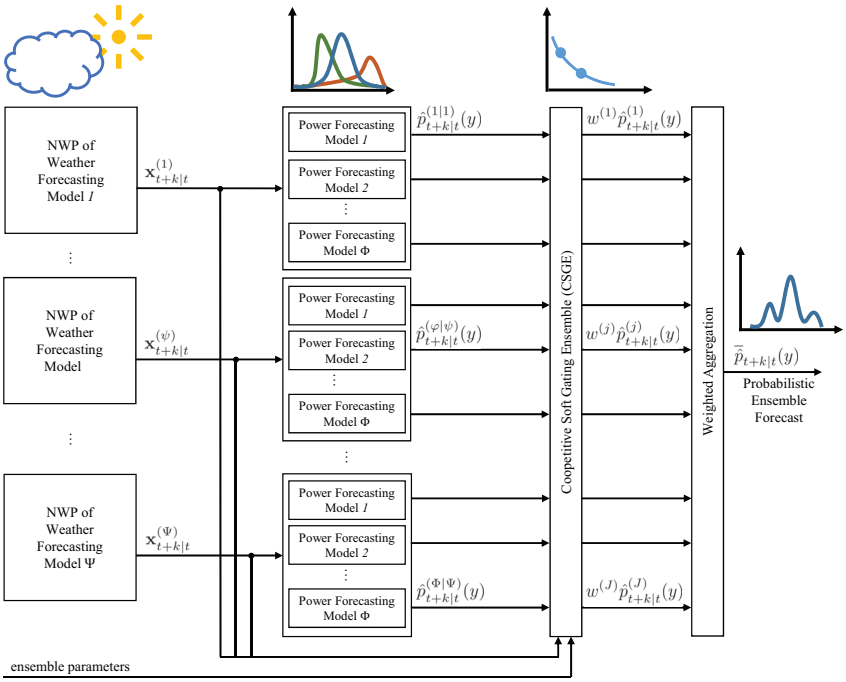


Figure 7.1: Overview of the probabilistic coopetitive soft gating ensemble (PCSGE) model structure. The technique is constructed as a hierarchical ensemble. A number of weather forecasting models (WM), which can be weather ensembles such as EPS, MME, or TLE, are each used for the creation of forecasts with a number of probabilistic power forecasting models (PM). In a post-processing step, each WM-PM combination is assigned a weight using Eq. 4.8. In the final step, a weighted aggregation is performed using Eq. 7.2.

CSGE algorithm laid out in Section 4.4 can be adopted.

Exemplary Forecast

When using probabilistic forecasts, a probability distribution is constructed for each point in time. An example of forecasts created by a probabilistic forecasting technique is shown in Fig. 7.2. The figure shows a day-ahead forecast that is predicted using the proposed ensemble technique. The time axis indicates 100 consecutive hours in which a rolling forecast is created. For the distribution forecast, the grey shaded areas indicate the 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.99 quantiles of the predictive distribution, where the solid line with square markers represents the median forecast. As can be seen from the figure, the probabilistic forecast estimates the amount of expected uncertainty of a forecast for each point in time. As can further be seen, the extreme quantiles are estimated very conservatively in this case, as the forecast of the 0.01 quantile is close to zero.

An exemplary forecast for a *particular* point in time of the trained ensemble technique

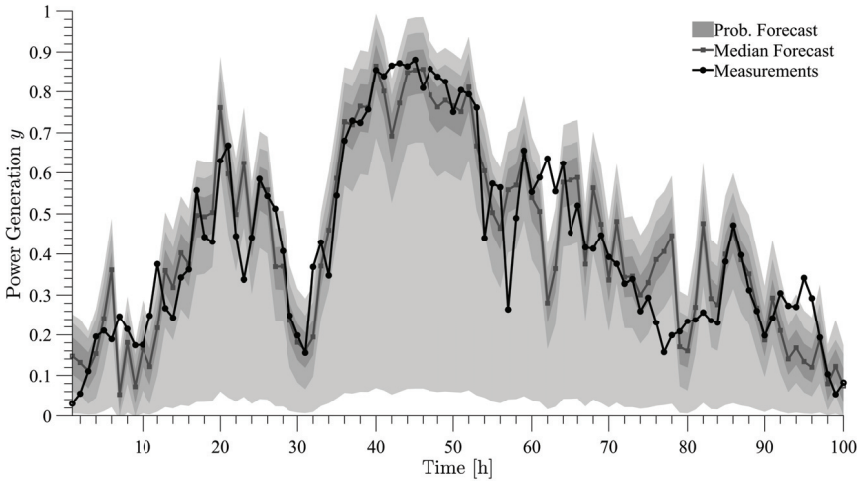


Figure 7.2: Example of a probabilistic distribution forecast. In contrast to conventional point forecasting, a predictive distribution is created for each point in time. The grey colored bars indicate the 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.99 quantiles of the predictive distribution. The square line denotes the median forecast. The actual power measurement are denoted with the circle line. As can be seen from the figure, the 0.01 quantile estimation is very conservative.

is visualized in Fig. 7.3. The figure shows three weather models (WM), $\Psi = 3$, each of which is predicted using three parametric probabilistic forecasting models (PM), i.e., $\Phi = 3$. Each PM-WM combination is denoted in the figure with $\hat{p}^{(\varphi|\psi)}(y)$. In the example, PM 1 is a homoscedastic (constant variance) linear regression model, PM 2 is a homoscedastic polynomial regression model, and PM 3 is a heteroscedastic (varying variance) support vector regression (SVR) model. For the sake of simplicity, all forecast distributions of the base predictors are assumed to be in the form of a normal distribution. Details on the form of probability representation of these models can be found in [18, 228] and in Section 5.6.1. A refined prediction that forms the ensemble prediction $\tilde{p}(y)$ is created from Eq. 7.2 as indicated by the black thick line that shows the ensemble forecast over the random variable y which represents the generated power. The weights are determined according to Eq. 4.8.

7.3 Experimental Results

This section investigates the performance of the proposed PCSGE technique. The experimental setup and the comparison techniques are described in Section 7.3.1. The performance of the investigated techniques for day-ahead forecasting is shown in Section 7.3.2, while the performance for intraday forecasting is examined in Section 7.3.3.

7.3.1 Experimental Setup

For the evaluation, each data set from the data set collection described in Section 4.5.1 is split into a training and a test subset in a 5-fold cross-validation with a training data set that

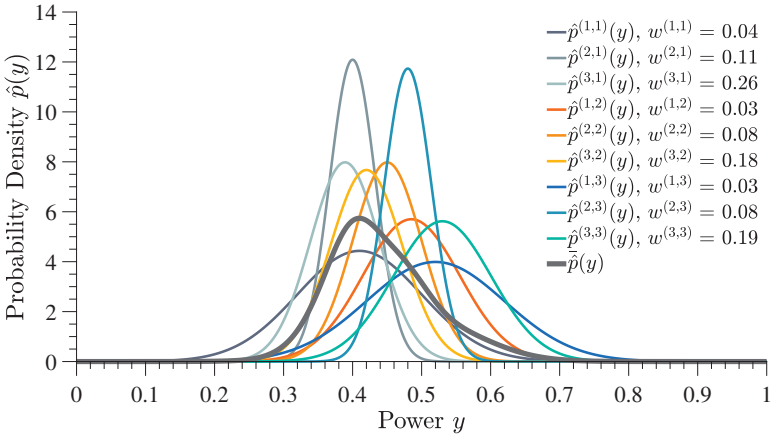


Figure 7.3: Example of the ensemble combination using Eq. 7.2 of a combination of three parametric forecasting models and three weather models. For the sake of simplicity, parametric Gaussian functions are chosen in the visualized case. The colored lines show the single base predictors, the black thick line denotes the refined ensemble forecast which is constructed from a weighted aggregation of the base predictors. The respective ensemble member weights are computed using Eq. 4.8.

includes (4/5) of the data and a test data set (1/5). The results presented in the case studies below thus are the results regarding five repetitions with each data set. When using the PCSGE technique, the training data set is further split into three sets of equal size which are called *parameter set* (1/3), *optimization set* (1/3) and *validation set* (1/3) for the sake of clarity. The single power forecasting models for each weather forecasting model are trained using the parameter set. The parameter optimization of the PCSGE technique is then performed on the optimization set (that serves, e.g., as historic data set for the local soft gating) and is finally optimized regarding $\boldsymbol{\eta}$ with the validation set. The parameter combination which performed best on the validation set is chosen as final model parameterization which is used to compute the final model quality on the test set. This process is conducted for each of the five folds.

The following probabilistic forecasting models (non-ensembles) are utilized for comparison in the experiments:

1. HoLR: This model is a simple parametric homoscedastic (constant variance) forecasting model that uses a linear regression model for the estimation of the expectation value such as described in Section 5.6.1. As probability distribution, a normal distribution (that is truncated to the interval $[0, o_{\text{inst}}]$) is chosen. The variance parameter σ is optimized to yield the minimum CRPS value on the training data set. This technique is, e.g., utilized in [18]. This model is also used as the baseline technique for the computation of the skill score.

2. HeSVR: This technique is a heteroscedastic (varying spread) parametric forecasting model which is also explained in Section 5.6.1. A support vector regression (SVR) model is chosen for the estimation of the expectation value. The uncertainty is assumed to have the shape of a beta distribution to better model the power interval of $[0, o_{\text{inst}}]$,

consequently there is no need for truncation of the probability distribution. The parameters a, b of the beta distribution are each estimated as functions of the expectation value (in the form $a = g(\mu)$ and $b = h(\mu)$ with $a, b \in \mathbb{R}^+$) that is created from the SVR technique (where $\mu = f(\mathbf{x}_{t+k|t})$). The parameters of g and h are optimized to minimize the CRPS. Heteroscedastic techniques have, e.g., been used in [228].

3. ELMQR: As sophisticated non-parametric probabilistic forecasting technique, an extreme learning machine is used to forecast the quantiles $0.1, 0.2, \ldots, 0.9$ of the predictive distribution. The technique therefore is based on quantile regression described in Section 5.6.4. In principle, each forecasted quantile can be predicted individually, however, for the evaluation, an entire predictive distribution is constructed using the scheme laid out in Section 5.2. For the ELM 300 neurons and a rectified linear unit (ReLU) activation function are used. ELMQR has, for instance, been proposed in [239] for power forecasting.

Furthermore, the following probabilistic ensemble techniques are included:

4. Ens. HoLR: This technique uses the HoLR technique in an ensemble with multiple weather models. The forecast of each weather model is averaged (using Eq. 7.2 and $w^{(j)} = \frac{1}{J}$) to create the ensemble forecast such as performed in [104].

5. Ens. HeSVR: This technique uses the HeSVR technique in an ensemble with multiple weather models. The weighting of the ensemble is determined using the coopetitive soft gating function using a parameter of $\eta = 2$ (such as performed in *Ens We.* of Section 4.5.3). The final aggregation is performed using Eq. 7.2.

6. BMA SVR: Bayesian model averaging (BMA) is a state of the art aggregation technique for multi-model ensembles. BMA is a realization of an ensemble dressing model and is described in Section 5.7.3. The BMA SVR technique uses an SVR forecast for each weather model. BMA utilizes a linear correction term on the forecasts of the expectation values to eliminate systematic errors in the forecast. Furthermore, the spread of each forecast is trained directly using a normal distribution and the iterative expectation maximization (EM) algorithm [167]. BMA has, e.g., been utilized in [199, 219] for power forecasting.

Finally, the following variants of the proposed PCSGE technique are utilized:

7. PCSGE (V1): The PCSGE technique is applied using a single weather model (WM) and techniques 1.-3. of the above probabilistic power forecasting models (PM). As there is only one weather forecasting model, the number of weighting dimensions is reduced to the three power forecasting model based weighting factors $w^{(\varphi|1)} = w_g^{(\varphi|1)} \cdot w_l^{(\varphi|1)} \cdot w_k^{(\varphi|1)}$ for each power forecasting model. The local weight determination of PCSGE is performed using a range search as laid out below.

8. PCSGE (V2): The PCSGE technique is used with multiple weather models (WM) and a single (best) power forecasting model, the HeSVR technique (PM). As there is only one power forecasting model, the number of weighting dimensions is reduced to the three weather forecasting model based weighting factors $w^{(\psi)} = w_g^{(\psi)} \cdot w_l^{(\psi)} \cdot w_k^{(\psi)}$ for each weather forecasting model. The local weight determination of PCSGE is performed using a range search as laid out below.

9. PCSGE (V3): The PCSGE technique is used with multiple weather models (WM) and power forecasting models 1.-3. (PM). The PCSGE technique consequently uses all six weighting factors. The local weight determination of PCSGE is performed using a range search as laid out below.

For the PCSGE techniques, the regularization parameter ζ (described in Section 4.4.4) is set empirically in a way which avoids overfitting of the model. A nearest neighbor range search is used for the local weight determination with NWP features being weighted according to a feature importance weighting prior to the nearest neighbor search. This process is conducted in the following way:

- The features of each weather model contain seven features which have been selected by domain experts for wind power forecasting. All present features can thus be assumed to be relevant, consequently no feature *selection* has to be performed. However, for the nearest neighbor search, a feature *weighting* may improve the quality of the local weighting.
- A computationally inexpensive filter approach for filter weighting is chosen. As has been pointed out in [141, 220], the Relieff algorithm [140] can be used for creating feature weights for a nearest neighbor search. After computing Relieff, the resulting features weights created by Relieff (which in other applications can be used for feature ranking) are used in this case for feature scaling to achieve the effect of a Mahalanobis distance computation (assuming no covariance).
- A simple nearest neighbor range search (with Euclidean distance as distance metric performed on the scaled features) is used to find the nearest neighbors. For determining the range parameter of the range search, prior to the ensemble computation, the training data set is split into a training and test data set and the range parameter is chosen in a way that minimizes the CRPS score. This is performed by using the identified neighbors as basis for the computation of an analog ensemble (see Section 5.6.3).
- The weather situations identified as neighbors for a certain query NWP are then used for the computation of the local weighting based on Eq. 4.18.

For all models, we use the CRPS [161] as scoring rule S for model training (training of individual models and the ensemble) and evaluation which is defined by

$$S_{\mathrm{CRPS}}(\hat{P}(y), o) = \int_{-\infty}^{+\infty} \left(\hat{P}(y) - H(y-o)\right)^2 \mathrm{d}y \tag{7.6}$$

for a single forecast-observation pair with H being a Heaviside step function at the location of the observation o. As described above, this scoring rule is utilized as score within the ensemble (i.e., utilized in Eq. 4.11). More details on the CRPS are also given in Section 6.1.1.

7.3.2 Case Study: Day-Ahead Performance on Single and Multiple Weather Forecasting Models

This case study aims at analyzing the forecasting performance for a day-ahead forecast (k_{min} = 25 h, k_{max} = 48 h, Δ = 1 h). The algorithms thus use (up to) the three day-ahead weather models that are available in the data set collection described in Section 4.5.1. This section examines the forecasting performance of the PCSGE technique in comparison to other

state of the art forecasting models using the methodology laid out in Section 7.3.1. The results of the experiments are shown in Table 7.1. The table denotes the results of 185 experimental runs for each forecasting model (5-fold cross-validation for each of the 37 wind farms). In the table, the minimum and maximum values of the CRPS, as well as the 10% and 90% quantile, and the median error values are given. Furthermore, the mean error, the standard deviation of the errors, and the skill score are denoted.

Table 7.1: Performance comparison regarding the distribution of CRPS scores, the mean error, the standard deviation of errors, and the skill score of the experiments for day-ahead forecasting. The color coding indicates the quality of each power forecasting model from high quality (green) to low quality (red). Bold text highlights the best achieved score for single and multiple weather models (WM) individually.

	Single WM				**Multiple WM**				
	HoLR	HeSVR	ELMQR	PCSGE (V1)	Ens. HoLR	Ens. HeSVR	BMA SVR	PCSGE (V2)	PCSGE (V3)
Max	0.106	**0.095**	0.098	**0.095**	0.110	0.098	0.096	**0.090**	0.091
90% Quant.	0.097	0.082	0.085	**0.079**	0.101	0.084	0.081	0.078	**0.077**
Median	0.076	0.060	0.060	**0.058**	0.078	0.060	0.057	0.058	**0.056**
10% Quant.	0.052	0.038	0.038	**0.037**	0.052	0.038	0.037	0.037	**0.036**
Min	0.038	0.026	**0.019**	0.022	0.038	0.028	**0.021**	0.023	**0.021**
Mean	0.075	0.060	0.061	**0.058**	0.076	0.060	0.058	0.057	**0.056**
Std. Dev.	0.016	0.016	0.016	**0.015**	0.017	0.016	0.016	**0.014**	0.015
Skill	0.0%	19.9%	18.8%	**22.1%**	-2.3%	19.3%	22.3%	23.4%	**24.9%**

As can be seen from the table, when considering models that are based on the single best weather model, the HeSVR and the ELMQR perform about equally well, where the HeSVR technique yields better results regarding the maximum error and the 90% quantile, while the ELMQR technique is able to yield a better minimum error. Both are able to outperform the HoLR technique by 19.9 % and 18.8 % (computed using the Skill, see Eq. 4.33), respectively. The best technique based on a single WM, however, is the PCSGE (V1) technique, that yields superior results in every metric but the minimum error. It performs 22.1% better in comparison to the baseline technique (skill). The standard deviation of errors is very similar among the techniques, the PCSGE (V1) technique however still yields the best result.

When using multiple WM, somewhat surprisingly, the performance of the HoLR technique *decreases* in comparison to the HoLR technique that is solely based on the single best weather model. The same is true for the Ens. HeSVR technique, that, while it performs stronger than HoLR, is not able to exceed the performance of the HeSVR technique that is based on only a single weather model. This is not in line with the behavior of the multi-model combination of deterministic forecasts, where even simple techniques, such as the averaging of forecasts from different WM, yield better results in comparison to techniques that only use a single WM on average (see Tables 4.7 and 4.8 for details). This may be due to the fact that the probabilistic forecast is already specified in the most precise way by the single best weather model. The static inclusion of additional weaker weather models may only hinder the ability of the ensemble model to create sharp forecasts which in turn leads to worse CRPS values.

However, when using more sophisticated weighting techniques, such as performed in the BMA SVR technique, the model can nevertheless yield superior results that in the case of BMA SVR is 22.3 % better than the baseline technique. The best results are achieved by the PCSGE (V2) and PCSGE (V3) techniques, that are able to improve upon the baseline technique by 23.4 % and 24.9 %, respectively. In particular the PCSGE (V3) technique is able to yield the

overall best results regarding all metrics but the maximum error and the standard deviation of the CRPS values.

The ranked performance of the probabilistic forecasting techniques of the 185 experiments is evaluated in Fig. 7.4. The ranking is performed individually on the CRPS scores of each of the 185 experiments using the methodology laid out in Section 2.8.3.

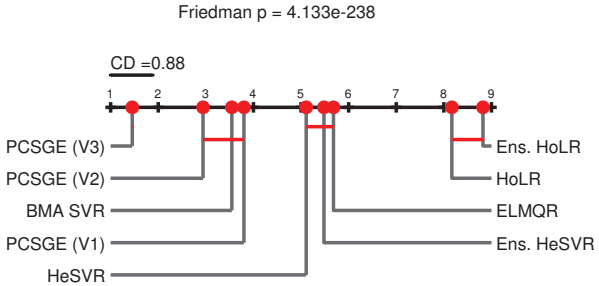


Figure 7.4: Evaluation of ranked performance of the day-ahead forecast using the Nemenyi post-hoc test. The algorithms have to exceed the value of the critical distance (CD) in order to be assumed to be statistically different. As can be seen from the figure, the PCSGE (V3) technique has the best ranked performance and is significantly better than the other comparison techniques on $\alpha = 0.05$.

As can be seen from the figure, the PCSGE (V3) technique is able achieve the best mean rank, followed by the PCSGE (V2) technique and the BMA SVR technique. The Ens. HoLR has the worst overall rank. The result of the Friedman test (see Section 2.8.3) indicate that it is very likely that the mean ranks are different as the Friedman p value is smaller than the significance level of $\alpha = 0.05$. When performing the Nemenyi post-hoc test (also described in Section 2.8.3), it can be stated that the PCSGE (V3) technique is significantly better regarding the mean rank than the comparison techniques, as it exceeds the critical distance regarding the difference in the mean rank of 0.88. However, the mean ranks of the PCSGE (V2), BMA SVR, and PCSGE (V1) techniques are not significantly different. This can also be stated for the group of HeSVR, Ens. HeSVR, and ELMQR.

7.3.3 Case Study: Intraday Performance on Single and Multiple Weather Forecasting Models

This case study analyzes the forecasting performance for an intraday forecast ($k_{\text{min}} = 1$ h, $k_{\text{max}} = 24$ h, $\Delta = 1$ h). Therefore, when using a multi-model ensemble technique, all four available weather forecasting models (one intraday weather forecasting model, three day-ahead weather forecasting models) can be used. The evaluation methodology is the one laid out in Section 7.3.1. In Table 7.2, the error distribution of the CRPS regarding the maximum and minimum error, the 10 % and 90 % quantile, and the median error are denoted. Furthermore, the mean error, the standard deviation of errors, and the skill score are given.

The table indicates a similar result as the experiment of Section 7.3.2. As can be seen from Tables 7.1 and 7.2, all models benefit from the inclusion of the intraday weather forecasting model. HeSVR and ELMQR perform roughly equally well throughout the distribution of errors except for the minimum error. The skill of the techniques in comparison to the HoLR

Table 7.2: Performance comparison regarding the distribution of CRPS scores, the mean error, the standard deviation of errors, and the skill score of the experiments for intraday forecasting of Section 7.3.2. The color coding indicates the quality of each power forecasting technique from high quality (green) to low quality (red). Bold text highlights the best achieved score for a single and multiple weather models (WM) individually.

	Single WM				**Multiple WM**				
	HoLR	HeSVR	ELMQR	PCSGE (V1)	Ens. HoLR	Ens. HeSVR	BMA SVR	PCSGE (V2)	PCSGE (V3)
Max	0.105	0.090	0.091	**0.086**	0.107	0.092	0.088	0.086	**0.085**
90% Quant.	0.092	0.076	0.076	**0.073**	0.096	0.078	0.074	0.071	**0.070**
Median	0.070	**0.055**	**0.055**	**0.055**	0.074	0.056	0.053	0.053	**0.052**
10% Quant.	0.050	0.038	0.037	**0.036**	0.050	0.037	**0.035**	**0.035**	**0.035**
Min	0.037	0.026	**0.020**	0.022	0.038	0.027	0.021	0.023	**0.020**
Mean	0.070	0.056	0.056	**0.054**	0.073	0.056	0.053	0.052	**0.051**
Std. Dev.	0.015	**0.014**	0.015	**0.014**	0.016	0.015	0.014	**0.013**	**0.013**
Skill	0.0%	20.6%	20.4%	**23.2%**	-3.9%	19.8%	24.3%	25.4%	**27.1%**

technique improves by 20.6 % and 20.4 %, respectively. Similar to the day-ahead experiments, the PCSGE (V1) technique is able to outperform all other evaluated probabilistic forecasting models that are based on a single weather forecasting model.

Interestingly, the performance of both the Ens. HoLR and the Ens. HeSVR technique again decreases (similar to the day-ahead experiments) when using day-ahead forecasting models in addition to their counterparts that are solely based on the intraday forecasts (HoLR and HeSVR). This may be due to the static weighting in the ensembles that is particularly dominant for this forecasting task as there is one very strong weather forecasting model (the intraday weather model). The BMA SVR technique, however, is able to better utilize the information present in the ensemble and is able to yield a skill of 24.3 %. The performance of BMA SVR is only exceeded by the PCSGE (V2) and PCSGE (V3) techniques. In particular the PCSGE (V3) technique is able to yield the overall best results with 27.1 % improvement of the forecasting skill.

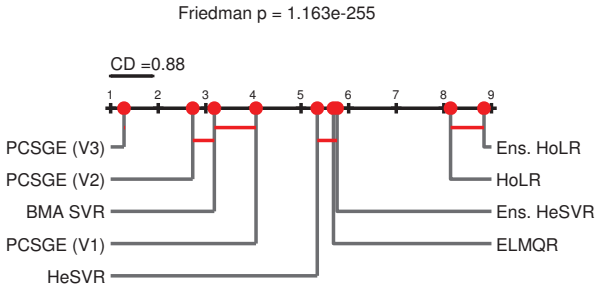


Figure 7.5: Evaluation of ranked performance of the intraday forecast using the Nemenyi post-hoc test. The algorithms have to exceed the value of the critical distance (CD) in order to be assumed to have a statistically different mean rank. As can be seen, the CSGE (V3) technique has the best ranked performance and is significantly different from the other comparison techniques on $\alpha = 0.05$.

The ranked performance is investigated in Fig. 7.5. As can be seen from the figure, the PCSGE (V3) technique has the overall best rank, followed by the PCSGE (V2) and the BMA SVR technique. HeSVR and ELMQR again perform similarly regarding the mean rank, while HoLR and Ens. HoLR perform worst. The Friedman p value indicates that the mean ranks are not identical. Using the Nemenyi post-hoc test it can be stated that the PCSGE (V3) technique is significantly better than all other forecasting models. PCSGE (V2) and BMA SVR are not significantly different, however, PCSGE (V2) has a significant difference regarding the mean rank in comparison to PCSGE (V1).

A comparison of the performance of a day-ahead and an intraday forecast regarding the mean CRPS is given in Table 7.3. On average, the performance on the intraday time horizon is 6.9 % better than the performance of the day-ahead forecast. As expected, all techniques can improve upon their forecasting quality in comparison to the day-ahead forecast. The most significant improvements can be observed for the multi-model ensemble techniques, namely BMA SVR, PCSGE (V2), and PCSGE (V3), each with over 8 % improvement.

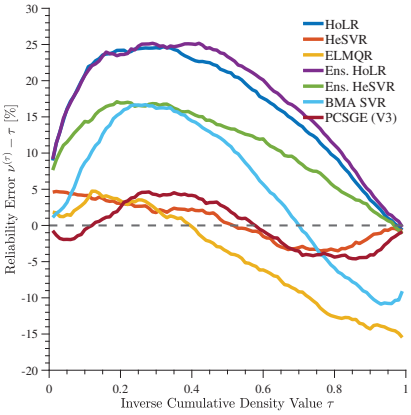
Table 7.3: Performance comparison of probabilistic day-ahead and intraday forecasts regarding the improvement of the mean value of the CRPS of the day-ahead forecast to the intraday forecast.

	Single WM				**Multiple WM**				
	HoLR	HeSVR	ELMQR	PCSGE (V1)	Ens. HoLR	Ens. HeSVR	BMA SVR	PCSGE (V2)	PCSGE (V3)
Day-Ahead	0.075	0.060	0.061	0.058	0.076	0.060	0.058	0.057	0.056
Intraday	0.070	0.056	0.056	0.054	0.073	0.056	0.053	0.052	0.051
Skill	5.7%	6.5%	7.6%	7.0%	4.2%	6.3%	8.2%	8.2%	8.5%

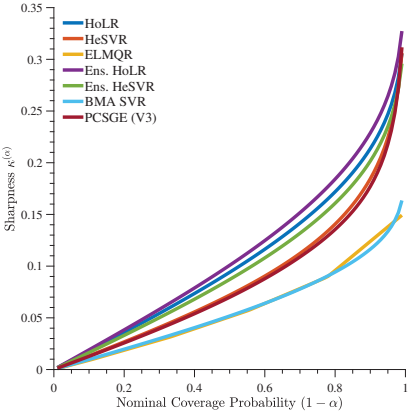
7.3.4 Case Study: Reliability and Sharpness Properties

In addition to the evaluation of the forecasting quality regarding probabilistic scoring rules, we aim to perform some diagnostic evaluation using the measures of reliability and sharpness. This section describes the reliability and sharpness properties of two wind farms using a set of the forecasting algorithms which have also been utilized in Section 7.3.1 on the intraday forecasting horizon. For the investigation of reliability, we use the visual verification techniques presented in Section 5.4. This means that for reliability, we use the modified version of the QQ plot (reliability diagram) of Section 5.4.1, for sharpness, the sharpness diagram of Section 5.4.2 is utilized. Both diagrams show the average values of all forecasts and observations of the investigated data set. Again, it should be noted that these measures "only" give an insight to characteristics of forecasting models and should not be interpreted in the same way as an actual quality estimate of a scoring rule such as the CRPS.

The first case study shows the reliability and sharpness values of windfarm $wf5$ of the EuropeWindFarm data set. The two diagrams are shown in Fig. 7.6. The reliability diagram of Fig. 7.6.1 shows the reliability error for each quantile value τ. To recall, the optimum value of the reliability is reached if the observed frequency of occurrence equals the assumed probability τ of an observation being below the quantile forecast, which is visualized in the form of the dotted line with value of 0 in the diagram. As can be seen in the diagram, some of the utilized forecasting models, in particular the homoscedastic forecasting models, expose a significant reliability error. While the techniques estimate the extreme quantiles reasonably well, they exhibit a significant error regarding all other quantile values. As can



(1) Reliability diagram of *wf5*.

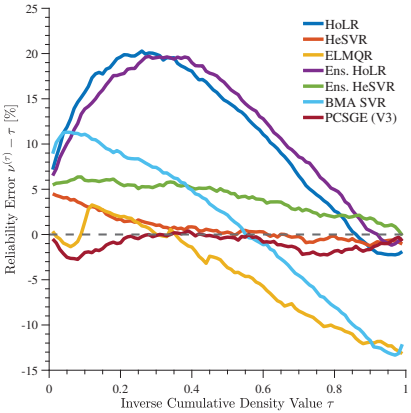


(2) Sharpness diagram of *wf5*.

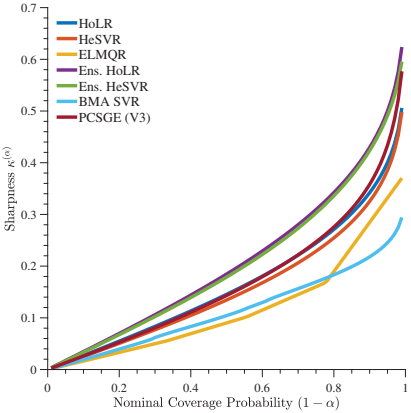
Figure 7.6: Reliability and sharpness diagrams of the wind farm *wf5*. As can be seen, some of the utilized algorithms (e.g., HoLR, BMA) are not able to create reliable predictions on all probability levels. ELMQR creates very sharp forecasts but makes reliability errors, which indicate an overly sharp forecast. HeSVR and PCSGE (V3) create the most reliable forecasts.

be seen in the sharpness diagram of Fig. 7.6.2, these techniques also create the least sharp distributions. This is somewhat understandable, as the normal distribution assumption of these techniques are an oversimplification of the actual distribution of observations. The HeSVR technique, on the other hand, specifies a significant improvement and is among the most reliable models. This means that a heteroscedastic Beta distribution is well suited to represent the distribution of power measurements of this wind farm in combination with the SVR technique. It furthermore is able to create sharper forecasts. In agreement with Table 7.2, the weighted HeSVR ensemble yields worse results than the HeSVR technique on the single best weather model only. This error can be attributed to both reliability and sharpness errors. The nonparametric ELMQR technique shows a reliability error that gets worse with increasing value of τ. This means that the highest quantile forecasts (e.g., the quantile forecasts $\hat{y}^{(0.6)}, \ldots, \hat{y}^{(0.95)}$) issue too low forecasts. On the other hand, ELMQR creates very sharp forecasting distributions. A very sharp model with reliability error indicates that ELMQR creates overly sharp forecasts in this case as the quantile forecasts are somewhat overfit to the training data set. The same oversharpening characteristics can be stated for the BMA SVR technique, which creates very sharp forecasts with reliability error. While all other ensemble techniques (even the BMA technique) exhibit a reliability error, the PCSGE (V3) technique is better able to create reliable forecasts. The sharpness slightly exceeds the sharpness of the HeSVR technique on all coverage probability levels.

For the second case study, the reliability and sharpness diagrams for the wind farm *wf1* is shown in Fig. 7.7. It can be observed in Fig. 7.7.2 that all forecasting models create less sharp forecasts in comparison to the case study of *wf5*. This indicates that it is harder to create accurate (and sharp) forecasts for this data set. Regarding the individual performance of the forecasting models it can again be observed that the HoLR and Ens. HoLR techniques are not able to yield reliable results. The HeSVR technique again is able to yield sharp and



(1) Reliability diagram of *wf1*.



(2) Sharpness diagram of *wf1*.

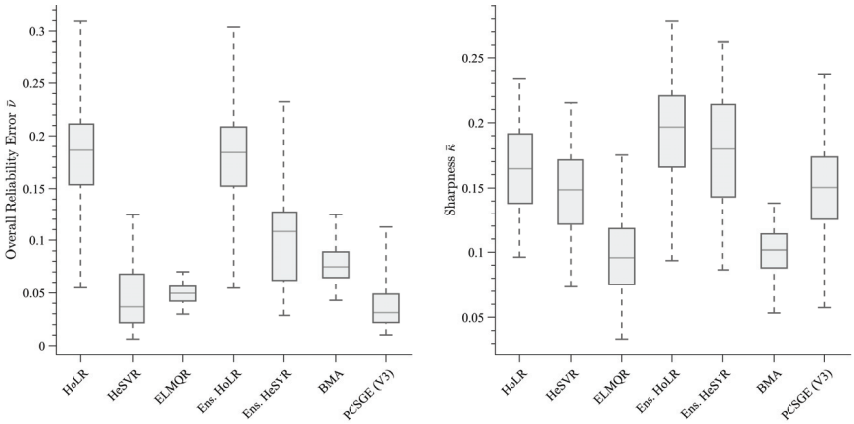
Figure 7.7: Reliability and sharpness diagrams of the wind farm *wf1*. Ens. HeSVR creates a biased forecast, as the issued forecasts are too high on all probability levels. ELMQR and BMA expose the same tendency to create overly sharp, but unreliable forecasts. HeSVR and PCSGE (V3) again create the most reliable forecasts.

reliable results and is even able to exceed the sharpness of the PCSGE technique. There still is a low reliability error for low cdf values of up to 4%. The Ens. HeSVR technique creates *biased* predictions which can be seen as the predictions are too high on all probability levels. ELMQR shows the same behavior as in the first case study, which means that it creates very sharp predictions, but with a reliability error for high quantile values. BMA also exposes a similar error characteristic as in the first case study, however, the reliability error for low quantile values is around 9%. The PCSGE technique again is able to yield results with high reliability, however, performs less sharp than the HeSVR technique. It should be mentioned that, as the PCSGE technique is composed of several base predictors, its performance is always closely coupled with the base predictors it is composed of.

These two case studies show a diagnostic evaluation of the investigated forecasting models regarding reliability and sharpness. In these two case studies, both HeSVR and PCSGE (V3) achieved the seemingly best results. However, it should again be stated that reliability and sharpness should not be interpreted in the sense of a quality estimate. For instance, the two diagrams are not able to show the contribution of either measure to an overall scoring rule, the sharpness plot even does not incorporate the corresponding observations at all. In the following section, we furthermore perform an investigation of the reliability and sharpness values for all of the investigated wind farms of the EuropeWindFarm data set.

7.3.5 Case Study: Overall Reliability and Sharpness Analysis

To give an overall impression of the reliability and sharpness properties, this experiment computes the overall reliability $\bar{\nu}$ (computed using Eq. 5.11) and sharpness $\bar{\kappa}$ (computed using Eq. 5.15) for all of the 37 wind farms in the EuropeWindFarm data set. The reliability values are given in Tables 7.4 while the sharpness values are shown in Tables 7.5. A graphical



(1) Box plot of overall reliability $\bar{\nu}$.

(2) Box plot of overall sharpness $\bar{\kappa}$.

Figure 7.8: Box plots of overall reliability $\bar{\nu}$ in Fig. 7.8.1 and overall sharpness $\bar{\kappa}$ in Fig. 7.8.2. The results show the error distribution of the overall scores regarding each of the 37 wind farms of the EuropeWindFarm data sets. HeSVR and PCSGE (V3) achieve the best results regarding overall reliability, whereas ELMQR and BMA yield the sharpest predictions.

visualization of the errors is shown in Fig. 7.8.

In Fig. 7.8.1 the distribution of overall reliability values $\bar{\nu}$ for all 37 wind farms is given. As can be seen from the figure and Table 7.4, HoLR and Ens. HoLR achieve the worst reliability values, both in the range of approximately 0.055 and 0.31. HeSVR achieves the overall minimum reliability value and the minimum value of the 10% quantile of the 37 reliability values. ELMQR shows a very consistent behavior regarding the reliability values and yields the best maximum reliability error and by far the smallest spread of reliability errors. Ens. HeSVR yields worse reliability errors in comparison to HeSVR that is based on the single best weather model. BMA, while yielding strong results regarding the scoring rule values (see, e.g., Table 7.2), only exhibits average reliability errors. The PCSGE (V3) technique yields the best reliability error values regarding the mean reliability error, the median reliability error and the 90% quantile of the reliability errors.

Table 7.4: Distribution of overall reliability errors $\bar{\nu}$ of the 37 wind farms of EuropeWindfarm. The color coding indicates the quality of each power forecasting technique from high quality (green) to low quality (red). Bold text highlights the best achieved score.

	Single WM			**Multiple WM**			
	HoLR	HeSVR	ELMQR	Ens. HoLR	Ens. HeSVR	BMA SVR	PCSGE (V3)
Max	0.310	0.124	**0.070**	0.304	0.232	0.124	0.113
90% Quant.	0.242	0.089	0.063	0.236	0.160	0.098	**0.061**
Median	0.186	0.037	0.050	0.184	0.108	0.075	**0.031**
10% Quant.	0.119	**0.013**	0.034	0.121	0.041	0.053	0.014
Min	0.055	**0.006**	0.030	0.055	0.028	0.043	0.010
Mean	0.182	0.045	0.049	0.180	0.100	0.077	**0.036**
Std. Dev.	0.053	0.029	**0.010**	0.053	0.049	0.017	0.021

The overall distribution of sharpness values is given in Fig. 7.8.2 and Table 7.5. As already has been observed in the case studies of Section 7.3.4, ELMQR and BMA yield the sharpest forecasts, where ELMQR achieves the best results regarding the minimum sharpness, 10 % quantile of sharpness values, and the mean and median sharpness values. BMA, on the other hand, achieves the best results regarding the 90 % quantile of sharpness values, the maximum sharpness value, and the standard deviation of sharpness values. HoLR, Ens. HoLR and Ens. HeSVR yield the worst sharpness values, while HeSVR and PCSGE (V3) achieve average results regarding the median and mean sharpness values.

Table 7.5: Distribution of overall sharpness values $\tilde{\kappa}$ of the 37 wind farms of EuropeWindfarm. The color coding indicates the value of $\tilde{\kappa}$ of each power forecasting technique from sharp (green) to not sharp (red). Bold text highlights the lowest achieved sharpness.

	Single WM			**Multiple WM**			
	HoLR	HeSVR	ELMQR	Ens. HoLR	Ens. HeSVR	BMA SVR	PCSGE (V3)
Max	0.233	0.215	0.175	0.278	0.262	**0.138**	0.237
90% Quant.	0.224	0.207	0.141	0.254	0.233	**0.126**	0.208
Median	0.165	0.149	**0.096**	0.197	0.180	0.102	0.150
10% Quant.	0.122	0.090	**0.058**	0.138	0.116	0.073	0.092
Min	0.096	0.074	**0.033**	0.094	0.086	0.053	0.057
Mean	0.166	0.146	**0.098**	0.193	0.178	0.099	0.150
Std. Dev.	0.037	0.039	0.031	0.044	0.044	**0.021**	0.041

These results give some expectation of the spread of the predictive distributions of the individual forecasting algorithms and the expected soundness of the created distributions. These two measures complement the results of Section 7.3.3. In combination with the results of the skill score, an expected behavior can be expressed:

- HoLR and Ens. HoLR: Both the forecast of the expectation value and the assumed parametric distribution are too simple to create an accurate probabilistic forecast.
- HeSVR: The form of the heteroscedastic parametric distribution in combination with SVR is well suited for creating accurate and statistically sound forecasts.
- Ens. HeSVR: In an intraday scenario, the additional introduced models irritate the creation of accurate predictive distributions. This can be observed regarding a decrease of all relevant measures (scoring rule, reliability, and sharpness).
- BMA SVR: The overall high quality of the CPRS is achieved by creating strong point estimates. This leads to the creation of very narrow forecasting distributions, which, however, should not be used for decision making tasks, as the issued spread is not sound on many probability levels.
- ELMQR: The overall performance of ELMQR is comparable to HeSVR, however, the way this performance is achieved is more similar to the BMA SVR technique that to HeSVR. Consequently, ELMQR creates very sharp forecasts that, however, are not statistically sound on particularly high probability levels.
- PCSGE (V3): In contrast to the BMA SVR technique, PCSGE (V3) focuses on creating broader, but statistically very sound forecasts that leads to an overall optimum result regarding the skill score.

7.4 Discussion and Conclusion of this Section

This section presents an extension of the CSGE technique to probabilistic forecasts which is called the probabilistic coopetitive soft gating ensemble (PCSGE). The technique extends the CSGE by using a scheme for the combination of probabilistic forecasts and including probabilistic scoring rules as optimization function for the ensemble training. Therein, the innovations of the CSGE technique are also utilized in the PCSGE technique. To recall, these innovations are (1) a hierarchical ensemble structure for the aggregation of multiple weather forecasting models and power forecasting models, (2) a novel weighting function with a low number of parameters, and (3) a multi-scheme weighting technique that weights the ensemble members dynamically by their overall quality, by their lead time-dependent quality, and their weather-situation dependent quality. It thus also has the same robust failure modes as the CSGE technique that retains the best possible performance when only some of the forecasting models are present.

In the experiments, we analyze the performance of the PCSGE technique in comparison to other state of the art probabilistic forecasting models and ensembles. It could be observed that the PCSGE technique outperforms the comparison techniques and is able to yield statistically significant better mean ranks than the comparison techniques over 185 experiments for each forecasting model. The 185 experiments are performed on the 37 data sets of wind farms in Europe. We have further shown that, unlike for point forecasts, for probabilistic forecasts sophisticated combination models are needed in order to achieve performance improvements when using multi-model ensembles in comparison to the single best weather forecasting model. In an analysis regarding reliability and sharpness it could be discovered that the proposed technique is among the most reliable of the investigated probabilistic forecasting models.

Chapter 8

Conclusion

This section summarizes the contents of this thesis in Section 8.1 and gives an outlook on future areas of research in Section 8.2.

8.1 Summary of Contents of this Thesis

This thesis investigates performance measures and ensemble architectures for deterministic as well as probabilistic forecasts with particular focus on the area of wind power forecasting and presents a novel ensemble technique that is able to be flexibly utilized for both, deterministic and probabilistic forecasts.

First, the fundamentals of numerical weather predictions, deterministic and probabilistic wind power forecasting, and ensemble techniques are laid out. For deterministic performance assessment, a novel categorization of error scores by basic error measure and normalization technique is introduced that simplifies the process of choosing an appropriate error score for a target application. Furthermore, the characteristics of commonly used error scores regarding their discrimination and abstraction abilities are investigated.

A novel ensemble technique called coopetitive soft gating ensemble (CSGE) is proposed. The technique is able to aggregate forecasts from a varying number of weather forecasting models and power forecasting models to an overall refined forecast. The main innovations of the technique are the weather situation and lead time-dependent dynamic weighting of the individual models using a novel low-dimensional weighting function. As has been shown in the experimental evaluation, the technique is able to outperform other state of the art power forecasting models and ensemble techniques.

In the probabilistic domain, the most popular forms of representation of probabilistic forecasts are analyzed. A scheme for the evaluation of probabilistic forecasts on a common basis is proposed by converting the multitude of existing forms of probabilistic forecasts to probability distributions. In a consecutive step, the most important forms of evaluation metrics for probability distributions are analyzed regarding particular types of occurring errors. Furthermore, the characteristics of the decomposition of these metrics are analyzed.

Based on an analysis of the probability distributions and error metrics, the coopetitive soft gating ensemble is extended to be able to be applied for probabilistic forecasts. As is being shown in the experimental evaluation, this probabilistic coopetitive soft gating ensemble (PCSGE) retains the same flexible structure of the CSGE technique and is able to yield superior results in comparison to a number of other state of the art probabilistic forecasting techniques and meteorological ensembles.

8.2 Directions of Future Research

This thesis presented a number of analyses and techniques that exceed the state of the art in deterministic and probabilistic power forecasting. Naturally, there still are a number of open questions that have not yet been addressed and that may be the goal of future research. In the following, some possible future research directions are discussed.

8.2.1 Deterministic and Probabilistic Error Measures

In the area of error measures for deterministic forecasts, the role of discrimination and abstraction of error scores has been analyzed. The discrimination and abstraction ability of scoring rules in the probabilistic domain, however, has yet to be investigated. Future research for probabilistic error scores may be conducted by introducing additional forms of normalization for enabling an even better comparability of wind farms. The normalization terms frequently used for deterministic error scores can be utilized to extend probabilistic scoring rules. Thereby, the comparability of scores reported on the basis of different data sets may increase, such as it is the case for deterministic scores. In particular the combination with interval or quantile score based measures may be of interest to better estimate extreme quantiles and thus determine extreme errors. A more detailed analysis of these properties may further increase the understandability and thus employment of probabilistic forecasts. Furthermore, the suitability of the investigated scoring rules considering their decomposition for the inclusion during model training of machine learning algorithms may be of further interest, e.g., to put more emphasis on the reliability or sharpness property to yield sharper, or more reliable forecasts.

For both deterministic and probabilistic error measures, the analyzed error measures have yet only been utilized for model evaluation and not for model training. Using other error metrics for model training may influence the model structure to show certain characteristics, such as

- being less prone to model overfitting to outliers,
- being able to fit the model more precisely by reducing the impact of samples with high errors that may arise, e.g., at weather ramps, or
- being able to quantify extreme errors by using quantile score based error functions.

The inclusion of a number of models trained with different error functions for the use as base predictors in an ensemble may also turn out advantageous.

8.2.2 Ensemble Techniques

In this thesis, an ensemble technique for the dynamic weighting of a set of base predictors based on weather situation and lead time-dependent weighting factors has been proposed. The presented characteristics of the ensemble (such as the robustness to partly missing data for a particular forecast) may be utilized to a further extent. A number of ideas are presented in the following.

Ensemble of Opportunity

The proposed ensemble technique is able to retain its performance given only a subset of forecasts of the forecasting models it was trained on in the first place for the creation of a

forecast (e.g., if no weather data are delivered in operation or a power forecasting model fails to create an output due to invalid inputs). Given this "failure mode" of the ensemble, a forecasting system applied in operational practice can then also be designed to work as an ensemble of opportunity. A number of power forecasting models and weather forecasting models can be prepared (pretrained) for application, however, not all power forecasting models (and/or not all weather forecasting models) have to be evaluated every time a forecast is created. The number of evaluated power forecasting models can be chosen dynamically. This number could depend on one or more of the following aspects:

- The estimated difficulty of the current forecasting task, e.g., derived from the weather situation.
- The uncertainty of the current prediction derived, e.g., from inner-ensemble disagreement.
- The expected overall weight of an individual base predictor in the ensemble, i.e., leading to the pruning of models with low weight if too high costs are present.
- The criticality / importance of a precise forecast to the power grid operation.
- The necessity of fast delivery time of the forecast to a client.
- The financial cost of a reported deviation from the true power generation (as, e.g., described in Section 2.9.2).
- The available computation capacity in a computing cluster.

Cost/Reward Functions

Given the availability of (hybrid) cloud computing solutions and e.g., computing concepts such as cheap preemptible virtual machines, the employment of cost/reward functions for financial optimizations can make sense, given the trade-off of investment in computing costs and gain in quality. The above-mentioned factors can have an impact on the design of the cost/reward function. This can, e.g., be utilized in a decision-making scenario such as the one laid out in Section 2.9.2.

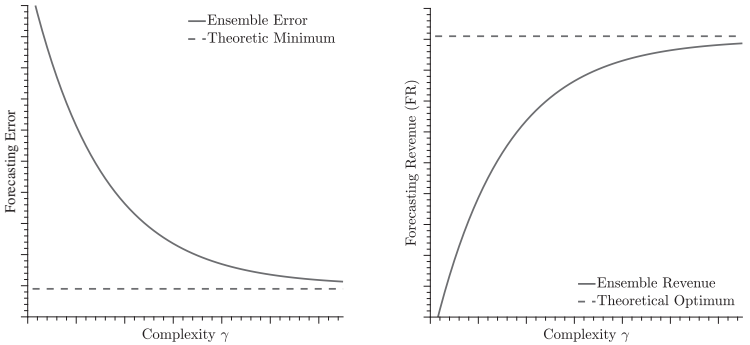
In the following, the basic principle of such a cost/reward function is sketched. As has been shown in the experiments in Chapters 4 and 7, the inclusion of more forecasting models can improve the forecasting quality. However, it can be assumed that the benefit of including additional models (which increases the complexity denoted here as γ of the ensemble model) converges to a point where the intrinsic uncertainty of a forecast is reached such as shown in Fig. 8.1.1, thus no further improvement can be achieved when including additional forecasting models in an ensemble.

If such a model is utilized in an operational forecasting system, e.g., for electricity trading, the forecasting revenue (FR) can be described based on Eq. 2.35 that now depends on the complexity γ in the way

$$\mathrm{FR}(\gamma) = \xi(o_{t+k}, \bar{\hat{y}}_{t+k|t}(\gamma)), \tag{8.1}$$

where ξ is the function of Eq. 2.35 that describes the revenue given a forecast. A schematic development of FR is shown in Fig. 8.1.2 which again converges towards a theoretical optimum of the forecasting revenue as the quality improvement converges to a theoretical minimum when increasing γ.

Nowadays, it is uneconomical for many companies to operate own computing clusters due to the high costs of equipment acquisition and maintenance which also led to the widespread adoption of cheaper, flexible, and scalable cloud computing resources. When using such



(1) Development of forecasting error. (2) Development of forecasting revenue.

Figure 8.1: Schematic representation of the forecasting error and the forecasting revenue given a varying complexity γ of the forecasting model. The assumption is that the inclusion of multiple models in an ensemble increases the forecasting quality, however, the quality gains converge towards a theoretic minimum. As the revenue (Eq. 2.35) is based on the forecasting error, it also converges to a theoretical optimum revenue.

resources for the computation of the proposed ensemble model, the computation costs (CC) which are assumed to be linear here depend on the complexity γ of the ensemble model which can then be expressed in the form of

$$\mathrm{CC}(\gamma) = C_{\text{fixed}} + \gamma \cdot C_{\text{variable}} \tag{8.2}$$

which may be composed of fixed costs C_{fixed} and variable costs C_{variable}. A visual representation of the computation costs is given in Fig. 8.2.1. The overall revenue (OR) then depends on both FR and CC in the form

$$\mathrm{OR}(\gamma) = \mathrm{FR}(\gamma) - \mathrm{CC}(\gamma), \tag{8.3}$$

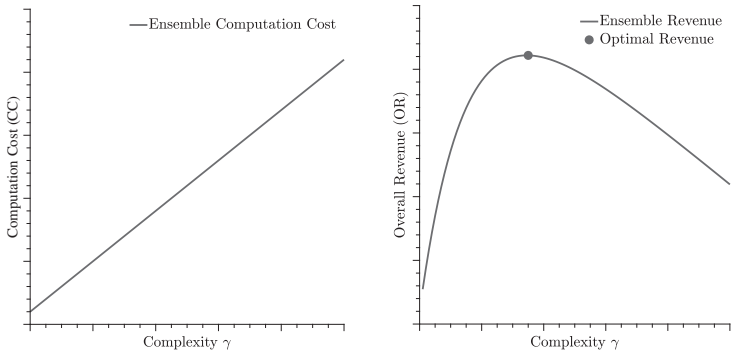
which is also visualized in Fig. 8.2.2. To optimize the overall revenue, the complexity γ with optimal OR has to be found over a time period with $n \in 1, \ldots, N$ situations for which a forecast has to be created with

$$\underset{\gamma}{\operatorname{argmax}} \sum_{n=1}^{N} \mathrm{OR}_n(\gamma), \tag{8.4}$$

which is also visualized in Fig. 8.2.2. The development of techniques to find this optimal complexity γ may be the goal of future research.

Variants of (P)CSGE

In this thesis, the similarity assessment of weather situations is based on simple nearest neighbor techniques. However, one can also think of more sophisticated techniques, such as kernel density methods. An alternative method may be to perform a reduction of the number of NWP features using feature selection or weighting of the NWP inputs which we have analyzed in [85]. Other possibilities for an improvement of the similarity assessment can be found using dimensionality reduction techniques such as principal component analysis (PCA), which we have shown to improve upon the raw data representation in a time series



(1) Development of computation cost.

(2) Development of overall revenue.

Figure 8.2: Schematic representation of the computation cost of the ensemble and the resulting overall revenue. The computation costs of Fig. 8.2.1 here are assumed to linearly depend on the complexity γ and a constant cost. The overall revenue (OR) of Fig. 8.2.2 is computed using Eq. 8.3 and consequently incorporates the forecasting revenue (FR) and the computation cost (CC). The overall revenue (OR) can be optimized when choosing an appropriate complexity γ such as shown in Fig. 8.2.2.

classification context in [87].

In this thesis, the overall optimization has been performed using a gradient-based interior point algorithm. However, one can also think of overall optimization using techniques such as simulated annealing, stochastic gradient descent, or particle swarms, possibly leading to an even better optimization.

Explicit Diversity Generation

The proposed ensemble technique is able to combine diverse base predictors to an overall refined forecast. However, while the diverse base predictors are used by the technique, the amount and type of diversity is not actively controlled. The quality of the ensemble forecast may further be improved by actively enforcing diverse base predictors. This may be introduced in a number of ways. A possibility to explicitly introduce diversity may be using standard data diversity principles, such as bootstrapping or boosting methods. This may be combined with structure or parameter diversity principles. Diverse predictors can also be generated by training them on different error metrics as mentioned in Section 8.2.1. Methods that separate data using expert knowledge may also be of interest, in which certain methodical categories may be formed, such as models which are, e.g., specialized on storm situations or snow events.

Universal Applicability

While the ensemble technique has been analyzed in the context of wind power ensemble forecasting in this thesis, the structure of the technique is generic in principle. This means that it can also be used on other domains such as other power forecasting domains (e.g., photovoltaic power forecasting), but also on entirely different domains, such as path predic-

tion of cars or vulnerable road users (VRU). The weather models that serve as inputs into the proposed ensemble technique in this thesis can therein be generalized to generic "sensors" that can be used for creating the forecast. We have already investigated the application area of path prediction of VRUs with multiple sensor inputs using neural networks and polynomial approximations in [100].

On-Line Weight Improvement

The proposed technique can be extended to update the weighting methods continuously "on-line" when observing novel model input data by permanently learning on novel observations. The respective power measurements of course have to be included in this setup as well in order to enable a meaningful feedback. The inclusion of novel observations into the ensemble has an effect in the sense that the model has to be partly (or incrementally) retrained. Depending on the chosen techniques (e.g., for the local weighting), the required complexity for the incorporation process of novel observations does vary. For instance, a simple nearest neighbor technique for local weighting does not have to be retrained, whereas other weights such as the lead time-dependent weighting do need to be recomputed. Of course, the base predictors may also benefit from more observations to be trained upon.

To reduce the computational costs, the retraining process of the ensemble should only be carried out from time to time. The particular point in time for retraining may be chosen in one of the following ways:

- Retraining based on a static schedule, e.g., the retraining is performed every month.
- Dynamic schedule based on the amount of historical data already available in the ensemble. In particular when little historical data is available, the ensemble may benefit more from the inclusion of new observations and may be conducted more frequently. Also, models with a high number of parameters may benefit from the inclusion of many samples.
- Dynamic retraining when observing exceptional weather events and events that are not yet well covered in the historic data set. This may be based on the analog ensemble technique that we have proposed in [85].
- Varying frequency of the retraining depending on the season, i.e., more frequent retraining in seasons with high variability, e.g., in the autumn.

Meteorological Event Detection

Some weather situations are exceptional in a way that makes it very hard for a forecasting algorithm to compute accurate forecasts. However, a machine learning model typically does not have judgmental capabilities regarding its output. Nowadays, forecasts from machine learning models are in some cases verified by a human expert to assess the plausibility of a forecast. However, situations that require special attention may also be detected from a computational model that performs the detection of special meteorological events.

We presented an event detection system in [82] that is able to detect points with special characteristics that may as well be used for the detection of meteorological events. The system is based on approximations that use combinations of orthogonal basis polynomials in sliding time windows that we have analyzed to be computationally very inexpensive [77, 78].

We described methods to define criteria for machine learning algorithms for these event detection models in [80].

All in all, the extensions to the proposed ensemble technique presented above may increase the real-world applicability of the techniques investigated in this thesis even further.

Appendix A

Acronym Definitions

This section defines common acronyms that are used throughout this thesis.

AE	Analog Ensemble
AKD	Affine Kernel Dressing
ANN	Artificial Neural Network
ARMA	Autoregressive Moving Average
BMA	Bayesian Model Averaging
BMWi	German Federal Ministry for Economic Affairs and Energy (Bundesministerium für Wirtschaft und Energie)
COP21	2015 United Nations Climate Change Conference
cdf	Cumulative Density Function
CSG	Coopetitive Soft Gating
CSGE	Coopetitive Soft Gating Ensemble
CRPS	Continuous Ranked Probability Score
DBN	Deep Belief Network
DWD	German Meteorological Office (Deutscher Wetterdienst)
ECC	Ensemble Copula Coupling
EEX	European Energy Exchange
ECMWF	European Centre for Medium-Range Weather Forecasts
EEG	German Renewable Energy Sources Act (Erneuerbare Energien Gesetz)
EL	Electric Load
ELM	Extreme Learning Machine
EMOS	Ensemble Model Output Statistics
EP	Expected Profit
EPS	Ensemble Prediction System
GCM	General Circulation Model
HAWT	Horizontal-Axis Wind Turbines
IGN	Ignorance Score
IPCC	Intergovernmental Panel on Climate Change
IWES	Fraunhofer Institute for Wind Energy and Energy System Technology
KDE	Kernel Density Estimation
KNN	k-Nearest Neighbor Algorithm
LSTM	Long Short-Term Memory Networks
MKL	Multiple Kernel Learning
MLP	Multi-Layer Perceptron

MME	Multi-Model Ensemble
MOS	Model Output Statistics
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
NWP	Numerical Weather Prediction
OC	Operating Cost
PCA	Principal Component Analysis
PME	Power Forecasting Model Ensemble
PCSGE	Coopetitive Soft Gating Ensemble
pdf	Probability Density Function
RE	Renewable Energies
SDE	Standard Deviation of Errors
SVM	Support Vector Machine
SVR	Support Vector Regression
PI	Prediction Interval
PIT	Probability Integral Transform
PM	Power Forecasting Model
PV	Photovoltaic Power
PSO	Particle Swarm Optimization
QR	Quantile Regression
QS	Quantile Score
UC	Unit Commitment
TLE	Time-Lagged Ensemble
WM	Weather Forecasting Model
WMO	World Meteorological Organization

Appendix B

Definition of Mathematical Symbols

This section describes the mathematical symbols which have a constant meaning throughout this thesis.

c_p	Aerodynamical rotor efficiency coefficient
d	Number of dimensions with $d = 1, \ldots, D$
e	Error value
j	Model identifier in an ensemble with $j = 1, \ldots, J$
k	Lead time with $k \in [k_{\min}, k_{\max}]$
n	Number of evaluated items with $n \in 1, \ldots, N$
o	Measured (true) power value
$o_{\text{inst.}}$	Installed nominal capacity of RE power plant
S	Scoring function
t	Point in time, in many cases the forecasting origin
w	Weighting value in an ensemble
$w^{(\phi\|\psi)}$	Weighting of PM ϕ computed on WM ψ
$\hat{p}(y)$	Pdf of generated power
$\mathbf{x}$	Numeric weather prediction (NWP)
$\mathbf{x}_{t+k\|t}$	NWP for point in time $t+k$ created at forecasting origin t
y	Generated power
$\hat{y}$	Forecast of generated power
$\hat{y}^{(\tau)}$	Quantile forecast of generated power
Δ	Time step
ϵ	Epsilon (very small value to avoid, e.g., division by zero)
ζ	Regularization parameter of (P)CSGE technique
η	Weighting parameter of CSG technique
$\boldsymbol{\theta}$	Governing parameter vector of predictive model
ξ	Overall economic revenue function
σ	Standard deviation or spread parameter
τ	Quantile identifier with $\tau = [0, 1]$
ϕ	Weather model identifier in ensemble with $\phi = 1, \ldots, \Phi$
ψ	Power forecasting model identifier with $\psi = 1, \ldots, \Psi$
υ	Time-lag indicator

Appendix C

Full Results of Error Score Distribution Comparison

This section contains supplementary information for the experiments regarding the error score distribution of Section 3.6. Table C.1 therein contains the modification of error scores which are used to create Fig. 3.1, while Table 3.5 is computed from Table C.2.

Table C.1: Overview of the artificial modification formulas used for the modification of original error results in order to achieve distributions with different characteristics, e.g., biased or skewed distributions from which Fig. 3.1 is created. Therein, $\hat{y}_n$ describes the n-th forecast, o_n denotes the corresponding observation, $\bar{o}$ is the mean observation, sgn is the sign function.

Error Distributions	**Artificial Modification Formula**
unmodified original	$\hat{y}_n^{\text{mod}} = \lvert\hat{y}_n\rvert$
Biased distribution	$\hat{y}_n^{\text{mod}} = \lvert\hat{y}_n + 0.1\rvert$
Skewed distribution	$\hat{y}_n^{\text{mod}} = \lvert\hat{y}_n - 0.3 \cdot (o_n - \bar{o})\rvert$
More spread	$\hat{y}_n^{\text{mod}} = \lvert\hat{y}_n \cdot 1.5\rvert$
Different kurtosis	$\hat{y}_n^{\text{mod}} = \lvert\text{sgn}(\hat{y}_n) \cdot 1.33 \cdot \lvert\hat{y}_n\rvert^{1.2}\rvert$

Table C.2: Error scores computed from the error distributions of Fig. 3.1. The value of each score is denoted for each of the error distributions. The colors denote the size of the respective error, where green means low error and yellow represents a high error. The NMSE is the variant which is in [38]. From the raw values of this table the relative changes in Table 3.5 are computed.

#	Error Distributions	Bias	MAE	RMSE	MSE	SDE	R^2	mRSE	KL	MASE	NMSE	MAPE
0	unmodified original	0.006	0.085	0.120	0.014	0.119	0.740	0.481	0.510	1.834	4.34	358.6
1	Biased distribution	0.106	0.132	0.160	0.025	0.119	0.537	0.655	0.681	2.856	2.35	1111.0
2	Skewed distribution	0.006	0.122	0.159	0.025	0.159	0.540	0.644	0.678	2.629	4.38	863.8
3	More spread (* 1.5)	0.010	0.126	0.176	0.031	0.176	0.433	0.711	0.753	2.714	9.74	536.4
4	Different kurtosis	0.004	0.076	0.119	0.014	0.119	0.741	0.479	0.509	1.650	3.58	285.1

Appendix D

Full Results of Analysis of Deterministic Error Scores

The following tables show the detailed results of the analysis of the EuropeWindFarm data set (see [76]) which are used as basis for Figs. 3.3 and 3.4 in Section 3. Table D.1 shows the scores of an extreme learning machine (ELM) on the EuropeWindfarm data set. Table D.2 shows the results of a simple linear regression model on the same data set. More details on the evaluation process can be found in Section 3.7. The tables show the results of the error scores for each wind farm. The colors denote the relative quality of each measure from low error (green) to high error (red). For an easier inspection of the error scores, the absolute value of the bias is given (A. Bias). A detailed analysis of the results of these tables are given in Sections 3.7 and 3.8.

Table D.1: Error scores for the EuropeWindFarm dataset using an extreme learning machine forecasting model. The colors indicate the error score values from low (green) to high (red). The last rows denote the average value of each score, their standard deviations, and the percent-wise difference.

Data	A.Bias	MAE	RMSE	MSE	SDE	R^2	mRSE	KL	MASE	NMSE	MAPE
wf1	0.011	0.082	0.115	0.013	0.114	0.719	0.497	0.530	1.866	0.328	340.5
wf2	0.032	0.138	0.181	0.033	0.179	0.566	0.609	0.659	1.828	0.647	705.3
wf3	0.000	0.062	0.100	0.010	0.100	0.760	0.441	0.490	1.649	0.439	1033.3
wf4	0.017	0.084	0.116	0.013	0.115	0.755	0.468	0.495	1.806	3.232	425.3
wf5	0.006	0.129	0.207	0.043	0.206	0.306	0.800	0.833	2.662	0.997	451.6
wf6	0.062	0.181	0.286	0.082	0.279	0.197	0.841	1.094	1.576	1.332	312.2
wf7	0.006	0.132	0.182	0.033	0.182	0.685	0.531	0.561	2.007	0.395	157.6
wf8	0.002	0.113	0.158	0.025	0.158	0.702	0.513	0.546	1.948	0.252	160.4
wf9	0.006	0.040	0.071	0.005	0.071	0.561	0.550	0.663	1.677	0.133	292.5
wf10	0.009	0.103	0.138	0.019	0.137	0.719	0.499	0.530	1.948	0.654	421.5
wf11	0.039	0.129	0.189	0.036	0.185	0.639	0.566	0.601	2.147	0.357	385.7
wf12	0.041	0.113	0.165	0.027	0.160	0.651	0.548	0.590	1.859	0.500	1178.0
wf13	0.017	0.063	0.093	0.009	0.091	0.691	0.502	0.556	1.742	0.311	465.1
wf14	0.061	0.103	0.151	0.023	0.138	0.496	0.650	0.710	2.069	0.450	554.5
wf15	0.017	0.079	0.115	0.013	0.114	0.704	0.499	0.544	1.639	0.291	1044.9
wf16	0.006	0.075	0.108	0.012	0.107	0.721	0.481	0.528	1.622	0.233	928.3
wf17	0.010	0.097	0.143	0.020	0.143	0.643	0.552	0.598	1.862	0.650	631.3
wf18	0.001	0.066	0.105	0.011	0.105	0.707	0.498	0.541	1.711	0.290	261.3
wf19	0.002	0.085	0.126	0.016	0.126	0.759	0.458	0.491	1.714	0.243	178.5
wf20	0.019	0.135	0.192	0.037	0.192	0.609	0.587	0.626	2.191	0.423	453.8
wf21	0.008	0.121	0.168	0.028	0.168	0.615	0.568	0.621	1.805	0.350	3670.4
wf22	0.006	0.061	0.090	0.008	0.090	0.695	0.493	0.552	1.668	0.222	1561.8
wf23	0.005	0.099	0.139	0.019	0.138	0.650	0.555	0.592	2.119	0.482	455.9
wf24	0.012	0.077	0.113	0.013	0.112	0.623	0.566	0.614	1.914	0.207	165.2
wf25	0.005	0.087	0.123	0.015	0.123	0.630	0.559	0.608	1.933	0.309	710.8
wf26	0.004	0.103	0.156	0.024	0.156	0.577	0.595	0.650	1.898	0.365	3528.1
wf27	0.003	0.087	0.134	0.018	0.134	0.621	0.539	0.615	1.572	0.353	1282.3
wf28	0.034	0.108	0.170	0.029	0.167	0.636	0.570	0.603	2.441	0.303	5497.2
wf29	0.015	0.063	0.093	0.009	0.092	0.729	0.473	0.520	1.762	0.303	279.1
wf30	0.047	0.131	0.193	0.037	0.188	0.599	0.598	0.633	2.177	3.650	320.0
wf31	0.045	0.168	0.227	0.051	0.222	0.482	0.689	0.719	3.069	0.580	5151.7
wf32	0.008	0.099	0.146	0.021	0.146	0.695	0.506	0.552	1.788	3.635	1496.4
wf33	0.006	0.076	0.113	0.013	0.112	0.650	0.531	0.592	1.820	0.324	827.8
wf34	0.009	0.117	0.160	0.026	0.160	0.691	0.517	0.555	1.799	0.395	355.8
wf35	0.006	0.094	0.133	0.018	0.133	0.804	0.423	0.443	2.109	0.457	1564.3
wf36	0.017	0.096	0.146	0.021	0.145	0.646	0.531	0.595	1.937	0.412	1173.8
wf37	0.008	0.088	0.125	0.016	0.125	0.718	0.484	0.531	1.646	1.043	2591.4
wf38	0.001	0.070	0.114	0.013	0.114	0.717	0.471	0.532	1.747	0.262	439.3
wf39	0.007	0.129	0.173	0.030	0.173	0.728	0.497	0.522	2.223	0.574	54769
wf40	0.034	0.088	0.140	0.020	0.136	0.658	0.535	0.585	1.898	0.483	403.4
wf41	0.005	0.065	0.114	0.013	0.114	0.504	0.603	0.705	1.892	0.237	342.7
wf42	0.023	0.158	0.206	0.043	0.205	0.526	0.634	0.689	1.819	0.765	126.2
wf43	0.010	0.083	0.118	0.014	0.117	0.835	0.384	0.406	1.691	0.324	20047
wf44	0.013	0.116	0.158	0.025	0.158	0.684	0.528	0.562	1.931	0.779	432.9
wf45	0.020	0.162	0.218	0.047	0.217	0.698	0.534	0.549	3.113	0.362	1179.9
Avg.	0.016	0.101	0.147	0.023	0.145	0.636	0.544	0.594	1.940	0.652	2639.0
Std.	0.016	0.031	0.042	0.014	0.041	0.156	0.083	0.107	0.331	0.799	8431.2
%	100	30.7	28.6	60.9	28.3	24.5	15.3	19.6	17.1	122.5	319.5

Table D.2: Error scores for the EuropeWindFarm dataset using a standard linear regression model. The colors indicate the error score values from low (green) to high (red). The last rows denote the average value of each score, their standard deviations, and the percent-wise difference.

Data	A.Bias	MAE	RMSE	MSE	SDE	$R\hat{2}$	mRSE	KL	MASE	NMSE	MAPE
wf1	0.011	0.085	0.118	0.014	0.118	0.701	0.516	0.547	1.935	0.266	282.4
wf2	0.023	0.137	0.176	0.031	0.174	0.593	0.590	0.638	1.810	0.509	456.3
wf3	0.004	0.076	0.121	0.015	0.121	0.646	0.541	0.595	2.024	0.306	1341.0
wf4	0.004	0.088	0.121	0.015	0.121	0.734	0.490	0.515	1.895	4.468	343.0
wf5	0.002	0.133	0.200	0.040	0.200	0.352	0.774	0.805	2.743	0.843	451.8
wf6	0.058	0.175	0.273	0.075	0.267	0.098	0.803	1.048	1.520	0.994	323.4
wf7	0.031	0.142	0.186	0.035	0.184	0.669	0.547	0.575	2.151	0.360	175.2
wf8	0.025	0.119	0.163	0.027	0.161	0.683	0.529	0.563	2.067	0.259	170.8
wf9	0.000	0.046	0.079	0.006	0.079	0.452	0.621	0.740	1.942	0.138	453.3
wf10	0.024	0.108	0.143	0.021	0.141	0.696	0.523	0.552	2.037	0.499	370.1
wf11	0.053	0.152	0.217	0.047	0.210	0.524	0.658	0.690	2.524	0.233	462.4
wf12	0.014	0.120	0.163	0.027	0.163	0.657	0.549	0.585	1.976	0.299	1037.8
wf13	0.012	0.060	0.090	0.008	0.089	0.708	0.487	0.541	1.672	0.320	351.8
wf14	0.063	0.104	0.151	0.023	0.137	0.497	0.650	0.709	2.097	0.477	338.2
wf15	0.006	0.087	0.123	0.015	0.123	0.661	0.539	0.583	1.804	0.285	793.4
wf16	0.000	0.079	0.113	0.013	0.113	0.690	0.511	0.556	1.717	0.235	778.9
wf17	0.013	0.106	0.150	0.023	0.149	0.606	0.584	0.628	2.032	0.505	800.1
wf18	0.001	0.072	0.113	0.013	0.113	0.661	0.539	0.582	1.864	0.209	241.6
wf19	0.014	0.099	0.138	0.019	0.137	0.711	0.508	0.538	2.010	0.235	172.4
wf20	0.024	0.144	0.201	0.041	0.200	0.571	0.620	0.655	2.337	0.360	457.8
wf21	0.004	0.120	0.166	0.027	0.165	0.628	0.560	0.610	1.783	0.325	2917.2
wf22	0.002	0.066	0.105	0.011	0.105	0.586	0.581	0.644	1.815	0.169	1222.6
wf23	0.013	0.097	0.141	0.020	0.141	0.636	0.568	0.603	2.085	0.220	409.1
wf24	0.019	0.077	0.115	0.013	0.113	0.608	0.577	0.626	1.913	0.124	164.6
wf25	0.003	0.089	0.128	0.016	0.127	0.601	0.580	0.632	1.992	0.193	699.4
wf26	0.002	0.106	0.151	0.023	0.151	0.605	0.578	0.629	1.950	0.273	2976.2
wf27	0.003	0.087	0.136	0.018	0.136	0.612	0.548	0.623	1.561	0.359	1174.6
wf28	0.028	0.115	0.174	0.030	0.171	0.621	0.586	0.615	2.591	0.251	6458.4
wf29	0.015	0.073	0.105	0.011	0.104	0.653	0.545	0.589	2.049	0.348	295.0
wf30	0.066	0.140	0.199	0.040	0.188	0.575	0.620	0.652	2.335	3.317	326.0
wf31	0.037	0.171	0.226	0.051	0.223	0.487	0.689	0.716	3.125	0.410	5001.2
wf32	0.012	0.101	0.148	0.022	0.147	0.690	0.515	0.557	1.832	2.415	1229.9
wf33	0.018	0.086	0.122	0.015	0.121	0.587	0.589	0.642	2.048	0.403	931.7
wf34	0.003	0.119	0.161	0.026	0.161	0.690	0.520	0.557	1.828	0.280	361.8
wf35	0.002	0.109	0.147	0.021	0.147	0.764	0.470	0.486	2.432	0.367	1826.7
wf36	0.020	0.103	0.154	0.024	0.152	0.608	0.569	0.626	2.094	0.446	1011.0
wf37	0.014	0.093	0.131	0.017	0.130	0.693	0.511	0.554	1.744	1.225	1155.1
wf38	0.007	0.079	0.129	0.017	0.129	0.639	0.540	0.601	1.973	0.243	302.0
wf39	0.002	0.133	0.185	0.034	0.185	0.689	0.526	0.558	2.288	0.536	22824.3
wf40	0.040	0.096	0.138	0.019	0.132	0.669	0.536	0.575	2.071	0.459	300.9
wf41	0.013	0.071	0.117	0.014	0.117	0.478	0.624	0.722	2.081	0.242	449.0
wf42	0.030	0.172	0.216	0.046	0.214	0.482	0.670	0.720	1.982	0.746	147.5
wf43	0.001	0.099	0.132	0.017	0.132	0.792	0.436	0.456	2.009	0.204	17937.8
wf44	0.006	0.124	0.161	0.026	0.161	0.672	0.543	0.573	2.061	0.578	381.2
wf45	0.019	0.169	0.214	0.046	0.213	0.709	0.527	0.540	3.254	0.216	898.5
Avg.	0.017	0.107	0.152	0.025	0.150	0.546	0.569	0.617	2.068	0.581	1804.5
Std.	0.017	0.031	0.039	0.013	0.038	0.139	0.070	0.094	0.344	0.813	4216.9
%	99.6	28.9	25.9	54.2	25.5	25.4	12.4	15.3	16.6	139.9	233.7

Appendix E

Proof of Location of Minimum Value of the Quantile Score

This section gives a proof of the minimum value of the quantile score (QS) being in the location of a specified quantile τ of a tuple of observations. As has been described in Eq. 5.24, the quantile score computes the error given a forecast y and an observation o with

$$\begin{aligned} \mathrm{QS}(y) &= \rho_\tau(o-y), && \text{(E.1)} \\ &= \tau \cdot |o-y| \cdot H(o-y) + (1-\tau) \cdot |o-y| \cdot H(y-o). && \text{(E.2)} \end{aligned}$$

Given a *sorted* tuple of observations $\mathbf{o} = (o_1, \ldots, o_N) \in \mathbb{R}^N$ with $n \in 1, \ldots, N$ and $o_n \leq o_{n+1}$, the QS can be written as

$$= \sum_{n=1}^{N} \tau \cdot |o_n - y| \cdot H(o_n - y) + (1-\tau) \cdot |o_n - y| \cdot H(y - o_n). \quad \text{(E.3)}$$

Assuming a observations within $\mathbf{o}$ for which $o_n < y$ is true, the QS can be written as

$$\mathrm{QS}(y) = \sum_{n=1}^{a} \Big(\tau \cdot |o_n - y| \cdot \underbrace{H(o_n - y)}_{=0} + (1-\tau) \cdot |o_n - y| \cdot \underbrace{H(y - o_n)}_{=1} \Big) \quad \text{(E.4)}$$

$$+ \sum_{n=a+1}^{N} \Big(\tau \cdot |o_n - y| \cdot \underbrace{H(o_n - y)}_{=1} + (1-\tau) \cdot |o_n - y| \cdot \underbrace{H(y - o_n)}_{=0} \Big), \quad \text{(E.5)}$$

$$= \sum_{n=1}^{a} \Big((1-\tau) \cdot |o_n - y| \Big) + \sum_{n=a+1}^{N} \Big(\tau \cdot |o_n - y| \Big). \quad \text{(E.6)}$$

As $o_n - y < 0$ in the first sum and $o_n - y \geq 0$ in the second sum is true, the term can be rewritten as

$$\mathrm{QS}(y) = \sum_{n=1}^{a} \Big((1-\tau) \cdot (y - o_n) \Big) + \sum_{n=a+1}^{N} \Big(\tau \cdot (o_n - y) \Big), \tag{E.7}$$

$$= \sum_{n=1}^{a} \Big((1-\tau) \cdot y \Big) + \underbrace{\sum_{n=1}^{a} \Big(-(1-\tau) \cdot o_n \Big)}_{=c_1} + \underbrace{\sum_{n=a+1}^{N} \Big(\tau \cdot o_n \Big)}_{=c_2} + \sum_{n=a+1}^{N} \Big(-\tau \cdot y \Big), \tag{E.8}$$

$$= a \cdot (1-\tau) \cdot y + c_1 + c_2 - (N-a) \cdot \tau \cdot y, \tag{E.9}$$

$$= ay - \cancel{a\tau y} - N\tau y + \cancel{a\tau y} + c_1 + c_2, \tag{E.10}$$

$$= ay - N\tau y + c_1 + c_2. \tag{E.11}$$

For determining the location of y with minimum error of the QS, it can be derived with

$$\frac{\mathrm{dQS}}{\mathrm{d}y} = a - N \cdot \tau, \tag{E.12}$$

the actual turning point can then be found by requiring the derivative to be zero, i.e.,

$$\frac{\mathrm{dQS}}{\mathrm{d}y} \stackrel{!}{=} 0, \tag{E.13}$$

$$0 = a - N \cdot \tau, \tag{E.14}$$

which can also be expressed as

$$a = \tau \cdot N. \tag{E.15}$$

This means that QS reaches its minimum value if the value of y is chosen so that there are $\tau \cdot N$ observations for which $o_n < y$ is true. It thereby reaches the minimum value independently of the actual values of $\mathbf{o}$, but only is influenced by the order of the elements in $\mathbf{o}$ with respect to $\tau \cdot N$.

Appendix F

Proof of Relationship of Interval Score and Combination of Quantile Scores

This section gives a proof of the relation of the interval score to the quantile score. As is described in Eq. 5.32, the relationship is said to be

$$\frac{\alpha}{2} \cdot \mathrm{IS}_\alpha(o) = \rho_{\tau_{\hat{l}}}(o - \hat{y}^{(\tau_{\hat{l}})}) + \rho_{\tau_{\hat{u}}}(o - \hat{y}^{(\tau_{\hat{u}})}). \tag{F.1}$$

To recall, the relationship of the nominal confidence with α and the quantiles $\tau_{\hat{l}}$ and $\tau_{\hat{u}}$ is

$$\tau_{\hat{l}} = 1 - \tau_{\hat{u}} = \frac{\alpha}{2}. \tag{F.2}$$

Using the above relationship of Eq. F.2, the left hand side of Eq. F.1 which for the sake of clarity is here called scaled interval score (SIS) can be denoted as

$$\begin{aligned}
\mathrm{SIS}(o) &= \tau_{\hat{l}} \cdot \mathrm{IS}_\alpha(o), && \text{(F.3)}\\
&= \tau_{\hat{l}} \cdot \Big((\hat{u} - \hat{l}) + \frac{2}{\alpha} \cdot (o - \hat{u}) \cdot H(o - \hat{u}) + \frac{2}{\alpha} \cdot (\hat{l} - o) \cdot H(\hat{l} - o) \Big), && \text{(F.4)}\\
&= \tau_{\hat{l}} \cdot \Big((\hat{u} - \hat{l}) + \frac{1}{\tau_{\hat{l}}} \cdot (o - \hat{u}) \cdot H(o - \hat{u}) + \frac{1}{\tau_{\hat{l}}} \cdot (\hat{l} - o) \cdot H(\hat{l} - o) \Big), && \text{(F.5)}\\
&= \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}) + (o - \hat{u}) \cdot H(o - \hat{u}) + (\hat{l} - o) \cdot H(\hat{l} - o). && \text{(F.6)}
\end{aligned}$$

The right hand side of Eq. F.1 denoted here as double quantile score (DQS) with $\hat{y}^{(\tau_{\hat{l}})} = \hat{l}$ and $\hat{y}^{(\tau_{\hat{u}})} = \hat{u}$ can be written as

$$\begin{aligned}
\mathrm{DQS}(o) &= \rho_{\tau_{\hat{l}}}(o - \hat{l}) + \rho_{\tau_{\hat{u}}}(o - \hat{u}), && \text{(F.7)}\\
&= \rho_{\tau_{\hat{l}}}(o - \hat{l}) + \rho_{(1-\tau_{\hat{l}})}(o - \hat{u}), && \text{(F.8)}\\
&= \underbrace{\tau_{\hat{l}} \cdot |o - \hat{l}| \cdot H(o - \hat{l}) + (1 - \tau_{\hat{l}}) \cdot |o - \hat{l}| \cdot H(\hat{l} - o)}_{\rho_{\tau_{\hat{l}}}(o - \hat{l})} && \text{(F.9)}\\
&\quad + \underbrace{(1 - \tau_{\hat{l}}) \cdot |o - \hat{u}| \cdot H(o - \hat{u}) + (1 - (1 - \tau_{\hat{l}})) \cdot |o - \hat{u}| \cdot H(\hat{u} - o)}_{\rho_{(1-\tau_{\hat{l}})}(o - \hat{u})}. && \text{(F.10)}
\end{aligned}$$

Regarding the position of the observation o, three different positional categories can be identified. The proofs of the individual categories indicated by the bold text are given in the following.

Observation below the lower interval border $(o < \hat{l})$ **:**
If the observation o is below the location of $\hat{l}$, the scores can be shown to be equal with

$$\mathrm{SIS}(o) = \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}) + (o - \hat{u}) \cdot \underbrace{H(o - \hat{u})}_{=0} + (\hat{l} - o) \cdot \underbrace{H(\hat{l} - o)}_{=1}, \tag{F.11}$$

$$= \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}) + \hat{l} - o. \tag{F.12}$$

For the DQS, the equivalence can be shown with

$$\mathrm{DQS}(o) = \tau_{\hat{l}} \cdot |o - \hat{l}| \cdot \underbrace{H(o - \hat{l})}_{=0} + (1 - \tau_{\hat{l}}) \cdot |o - \hat{l}| \cdot \underbrace{H(\hat{l} - o)}_{=1} \tag{F.13}$$

$$+ (1 - \tau_{\hat{l}}) \cdot |o - \hat{u}| \cdot \underbrace{H(o - \hat{u})}_{=0} + (1 - (1 - \tau_{\hat{l}})) \cdot |o - \hat{u}| \cdot \underbrace{H(\hat{u} - o)}_{=1}, \tag{F.14}$$

$$= (1 - \tau_{\hat{l}}) \cdot |o - \hat{l}| + (1 - (1 - \tau_{\hat{l}})) \cdot |o - \hat{u}|. \tag{F.15}$$

As $o - \hat{l}$ and $o - \hat{u}$ are both < 0, the absolute values can be rewritten as

$$= (1 - \tau_{\hat{l}}) \cdot (\hat{l} - o) + (1 - (1 - \tau_{\hat{l}})) \cdot (\hat{u} - o), \tag{F.16}$$

$$= -o + \hat{l} + \tau_{\hat{l}} o - \tau_{\hat{l}} \hat{l} - \tau_{\hat{l}} o + \tau_{\hat{l}} \hat{u}, \tag{F.17}$$

$$= -o + \hat{l} - \tau_{\hat{l}} \hat{l} + \tau_{\hat{l}} \hat{u}, \tag{F.18}$$

$$= \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}) + \hat{l} - o. \tag{F.19}$$

Observation in the interval $\hat{l} \leq o \leq \hat{u}$ **:**
For the case that the observation o is in the interval $[\hat{l}, \hat{u}]$, SIS is computed with

$$\mathrm{SIS}(o) = \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}) + (o - \hat{u}) \cdot \underbrace{H(o - \hat{u})}_{=0} + (\hat{l} - o) \cdot \underbrace{H(\hat{l} - o)}_{=0}, \tag{F.20}$$

$$= \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}). \tag{F.21}$$

The equivalence of the DQS can be shown with

$$\mathrm{DQS}(o) = \tau_{\hat{l}} \cdot |o - \hat{l}| \cdot \underbrace{H(o - \hat{l})}_{=1} + (1 - \tau_{\hat{l}}) \cdot |o - \hat{l}| \cdot \underbrace{H(\hat{l} - o)}_{=0} \tag{F.22}$$

$$+ (1 - \tau_{\hat{l}}) \cdot |o - \hat{u}| \cdot \underbrace{H(o - \hat{u})}_{=0} + (1 - (1 - \tau_{\hat{l}})) \cdot |o - \hat{u}| \cdot \underbrace{H(\hat{u} - o)}_{=1}, \tag{F.23}$$

$$= \tau_{\hat{l}} \cdot |o - \hat{l}| + (1 - (1 - \tau_{\hat{l}})) \cdot |o - \hat{u}|. \tag{F.24}$$

As $o - \hat{u}$ is < 0, the absolute value can be rewritten as

$$= \tau_{\hat{l}} \cdot (o - \hat{l}) + (1 - (1 - \tau_{\hat{l}})) \cdot (\hat{u} - o), \tag{F.25}$$

$$= \tau_{\hat{l}} o - \tau_{\hat{l}} \hat{l} + \tau_{\hat{l}} \hat{u} - \tau_{\hat{l}} o, \tag{F.26}$$

$$= \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}). \tag{F.27}$$

Observation above the upper interval border ($\hat{u} < o$) **:**
For the final case that the observation o is above the value of the upper interval border $\hat{u}$, the proof can be conducted with

$$\mathrm{SIS}(o) = \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}) + (o - \hat{u}) \cdot \underbrace{H(o - \hat{u})}_{=1} + (\hat{l} - o) \cdot \underbrace{H(\hat{l} - o)}_{=0}, \tag{F.28}$$

$$= \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}) + o - \hat{u}. \tag{F.29}$$

$$\mathrm{DQS}(o) = \tau_{\hat{l}} \cdot |o - \hat{l}| \cdot \underbrace{H(o - \hat{l})}_{=1} + (1 - \tau_{\hat{l}}) \cdot |o - \hat{l}| \cdot \underbrace{H(\hat{l} - o)}_{=0} \tag{F.30}$$

$$+ (1 - \tau_{\hat{l}}) \cdot |o - \hat{u}| \cdot \underbrace{H(o - \hat{u})}_{=1} + (1 - (1 - \tau_{\hat{l}})) \cdot |o - \hat{u}| \cdot \underbrace{H(\hat{u} - o)}_{=0}, \tag{F.31}$$

$$= \tau_{\hat{l}} \cdot |o - \hat{l}| + (1 - \tau_{\hat{l}}) \cdot |o - \hat{u}|. \tag{F.32}$$

As o is larger than $\hat{l}$ and $\hat{u}$, the absolute values can be written as

$$= \tau_{\hat{l}} \cdot (o - \hat{l}) + (1 - \tau_{\hat{l}}) \cdot (o - \hat{u}), \tag{F.33}$$

$$= \tau_{\hat{l}} o - \tau_{\hat{l}} \hat{l} + o - \hat{u} - \tau_{\hat{l}} o + \tau_{\hat{l}} \hat{u}, \tag{F.34}$$

$$= \tau_{\hat{l}} \cdot (\hat{u} - \hat{l}) + o - \hat{u}. \tag{F.35}$$

As has thus been shown, for all possible values of o, the left hand side and the right hand side of Eq. F.1 are equivalent, which gives a proof of the relationship.

Appendix G

Proof of Optimality of IS with Respect To Nominal Confidence

This section gives a proof of the convergence of the interval bounds $\hat{l}$ and $\hat{u}$ to achieve the minimum value of the interval score (IS). It also gives a proof that the nominal confidence $(1-\alpha)$ actually specifies the amount of values in the prediction interval $[\hat{l}, \hat{u}]$.

As laid out in Section 5.6.5 and Eq. 5.26, the interval score with nominal confidence $(1-\alpha)$ is defined as

$$\text{IS}_\alpha = (\hat{u} - \hat{l}) + \frac{2}{\alpha} \cdot (o - \hat{u}) \cdot H(o - \hat{u}) + \frac{2}{\alpha} \cdot (\hat{l} - o) \cdot H(\hat{l} - o) \tag{G.1}$$

As has been proven in Appendix F, this is equivalent to

$$\text{IS}_\alpha = \frac{2}{\alpha} \cdot \Big(\rho_{\tau_{\hat{l}}}(o - \hat{l}) + \rho_{\tau_{\hat{u}}}(o - \hat{u}) \Big). \tag{G.2}$$

Given the relationship

$$\tau = \tau_{\hat{l}} = 1 - \tau_{\hat{u}} = \frac{\alpha}{2} \tag{G.3}$$

that we denote here as a simple τ for the sake of better readability, we can rewrite IS_α as

$$\text{IS}_\alpha = \frac{1}{\tau} \cdot \Big(\rho_\tau(o - \hat{l}) + \rho_{(1-\tau)}(o - \hat{u}) \Big). \tag{G.4}$$

For a tuple of *sorted* observations $\mathbf{o} = (o_1, \ldots, o_N) \in \mathbb{R}^N$ with $n \in 1, \ldots, N$ with $o_n \leq o_{n+1}$, the IS can be written as

$$\text{IS}_\alpha = \sum_{n=1}^{N} \left(\frac{1}{\tau} \cdot \Big(\rho_\tau(o_n - \hat{l}) + \rho_{(1-\tau)}(o_n - \hat{u}) \Big) \right). \tag{G.5}$$

Given the structure of the IS, an observation can either be (1) below the lower interval border $o_n < \hat{l}$, (2) in the interval $\hat{l} \leq o_n \leq \hat{u}$, or (3) above the upper interval bound $\hat{u} < o_n$. The index in $\mathbf{o}$ that defines the cutting point between case (1) and (2) is denoted as a, whereas the index of the cutting point between case (2) and (3) is denoted as b.

Case (1) can then be written as

$$\mathrm{IS}_\alpha^{(1)} = \sum_{n=1}^{a} \left(\frac{1}{\tau} \cdot \Big(\tau \cdot |o_n - \hat{l}| \cdot \underbrace{H(o_n - \hat{l})}_{=0} + (1-\tau) \cdot |o_n - \hat{l}| \cdot \underbrace{H(\hat{l} - o_n)}_{=1} \right. \tag{G.6}$$

$$\left. + (1-\tau) \cdot |o_n - \hat{u}| \cdot \underbrace{H(o_n - \hat{u})}_{=0} + (1-(1-\tau)) \cdot |o_n - \hat{u}| \cdot \underbrace{H(\hat{u} - o_n)}_{=1} \Big) \right), \tag{G.7}$$

$$= \frac{1}{\tau} \sum_{n=1}^{a} \Big((1-\tau) \cdot |o_n - \hat{l}| + \tau \cdot |o_n - \hat{u}| \Big). \tag{G.8}$$

As $o_n - \hat{l}$ and $o_n - \hat{u}$ are both < 0, the absolute values can be rewritten as

$$= \frac{1}{\tau} \sum_{n=1}^{a} \Big((1-\tau) \cdot (\hat{l} - o_n) + \tau \cdot (\hat{u} - o_n) \Big), \tag{G.9}$$

which can further be simplified to

$$= \frac{1}{\tau} \sum_{n=1}^{a} \Big(\hat{l} - o_n - \tau \hat{l} + \cancel{\tau o_n} + \tau \hat{u} - \cancel{\tau o_n} \Big), \tag{G.10}$$

$$= \frac{1}{\tau} \Big(\sum_{n=1}^{a} (1-\tau) \hat{l} + \sum_{n=1}^{a} \tau \hat{u} + \underbrace{\sum_{n=1}^{a} -o_n}_{=c_1} \Big), \tag{G.11}$$

$$= \frac{1}{\tau} \Big(a(1-\tau) \hat{l} + a\tau \hat{u} + c_1 \Big). \tag{G.12}$$

Case (2) can be denoted as

$$\mathrm{IS}_\alpha^{(2)} = \sum_{n=a+1}^{b} \left(\frac{1}{\tau} \cdot \Big(\tau \cdot |o_n - \hat{l}| \cdot \underbrace{H(o_n - \hat{l})}_{=1} + (1-\tau) \cdot |o_n - \hat{l}| \cdot \underbrace{H(\hat{l} - o_n)}_{=0} \right. \tag{G.13}$$

$$\left. + (1-\tau) \cdot |o_n - \hat{u}| \cdot \underbrace{H(o_n - \hat{u})}_{=1} + (1-(1-\tau)) \cdot |o_n - \hat{u}| \cdot \underbrace{H(\hat{u} - o_n)}_{=0} \Big) \right), \tag{G.14}$$

$$= \frac{1}{\tau} \sum_{n=a+1}^{b} \Big(\tau \cdot |o_n - \hat{l}| + \tau \cdot |o_n - \hat{u}| \Big). \tag{G.15}$$

As $o_n - \hat{u}$ is < 0, the absolute value can be rewritten as

$$= \frac{1}{\tau} \sum_{n=a+1}^{b} \Big(\tau \cdot (o_n - \hat{l}) + \tau \cdot (\hat{u} - o_n) \Big), \tag{G.16}$$

which can be simplified to

$$= \frac{1}{\tau} \sum_{n=a+1}^{b} \Big(\cancel{-o_n} - \tau \hat{l} + \tau \hat{u} - \cancel{-o_n} \Big), \tag{G.17}$$

$$= \frac{\cancel{\tau}}{\cancel{\tau}} \sum_{n=a+1}^{b} \Big(\hat{u} - \hat{l} \Big), \tag{G.18}$$

$$= (b-a)(\hat{u} - \hat{l}) \tag{G.19}$$

Case (3) can be written as

$$\mathrm{IS}_{\alpha}^{(3)} = \sum_{n=b+1}^{N} \Bigg(\frac{1}{\tau} \cdot \Big(\tau \cdot |o_n - \hat{l}| \cdot \underbrace{H(o_n - \hat{l})}_{=1} + (1-\tau) \cdot |o_n - \hat{l}| \cdot \underbrace{H(\hat{l} - o_n)}_{=0} \tag{G.20}$$

$$+ (1-\tau) \cdot |o_n - \hat{u}| \cdot \underbrace{H(o_n - \hat{u})}_{=0} + (1-(1-\tau)) \cdot |o_n - \hat{u}| \cdot \underbrace{H(\hat{u} - o_n)}_{=1} \Big) \Bigg), \tag{G.21}$$

$$= \frac{1}{\tau} \sum_{n=b+1}^{N} \Big(\tau \cdot |o_n - \hat{l}| + (1-\tau) \cdot |o_n - \hat{u}| \Big). \tag{G.22}$$

As $o - \hat{l}$ and $o - \hat{u}$ are both < 0, the absolute values can be rewritten as

$$= \frac{1}{\tau} \sum_{n=b+1}^{N} \Big(\tau \cdot (o_n - \hat{l}) + (1-\tau) \cdot (o_n - \hat{u}) \Big), \tag{G.23}$$

which can be simplified to

$$= \frac{1}{\tau} \sum_{n=b+1}^{N} \Big(\cancel{-o_n} - \tau \hat{l} + o_n - \hat{u} - \cancel{-o_n} + \tau \hat{u} \Big), \tag{G.24}$$

$$= \frac{1}{\tau} \Bigg(\sum_{n=b+1}^{N} (\tau - 1) \hat{u} - \sum_{n=b+1}^{N} \tau \hat{l} + \underbrace{\sum_{n=b+1}^{N} o_n}_{=c_2} \Bigg), \tag{G.25}$$

$$= \frac{1}{\tau} \Big((N-b) \big((\tau - 1) \hat{u} - \tau \hat{l} \big) + c_2 \Big). \tag{G.26}$$

The partially defined cases can then be combined to the overall IS function with

$$\mathrm{IS}_{\alpha} = \mathrm{IS}_{\alpha}^{(1)} + \mathrm{IS}_{\alpha}^{(2)} + \mathrm{IS}_{\alpha}^{(3)}, \tag{G.27}$$

$$= \frac{1}{\tau} \Big(a(1-\tau) \hat{l} + a\tau \hat{u} + c_1 \Big) + (b-a)(\hat{u} - \hat{l}) + \frac{1}{\tau} \Big((N-b) \big((\tau - 1) \hat{u} - \tau \hat{l} \big) + c_2 \Big). \tag{G.28}$$

The values of $\hat{l}$ and $\hat{u}$ that minimize the overall IS can then be found via partial differentiation of each variable.

The **optimum value** of $\hat{l}$ can be found via

$$\frac{\partial \mathrm{IS}_\alpha}{\partial \hat{l}} = \frac{1}{\tau} a(1-\tau) - (b-a) + \frac{1}{\tau}(N-b)(-\tau), \tag{G.29}$$

$$= \frac{1}{\tau} a - \cancel{a} - \cancel{b} + \cancel{a} + \cancel{b} - N, \tag{G.30}$$

$$= \frac{1}{\tau} a - N. \tag{G.31}$$

The extremum (turning point) can then be found by equating the derivative to zero, i.e.,

$$\frac{\partial \mathrm{IS}_\alpha}{\partial \hat{l}} \stackrel{!}{=} 0, \tag{G.32}$$

which leads to

$$0 = \frac{1}{\tau} a - N, \tag{G.33}$$

and can be rearranged to

$$a = \tau \cdot N. \tag{G.34}$$

This means that the IS has the minimum score value (regarding $\hat{l}$) if the lower interval border $\hat{l}$ is chosen so that there are $\tau \cdot N$ observations for which $o_n < \hat{l}$ is true.

The **optimum value** of $\hat{u}$ can be found with

$$\frac{\partial \mathrm{IS}_\alpha}{\partial \hat{u}} = \frac{1}{\tau} a\tau + b - a + \frac{1}{\tau}(N-b)(\tau - 1), \tag{G.35}$$

$$= \cancel{a} + \cancel{b} - \cancel{a} + N - \cancel{b} - \frac{1}{\tau}(N-b), \tag{G.36}$$

$$= N - \frac{1}{\tau}(N-b). \tag{G.37}$$

Again, the minimum value can be found by equating the derivate to zero, i.e.,

$$\frac{\partial \mathrm{IS}_\alpha}{\partial \hat{u}} \stackrel{!}{=} 0, \tag{G.38}$$

which leads to

$$0 = N - \frac{1}{\tau}(N-b). \tag{G.39}$$

and can be rearranged to

$$b = (1-\tau) \cdot N. \tag{G.40}$$

This again means that the minimum IS value regarding $\hat{u}$ is achieved when $\hat{u}$ is located so that there are $(1-\tau)\cdot N$ observations for which $o_n < \hat{u}$ is true.

In summary, this signifies that the IS gives the guarantee that it reaches its minimum value if the locations of $\hat{l}$ and $\hat{u}$ are set according to Eqs. G.34 and G.40. It is interesting that the optimum locations of both $\hat{l}$ and $\hat{u}$ do not depend on each other. It can furthermore be shown that $(1-\alpha)$ actually is the nominal confidence (NC) that specifies the coverage probability of the interval or the *relative* number of samples within the interval. The number of samples within the interval are specified with $b-a$, the *relative* number of samples can be expressed using $\text{NC} = \frac{b-a}{N}$ with

$$\text{NC} = \frac{(1-\tau)\cdot N - \tau \cdot N}{N}, \tag{G.41}$$

$$= 1 - 2\cdot\tau. \tag{G.42}$$

Given the relationship of Eq. G.3 with $\tau = \frac{\alpha}{2}$, the coverage probability indeed is

$$\text{NC} = 1 - 2\cdot\frac{\alpha}{2}, \tag{G.43}$$

$$= 1 - \alpha. \tag{G.44}$$

Bibliography

[1] AG Energiebilanzen e.V. (AGEB). Bruttostromerzeugung in Deutschland ab 1990 nach Energieträgern. Technical report, `http://www.ag-energiebilanzen.de/#20151112_brd_stromerzeugung1990-2014`, Berlin, Germany, 2017. Last accessed 2018-01-03.

[2] S. Alessandrini, S. Sperati, and P. Pinson. A comparison between the ECMWF and COSMO Ensemble Prediction Systems applied to short-term wind power forecasting on real data. *Applied Energy*, 107:271–280, 2013.

[3] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, 2016.

[4] R. Avnimelech and N. Intrator. Boosted Mixture of Experts: An Ensemble Learning Scheme. *Neural Computation*, 11(2):483–497, 1999.

[5] P. Bacher, H. Madsen, and H. A. Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83(10):1772–1783, 2009.

[6] A. B. Ballish and V. K. Kumar. Systematic differences in aircraft and radiosonde temperatures. *Bulletin of the American Meteorological Society*, 89(11):1689–1707, 2008.

[7] T. Barbounis and J. Theocharis. Locally recurrent neural networks for wind speed prediction using spatial correlation. *Information Sciences*, 177(24):5775–5797, 2007.

[8] R. Benedetti. Scoring Rules for Forecast Verification. *Monthly Weather Review*, 138(1):203–211, 2010.

[9] Y. Bengio. *Learning Deep Architectures for AI*, volume 2. now Publishers Inc., Delft, Netherlands, 2009.

[10] S. Bentzien and P. Friederichs. Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1924–1934, 2014.

[11] R. J. Bessa, A. Trindade, C. S. P. Silva, and V. Miranda. Probabilistic solar power forecasting in smart grids using distributed information. *International Journal of Electrical Power and Energy Systems*, 72:16–23, 2015.

[12] R. Binter. *Applied Probabilistic Models*. PhD thesis, London School of Economics, 2012.

[13] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer-Verlag, New York, USA, 2006.

[14] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[15] A. Botterud, Z. Zhou, J. Wang, J. Sumaili, H. Keko, J. Mendes, R. J. Bessa, and V. Miranda. Demand dispatch and probabilistic wind power forecasting in unit commitment and economic dispatch: A case study of Illinois. *IEEE Transactions on Sustainable Energy*, 4(1):250–261, 2013.

[16] Z. B. Bouallègue, P. Pinson, and P. Friederichs. Quantile forecast discrimination ability and value. *Quarterly Journal of the Royal Meteorological Society*, 141(693):3415–3424, 2015.

[17] M. Bouzerdoum, A. Mellit, and A. Massi Pavan. A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, 98(C):226–235, 2013.

[18] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting & Control*. John Wiley & Sons, New Jersey, USA, 1994.

[19] L. Breiman. Bagging Predictors. *Machine Learning*, 24(421):123–140, 1996.

[20] L. Breiman. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40(3):229–242, 2000.

[21] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[22] J. B. Bremnes. Probabilistic wind power forecasts using local quantile regression. *Wind Energy*, 7(1):47–54, 2004.

[23] J. Bröcker and L. A. Smith. From ensemble forecasts to predictive distributions. *Tellus A*, 60(4):663–678, 2008.

[24] J. Bröcker and L. A. Smith. Scoring Probabilistic Forecasts: The Importance of Being Proper. *Weather and Forecasting*, 22(2):382–388, 2007.

[25] G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: A survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

[26] G. Brown, J. Wyatt, and P. Tino. Managing Diversity in Regression Ensembles. *Journal of Machine Learning Research*, 6:1621–1650, 2005.

[27] S. S. Bucak, R. Jin, and A. K. Jain. Multiple kernel learning for visual object recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1354–1369, 2014.

[28] Bundesministerium für Wirtschaft und Energie (BMWi). Gesetz für den Ausbau erneuerbarer Energien (EEG 2017). `https://www.gesetze-im-internet.de/eeg_2014/BJNR106610014.html`. Last accessed 2018-01-03.

[29] Bundesministerium für Wirtschaft und Energie (BMWi). Zentrale Vorhaben Energiewende für die 18. Legislaturperiode. Technical report, `http://www.bmwi.de/Redaktion/DE/Downloads/0-9/10-punkte-energie-agenda.pdf`, Berlin, Germany, 2014. Last accessed 2017-12-20.

[30] Bundesverband WindEnergie. Fakten zur Windenergie: A bis Z. Technical Report 01, `https://www.wind-energie.de/sites/default/files/download/publication/z-fakten-zur-windenergie/bwe_abisz_3-2015_72dpi_final.pdf`, Berlin, Germany, 2014. Last accessed 2018-01-03.

[31] R. H. Byrd, J. C. Gilbert, and J. Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming, Series B*, 89(1):149–185, 2000.

[32] E. Cadenas, W. Rivera, R. Campos-Amezcua, and C. Heard. Wind speed prediction using a univariate ARIMA model and a multivariate NARX model. *Energies*, 9(2):1–15, 2016.

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

[34] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes. Ensemble selection from libraries of models. In *Proceedings of the International Conference on Machine Learning (ICML04)*, pages 137–144, 2004.

[35] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical methods for data analysis*. Wadsworth International Co. Inc., Pacific Grove, USA, 1983.

[36] A. Chandra and X. Yao. Ensemble learning using multi-objective evolutionary algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):417–445, 2006.

[37] W.-Y. Chang. A Literature Review of Wind Forecasting Methods. *Journal of Power and Energy Engineering*, 2(4):161–168, 2014.

[38] Z. Chen and Y. Yang. Assessing forecast accuracy measures. *Preprint Series*, pages 1–26, 2004.

[39] Y. Chu, B. Urquhart, S. M. Gohari, H. T. Pedro, J. Kleissl, and C. F. Coimbra. Short-term reforecasting of power output from a 48 MWe solar PV plant. *Solar Energy*, 112:68–77, 2015.

[40] N. Citroen, M. Ouassaid, and M. Maaroufi. Long term electricity demand forecasting using autoregressive integrated moving average model: Case study of Morocco. *Proceedings of International Conference on Electrical and Information Technologies, (ICEIT15)*, pages 59–64, 2015.

[41] B. T. Clough. *Unmanned Aerial Vehicles: Autonomous Control Challenges, a Researcher's Perspective*. Springer, Boston, USA, 1 edition, 2002.

[42] I. Colak, S. Sagiroglu, and M. Yesilbudak. Data mining and wind power prediction: A literature review. *Renewable Energy*, 46:241–247, 2012.

[43] A. Costa, A. Crespo, J. Navarro, G. Lizcano, H. Madsen, and E. Feitosa. A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, 12(6):1725–1744, 2008.

[44] J. D. Cox. *Storm watchers: the turbulent history of weather prediction from Franklin's kite to El Niño*. John Wiley & Sons, Hoboken, USA, 2002.

[45] P. P. Craig, A. Gadgil, and J. G. Koomey. What can history teach us? A Retrospective Examination of Long-Term Energy Forecasts for the United States. *Annual Review of Energy and the Environment*, 27(1):83–118, 2002.

[46] J. G. Da Silva Fonseca, T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, and K. Ogimoto. Photovoltaic power production forecasts with support vector regression: A study on the forecast horizon. In *Conference Record of the IEEE Photovoltaic Specialists Conference*, pages 2579–2583, Seattle, USA, 2011.

[47] I. Damousis, M. Alexiadis, J. Theocharis, and P. Dokopoulos. A Fuzzy Model for Wind Speed Prediction and Power Generation in Wind Parks Using Spatial Correlation. *IEEE Transactions on Energy Conversion*, 19(2):352–361, 2004.

[48] L. Delle Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight. Probabilistic Weather Prediction with an Analog Ensemble. *Monthly Weather Review*, 141(10):3498–3516, 2013.

[49] L. Deng and D. Yu. Deep Learning: Methods and Applications. *Foundations and Trends in Signal Processing*, 7(3-4):197–387, 2014.

[50] Deutscher Wetterdienst (DWD). ICON (Icosahedral Nonhydrostatic) Model. `https://www.dwd.de/EN/research/weatherforecasting/num_modelling/01_num_weather_prediction_modells/icon_description.html`. Last accessed 2018-01-03.

[51] Deutscher Wetterdienst (DWD). Nationaler Klimareport 2016, Klima - Gestern, heute und in der Zukunft. Technical report, `http://www.fortbildung-klimawandel.de/wp-content/uploads/2017/01/DWD_Nationaler_klimareport_2016.pdf`, Offenbach, Germany, 2016. Last accessed 2018-01-03.

[52] F. X. Diebold and J. A. Lopez. Forecast Evaluation and Combination. *Handbook of Statistics*, pages 241–268, 1996.

[53] F. X. Diebold and R. S. Mariano. Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 13(July):253–265, 1995.

[54] T. G. Dietterich. Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1857:1–15, 2000.

[55] J. Dobschinski. *Vorhersage der Prognosegüte verschieden großer Windpark-Portfolios*. Intelligent Embedded Systems, Kassel University Press, Kassel, Germany, 2016.

[56] A. Dolara, F. Grimaccia, S. Leva, M. Mussetta, and E. Ogliari. A physical hybrid artificial neural network for short term forecasting of PV plant power output. *Energies*, 8(2):1138–1153, 2015.

[57] Z. Dongmei, Z. Yuchen, and Z. Xu. Research on wind power forecasting in wind farms. In *Proceedings of the IEEE Power Engineering and Automation Conference (PEAM11)*, volume 1, pages 175–178, Wuhan, China, 2011.

[58] H. Du and L. A. Smith. Parameter estimation through ignorance. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 86(1):11–13, 2012.

[59] Q. Duan, N. K. Ajami, X. Gao, and S. Sorooshian. Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, 30(5):1371–1386, 2007.

[60] T. El-Fouly, E. El-Saadany, and M. Salama. One Day Ahead Prediction of Wind Speed and Direction. *IEEE Transactions on Energy Conversion*, 23(1):191–201, 2008.

[61] European Centre for Medium-Range Weather Forecasts (ECMWF). Medium-range forecasts. `http://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range/medium-range-forecast-charts`. Last accessed 2018-01-03.

[62] European Commission Joint Research Centre and PBL Netherlands Environmental Assessment Agency. Trends in global CO2 emissions: 2016 report. Technical report, `http://www.pbl.nl/sites/default/files/cms/publicaties/pbl-2016-trends-in-global-co2-emisions-2016-report-2315.pdf`, The Hague, Netherlands, 2016. Last accessed 2018-01-03.

[63] European Wind Energy Association (EWEA). Wind in power - 2016 European statistics. Technical report, `https://windeurope.org/wp-content/uploads/files/about-wind/statistics/WindEurope-Annual-Statistics-2016.pdf`, Brussels, Belgium, 2016. Last accessed 2018-01-03.

[64] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, and D. Amorim Fernández-Delgado. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

[65] D. Fisch, M. Jänicke, B. Sick, and C. Müller-Schloer. Quantitative emergence - A refined approach based on divergence measures. In *Proceedings of the 4th IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO2010)*, pages 94–103, Budapest, Hungary, 2010.

[66] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable Energy*, 37(1):1–8, 2012.

[67] Frankfurter Allgemeine Zeitung (FAZ). Stromnetz unter Druck: Rekordkosten wegen Noteingriffen. `http://www.faz.net/-iki-95elz`. Last accessed 2018-01-03.

[68] Fraunhofer IWES. Wind Energy Report Germany 2014. Technical report, `http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-354656-16.pdf`, Kassel, Germany, 2015. Last accessed 2018-01-03.

[69] P. Friederichs and A. Hense. Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. *Monthly Weather Review*, 135(6):2365–2378, 2007.

[70] J. Friedman. Greedy Function Approximation : A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[71] M. Friedman. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[72] H. C. Fu, Y. P. Lee, C. C. Chiang, and H. T. Pao. Divide-and-conquer learning and modular perceptron networks. *IEEE Transactions on Neural Networks*, 12(2):250–263, 2001.

[73] L. Fugon and G. Kariniotakis. Uncertainty Estimation of Wind Power Forecasts: Comparison of Probabilistic Modelling Approaches. In *European Wind Energy Conference and Exhibition (EWEC08)*, pages 1–10, Brussels, Belgium, 2008.

[74] C. Gallego-Castillo, A. Cuerva-Tejero, and O. Lopez-Garcia. A review on the recent history of wind power ramp forecasting. *Renewable and Sustainable Energy Reviews*, 52:1148–1157, 2015.

[75] A. Gandelli, F. Grimaccia, S. Leva, M. Mussetta, and E. Ogliari. Hybrid Model Analysis and Validation for PV energy production forecasting. *International Joint Conference on Neural Networks (IJCNN)*, 2014.

[76] A. Gensler. EuropeWindFarm Data Set Collection. `http://ies-research.de/Software`. Last accessed 2018-01-03.

[77] A. Gensler, T. Gruber, and B. Sick. Blazing Fast Time Series Segmentation Based on Update Techniques for Polynomial Approximations. In *Proceedings of the 13th IEEE International Conference on Data Mining Workshops (ICDMW13)*, pages 1002–1011, Dallas, USA, 2013.

[78] A. Gensler, T. Gruber, and B. Sick. Fast Feature Extraction For Time Series Analysis Using Least-Squares Approximations with Orthogonal Basis Functions. In *Proceedings of the International Workshop on Temporal Representation and Reasoning (TIME15)*, pages 29–37, Kassel, Germany, 2015.

[79] A. Gensler, J. Henze, B. Sick, and N. Raabe. Deep Learning for Solar Power Forecasting – An Approach Using Autoencoder and LSTM Neural Networks. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC16)*, pages 2858–2865, Budapest, Hungary, 2016.

[80] A. Gensler and B. Sick. Novel Criteria to Measure Performance of Time Series Segmentation Techniques. In *Proceedings of LWA/KDML: Workshop on Knowledge Discovery, Data Mining and Machine Learning*, volume 2, pages 29–37, Aachen, Germany, 2014.

[81] A. Gensler and B. Sick. Forecasting wind power – an ensemble technique with gradual coopetitive weighting based on weather situation. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN16)*, pages 4976–4984, Vancouver, Canada, jul 2016.

[82] A. Gensler and B. Sick. Performing Event Detection in Time Series with SwiftEvent: An Algorithm with Supervised Learning of Detection Criteria. *Springer Pattern Analysis and Applications (PAA)*, 1(1):1–20, 2017.

[83] A. Gensler and B. Sick. Probabilistic Wind Power Forecasting: A Multi-Scheme Ensemble Technique With Gradual Coopetitive Soft Gating. In *Proceedings of the 9th IEEE Symposium Series on Computational Intelligence (SSCI17)*, pages 1803–1812, Honolulu, USA, 2017.

[84] A. Gensler and B. Sick. A Multi-Scheme Ensemble Using Coopetitive Soft Gating With Application to Power Forecasting for Renewable Energy Generation. *ArXiv e-prints*, 1803.06344:1–22, 2018.

[85] A. Gensler, B. Sick, and V. Pankraz. An analog ensemble-based similarity search technique for solar power forecasting. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC16)*, pages 2850–2857, Budapest, Hungary, 2016.

[86] A. Gensler, B. Sick, and S. Vogt. A review of deterministic error scores and normalization techniques for power forecasting algorithms. In *Proceedings of the 8th IEEE Symposium Series on Computational Intelligence (SSCI16)*, pages 1–9, Athens, Greece, 2016.

[87] A. Gensler, B. Sick, and J. Willkomm. Temporal Data Analytics Based on Eigenmotif and Shape Space Representations of Time Series. In *Proceedings of the IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP14)*, pages 753–757, Xian, China, 2014.

[88] A. Gensler, S. Vogt, and B. Sick. Metaverification of Uncertainty Representations and Assessment Techniques for Power Forecasting Algorithms including Ensembles. *Renewable & Sustainable Energy Reviews*, 96:352–379, 2018.

[89] G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, and C. Draxl. The State-Of-The-Art in Short-Term Prediction of Wind Power: A Literature Overview, 2nd Edition. Technical report, `http://orbit.dtu.dk/files/7939719/Prediction_of_Wind_Power.pdf`, Lyngby, Denmark, 2011. Last accessed 2018-01-03.

[90] H. R. Glahn and D. A. Lowry. The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, 11(8):1203–1211, 1972.

[91] T. Gneiting. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

[92] T. Gneiting. Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27(2):197–207, 2011.

[93] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69(2):243–268, 2007.

[94] T. Gneiting and M. Katzfuss. Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151, 2014.

[95] T. Gneiting and A. E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[96] T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

[97] T. Gneiting and R. Ranjan. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011.

[98] T. Gneiting, L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235, 2008.

[99] M. Goldhammer. *Selbstlernende Algorithmen zur videobasierten Absichtserkennung von Fußgängern*, volume 9. Intelligent Embedded Systems Series, kassel university press GmbH, Kassel, Germany, 2017.

[100] M. Goldhammer, K. Doll, U. Brunsmann, A. Gensler, and B. Sick. Pedestrian's Trajectory Forecast in Public Traffic with Artificial Neural Networks. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR14)*, pages 4110–4115, Stockholm, Sweden, 2014.

[101] M. Gönen and E. Alpaydın. Multiple Kernel Learning Algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[102] D. Gopika and B. Azhagusundari. An Analysis on Ensemble Methods In Classification Tasks. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(7):7423–7427, 2014.

[103] C. W. J. Granger, H. White, and M. Kamstra. Interval forecasting. An analysis based upon ARCH-quantile estimators. *Journal of Econometrics*, 40(1):87–96, 1989.

[104] R. Hagedorn, F. J. Doblas-Reyes, and T. N. Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 57(3):219–233, 2005.

[105] T. M. Hamill. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, 129(3):550–560, 2001.

[106] A. U. Haque, M. H. Nehrir, and P. Mandal. A Hybrid Intelligent Model for Deterministic and Quantile Regression Approach for Probabilistic Wind Power Forecasting. *IEEE Transactions on Power Systems*, 29(4):1663–1672, jul 2014.

[107] L. O. Harvey, K. R. Hammond, C. M. Lusk, and E. F. Mross. The Application of Signal Detection Theory to Weather Forecasting Behavior. *Monthly Weather Review*, 120(5):863–883, 1992.

[108] H. Hersbach. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15(5):559–570, 2000.

[109] M. Hibon, S. Crone, and N. Kourentzes. Statistical Significance of Forecasting Methods. Technical report, `http://kourentzes.com/forecasting/wp-content/uploads/2014/04/ISF2012_Tests_Kourentzes.pdf`, Boston, USA, 2012. Last accessed 2018-01-03.

[110] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[111] S. Hochreiter and J. J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1–32, 1997.

[112] J. R. Holton and G. J. Hakim. *An Introduction to Dynamic Meteorology*. Elsevier Inc., Burlington, USA, 5 edition, 2012.

[113] H. Holttinen, P. Meibom, A. Orths, B. Lange, M. O'Malley, J. O. Tande, A. Estanqueiro, E. Gomez, L. Söder, G. Strbac, and Others. Impacts of large amounts of wind power on design and operation of power systems, results of IEA collaboration. *Wind Energy*, 14(2):179–192, 2011.

[114] T. Hong and S. Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

[115] T. Hong, P. Pinson, and S. Fan. Global Energy Forecasting Competition 2012. *International Journal of Forecasting*, 30(2):357–363, apr 2014.

[116] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, 2016.

[117] T. Hong, J. Wilson, and J. Xie. Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid*, 5(1):456–462, jan 2014.

[118] Q. Hu, R. Zhang, and Y. Zhou. Transfer learning for short-term wind speed prediction with deep neural networks. *Renewable Energy*, 85:83–95, 2016.

[119] G. B. Huang, D. H. Wang, and Y. Lan. Extreme learning machines: A survey. *International Journal of Machine Learning and Cybernetics*, 2(2):107–122, 2011.

[120] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, oct 2006.

[121] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra. Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, 39(6):535–576, 2013.

[122] H. Ishibuchi, T. Nakashima, and T. Morisawa. Voting in fuzzy rule-based systems for pattern classification problems. *Fuzzy Sets and Systems*, 103(2):223–238, 1999.

[123] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87, 1991.

[124] J. Jeon and J. W. Taylor. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107(497):66–79, 2012.

[125] W. Ji and K. C. Chee. Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. *Solar Energy*, 85(5):808–817, 2011.

[126] I. T. Jolliffe. The impenetrable hedge: A note on propriety, equitability and consistency. *Meteorological Applications*, 15(1):25–29, 2008.

[127] I. T. Jolliffe and D. B. Stephenson. *Forecast Verification - A Practitioner's Guide in Atmospheric Science*. Wiley-Blackwell, Chichester, 2nd edition, 2012.

[128] G. Juban, J., Fugon, L., Kariniotakis. Probabilistic Short-Term Wind Power Forecasting Based on Kernel Density Estimators. *European Wind Conference and Exhibition*, pages 1–11, 2007.

[129] J. Jung and R. P. Broadwater. Current status and future advances for wind speed and power forecasting. *Renewable and Sustainable Energy Reviews*, 31:762–777, 2014.

[130] C. Junk, L. Delle Monache, and S. Alessandrini. Analog-Based Ensemble Model Output Statistics. *Monthly Weather Review*, 143(7):2909–2917, 2015.

[131] R. Jursa and K. Rohrig. Short-term wind power forecasting using evolutionary algorithms for the automated specification of artificial intelligence models. *International Journal of Forecasting*, 24(4):694–709, 2008.

[132] M. Keuls. The use of the "studentized range" in connection with an analysis of variance. *Euphytica*, 1(2):112–122, 1952.

[133] M. Khodayar and M. Teshnehlab. Robust Deep Neural Network for Wind Speed Prediction. In *Proceedings of the 4th Iranian Joint Congress on Fuzzy and Intelligent Systems*, pages 1–5, Zahedan, Iran, 2015.

[134] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks*, 22(9):1341–1356, 2011.

[135] E. M. Kleinberg. Stochastic discrimination. *Annals of Mathematics and Artificial Intelligence*, 1(1):207–239, 1990.

[136] L. Knorr-Held and E. Rainer. Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction. *Biostatistics*, 2(1):109–129, 2001.

[137] R. Koenker and G. Bassett. Regression Quantiles. *Econometrica*, 46(1):33, jan 1978.

[138] R. Kohavi and D. H. Wolpert. Bias plus variance decomposition for zero-one loss functions. *Proceedings of the 13th International Conference on Machine Learning (ICML96)*, pages 275–283, 1996.

[139] A. J. Koning, P. H. Franses, M. Hibon, and H. O. Stekler. The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, 21(3):397–409, 2005.

[140] I. Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In *Proceeding of the European Conference on Machine Learning (ECML94)*, pages 171–182, Catania, Italy, 1994.

[141] I. Kononenko, E. Šimec, and M. Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7(1):39–55, 1997.

[142] L. Landberg. Short-term prediction of local wind conditions. *Boundary-Layer Meteorology*, 70(1):171–195, 1994.

[143] L. Landberg. Short-term prediction of the power production from wind farms. *Journal of Wind Engineering & Industrial Aerodynamics*, 80(1):207–220, 1999.

[144] L. Landberg, L. Myllerup, O. Rathmann, E. L. Petersen, B. H. Jørgensen, J. Badger, and N. G. Mortensen. Wind resource estimation - An overview. *Wind Energy*, 6(3):261–271, 2003.

[145] M. Lange and U. Focken. New developments in wind energy forecasting. In *Proceedings of the IEEE Power and Energy Society General Meeting: Conversion and Delivery of Electrical Energy in the 21st Century*, pages 1–8, 2008.

[146] M. Lei, L. Shiyan, J. Chuanwen, L. Hongling, and Z. Yan. A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, 13(4):915–920, 2009.

[147] B. Li, X. Zhang, and J. E. Smerdon. Comparison between spatio-temporal random processes and application to climate model data. *Environmetrics*, 27(5):267–279, 2016.

[148] Y. Li, Y. Su, and L. Shu. An ARMAX model for forecasting the power output of a grid connected photovoltaic system. *Renewable Energy*, 66(C):78–89, 2014.

[149] Y. Li, J. Chen, and L. Feng. Dealing with uncertainty: A survey of theories and practices. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2463–2482, 2013.

[150] Y. Lin, M. Yang, C. Wan, J. Wang, and Y. Song. A Multi-model Combination Approach for Probabilistic Wind Power Forecasting. *arXiv preprint*, pages 1–8, 2017.

[151] C. Loebecke, P. C. Van Fenema, and P. Powell. Co-opetition and knowledge transfer. *Database for Advances in Information Systems*, 30(2):14–25, 1999.

[152] M. Lydia, A. Selvakumar, S. Kumar, and G. Kumar. Advanced algorithms for wind turbine power curve modeling. *IEEE Transactions on Sustainable Energy*, 4(3):827–835, 2013.

[153] M. D. Hanouz (World Economic Forum). The Global Risks Report 2017, 12th Edition. Technical report, `http://www3.weforum.org/docs/GRR17_Report_web.pdf`, Cologny, Switzerland, 2017. Last accessed 2018-01-03.

[154] M. van der Hoeven (International Energy Agency). Energy and Climate Change: World Energy Outlook Special Report. Technical report, `https://www.iea.org/publications/freepublications/publication/WEO2015SpecialReportonEnergyandClimateChange.pdf`, Paris, France, 2015. Last accessed 2018-01-03.

[155] H. Madsen, P. Pinson, G. Kariniotakis, H. A. Nielsen, and T. Nielsen. Standardizing the Performance Evaluation of Short-Term Wind Power Prediction Models. *Wind Engineering*, 29(6):475–489, 2005.

[156] Maison de la Simulation. Energy oriented Centre of Excellence (EOCOE) Project - Meteorology for Energy. `http://www.eocoe.eu/workpackages/meteorology-energy`. Last accessed 2018-01-03.

[157] R. Mallipeddi and P. N. Suganthan. Unit commitment - a survey and comparison of conventional and nature inspired algorithms. *International Journal of Bio-Inspired Computation*, 6(2):71, 2014.

[158] D. R. Mandel and A. Barnes. Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, 111(30):10984–10989, jul 2014.

[159] G. Mascaro, E. R. Vivoni, and R. Deidda. Implications of Ensemble Quantitative Precipitation Forecast Errors on Distributed Streamflow Forecasting. *Journal of Hydrometeorology*, 11(1):69–86, 2010.

[160] S. J. Mason and D. B. Stephenson. How Do We Know Whether Seasonal Climate Forecasts are Any Good? *Seasonal Climate: Forecasting and Managing Risk*, 82:265–296, 2008.

[161] J. E. Matheson and R. L. Winkler. Scoring Rules for Continuous Probability Distributions. *Management Science*, 22(10):1087–1096, 1976.

[162] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa. Ensemble approaches for regression. *ACM Computing Surveys*, 45(1):1–40, 2012.

[163] Met Office. Met Office Numerical Weather Prediction models. `http://www.metoffice.gov.uk/research/modelling-systems/unified-model/weather-forecasting`. Last accessed 2018-01-03.

[164] S. Mirasgedis, Y. Sarafidis, E. Georgopoulou, D. Lalas, M. Moschovits, F. Karagiannis, and D. Papakonstantinou. Models for mid-term electricity demand forecasting incorporating weather influences. *Energy*, 31(2-3):208–227, 2006.

[165] M. P. Mittermaier. Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quarterly Journal of the Royal Meteorological Society*, 133(October):1487–1500, 2007.

[166] C. Monteiro, T. Santos, L. A. Fernandez-Jimenez, I. J. Ramirez-Rosado, and M. S. Terreros-Olarte. Short-term power forecasting model for photovoltaic plants based on historical similarity. *Energies*, 6(5):2624–2643, 2013.

[167] T. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, 1996.

[168] A. H. Murphy. The Finley Affair: A Signal Event in the History of Forecast Verification. *Weather and Forecasting*, 11(March):3–20, 1996.

[169] A. H. Murphy and R. L. Winkler. Reliability of Subjective Probability Forecasts of Precipitation and Temperature. *Applied Statistics*, 26(1):41, 1977.

[170] NASA Goddard Institute for Space Studies. GISS Surface Temperature Analysis (GISTEMP). `https://data.giss.nasa.gov/gistemp/`. Last accessed 2018-01-03.

[171] National Aeronautics and Space Administration (NASA). NASA, NOAA Analyses Reveal Record-Shattering Global Warm Temperatures in 2015. `https://goo.gl/4NNrPn`. Last accessed 2018-01-03.

[172] National Oceanic and Atmospheric Administration (NOAA). METAR Aviation Routine Weather Report. `https://www.aviationweather.gov/metar`. Last accessed 2018-01-03.

[173] M. Negnevitsky, P. Johnson, S. Member, S. Santoso, and S. Member. Short term wind power forecasting using hybrid intelligent systems. In *Proceedings of the IEEE Power Engineering Society General Meeting*, pages 1–4, Tampa, USA, 2007.

[174] J. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7(4):308–313, 1965.

[175] P. Nemenyi. Distribution-free multiple comparisons. In *Biometrics*, volume 18, page 263, 1962.

[176] H. A. Nielsen, H. Madsen, and T. S. Nielsen. Using quantile regression to extend an existing wind power forecasting system With probabilistic forecasts. *Wind Energy*, 9(1-2):95–108, 2006.

[177] H. A. Nielsen, H. Madsen, T. S. Nielsen, J. Badger, G. Giebel, L. Landberg, K. Sattler, and H. Feddersen. Wind Power Ensemble Forecasting. In *Proceedings of the Global Wind Power Conference*, pages 28–31, 2004.

[178] W. S. Noble. Support vector machine applications in computational biology. *Kernel Methods in Computational Biology*, pages 71–92, 2004.

[179] A. O'Hagan. Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. *Environmental Modelling and Software*, 36:35–48, 2012.

[180] I. Orlanski. A Rational Subdivision of Scales for Atmospheric Processes. *American Meteorological Society*, 56:527–530, 1975.

[181] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, Massachusetts, 2012.

[182] N. Padhy. Unit Commitment—A Bibliographical Survey. *IEEE Transactions on Power Systems*, 19(2):1196–1205, 2004.

[183] T. N. Palmer and D. L. T. Anderson. The Prospects for Seasonal Forecasting - a Review Paper. *Quarterly Journal of the Royal Meteorological Society*, 120(518):755–793, 1994.

[184] H. T. C. Pedro and C. F. M. Coimbra. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, 86(7):2017–2028, 2012.

[185] M. Pierro, F. Bucci, M. De Felice, E. Maggioni, D. Moser, A. Perotto, F. Spada, and C. Cornaro. Multi-Model Ensemble for day ahead prediction of photovoltaic power generation. *Solar Energy*, 134:132–146, 2016.

[186] P. Pinson and R. Girard. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96:12–20, 2012.

[187] P. Pinson, J. Juban, and G. N. Kariniotakis. On the quality and value of probabilistic forecasts of wind generation. In *Proceedings of the 9th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS06)*, Stockholm, Sweden, 2006.

[188] P. Pinson and G. Kariniotakis. On-line assessment of prediction risk for wind power production forecasts. *Wind Energy*, 7(2):119–132, 2004.

[189] P. Pinson, H. Nielsen, H. Madsen, and G. Kariniotakis. Skill forecasting from ensemble predictions of wind power. *Applied Energy*, 86(7-8):1326–1334, 2009.

[190] P. Pinson. *Estimation of the Uncertainty in Wind Power Forecasting*. PhD thesis, Ecole des Mines de Paris, 2006.

[191] P. Pinson and G. Kariniotakis. Conditional prediction intervals of wind power generation. *IEEE Transactions on Power Systems*, 25(4):1845–1856, 2010.

[192] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind Energy*, 12(1):51–62, 2009.

[193] P. Pinson, H. A. Nielsen, J. K. Müller, H. Madsen, and G. N. Kariniotakis. Non-parametric probabilistic forecasts of wind power: Required properties and evaluation. *Wind Energy*, 10(6):497–516, 2007.

[194] V. Pisetta, P. E. Jouve, and D. A. Zighed. Learning with ensembles of randomized trees: New insights. *Lecture Notes in Computer Science, subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics (LNAI)*, 6323(PART 3):67–82, 2010.

[195] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[196] Z. Qin, W. Li, and X. Xiong. Estimating wind speed probability distribution using kernel density method. *Electric Power Systems Research*, 81(12):2139–2146, 2011.

[197] Qu Xiaoyun, Kang Xiaoning, Zhang Chao, Jiang Shuai, and Ma Xiuda. Short-term prediction of wind power based on deep Long Short-Term Memory. In *Proceedings of the IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pages 1148–1152, Xian, China, 2016.

[198] H. Quan, D. Srinivasan, and A. Khosravi. Short-term load and wind power forecasting using neural network-based prediction intervals. *IEEE Transactions on Neural Networks and Learning Systems*, 25(2):303–315, 2014.

[199] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.

[200] B. Rajagopalan, U. Lall, and S. E. Zebiak. Categorical Climate Forecasts through Regularization and Optimal Combination of Multiple GCM Ensembles. *Monthly Weather Review*, 130(7):1792–1811, 2002.

[201] I. J. Ramirez-Rosado, L. A. Fernandez-Jimenez, C. Monteiro, J. Sousa, and R. Bessa. Comparison of two new short-term wind-power forecasting systems. *Renewable Energy*, 34(7):1848–1854, 2009.

[202] P. Ramsami and V. Oree. A hybrid method for forecasting the energy output of photovoltaic systems. *Energy Conversion and Management*, 95:406–413, 2015.

[203] M. Rana, I. Koprinska, and V. G. Agelidis. 2D-interval forecasts for solar power production. *Solar Energy*, 122(September):191–203, 2015.

[204] D. J. Rapp. Probabilistische Wettervorhersage. *Promet*, 37(3/4):1–109, 2011.

[205] Y. Ren, P. N. Suganthan, and N. Srikanth. A Comparative Study of Empirical Mode Decomposition-Based Short-Term Wind Speed Forecasting Methods. *IEEE Transactions on Sustainable Energy*, 6(1):236–244, 2015.

[206] Y. Ren, P. Suganthan, and N. Srikanth. Ensemble methods for wind and solar power forecasting - A state-of-the-art review. *Renewable and Sustainable Energy Reviews*, 50:82–91, 2015.

[207] Y. Ren, L. Zhang, and P. N. Suganthan. Ensemble Classification and Regression— Recent Developments, Applications and Future Directions. *IEEE Computational Intelligence Magazine*, 11(1):41–53, 2016.

[208] D. S. Richardson. Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126:649–667, 2000.

[209] L. Rokach. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics and Data Analysis*, 53(12):4046–4072, 2009.

[210] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[211] N. Rooney, D. Patterson, S. Anand, and A. Tsymbal. Dynamic integration of regression models. *Proceedings of the International Workshop on Multiple Classifier Systems*, 3181(1):164–173, 2004.

[212] M. S. Roulston and L. a. Smith. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, 130(6):1653–1660, 2002.

[213] S. Samarasinghe. *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. CRC Press, Boca Raton, USA, 1st edition, 2016.

[214] C. Sammut and G. I. Webb, editors. *Bias-Variance-Covariance Decomposition, Encyclopedia of Machine Learning*, page 111. Springer US, Boston, 2010.

[215] R. E. Schapire. The Strength of Weak Learnability. *Machine Learning*, 5(2):197–227, 1990.

[216] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26(5):1651–1686, 1998.

[217] R. Schefzik, T. L. Thorarinsdottir, and T. Gneiting. Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling. *Statistical Science*, 28(4):616–640, 2013.

[218] C. Schölzel and A. Hense. Probabilistic assessment of regional climate change in Southwest Germany by ensemble dressing. *Climate Dynamics*, 36(9-10):2003–2014, may 2011.

[219] J. M. Sloughter, T. Gneiting, and A. E. Raftery. Probabilistic Wind Speed Forecasting using Ensembles and Bayesian Model Averaging. *Journal of the American Statistical Association*, 105(489):25–35, 2010.

[220] O. Söder. kNN Classifiers: Filter-based feature weighting. `http://www.fon.hum.uva.nl/praat/manual/kNN_classifiers_1_1_1_1__Filter-based_feature_weighting.html`. Last accessed 2018-01-03.

[221] S. S. Soman, H. Zareipour, O. Malik, and P. Mandal. A review of wind power and wind speed forecasting methods with different time horizons. In *North American Power Symposium*, pages 1–8, Arlington, USA, 2010.

[222] T. F. Stocker and D. Qin. Climate Change 2013: The Physical Science Basis. Technical report, Cambridge University Press, New York, USA, 2013.

[223] Y. Tao, H. Chen, and C. Qiu. Wind Power Prediction and Pattern Feature Based on Deep Learning Method. In *Proceedings of the IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC14)*, pages 1–4, Kowloon, Hong Kong, 2014.

[224] A. Tascikaraoglu and M. Uzunoglu. A review of combined approaches for prediction of short-term wind speed and power. *Renewable and Sustainable Energy Reviews*, 34:243–254, 2014.

[225] L. J. Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.

[226] J. W. Taylor, P. E. McSharry, and R. Buizza. Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*, 24(3):775–782, 2009.

[227] J. Taylor and R. Buizza. Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems*, 17(3):626–632, aug 2002.

[228] T. L. Thorarinsdottir and M. S. Johnson. Probabilistic Wind Gust Forecasting Using Nonhomogeneous Gaussian Regression. *Monthly Weather Review*, 140:889–897, 2012.

[229] J. Tödter and B. Ahrens. Generalization of the Ignorance Score: Continuous Ranked Version and Its Decomposition. *Monthly Weather Review*, 140(6):2005–2017, 2012.

[230] N. Troldborg and J. Sørensen. A simple atmospheric boundary layer model applied to large eddy simulations of wind turbine wakes. *Wind Energy*, 17(April):657–669, 2014.

[231] N. Ueda and R. Nakano. Generalization error of ensemble estimators. In *Proceedings of the International Conference on Neural Networks (ICNN96)*, pages 90–95, Washington, USA, 1996.

[232] Unisys Weather. SYNOP Surface Synoptic Observations. `http://weather.unisys.com/wxp/Appendices/Formats/SYNOP.html`. Last accessed 2018-01-03.

[233] United Nations Framework Convention on Climate Change (UNFCCC), ICLEI World Secretariat. The Paris Climate Package : A Basic Guide for Local and Subnational Governments. Technical report, `http://e-lib.iclei.org/wp-content/uploads/2016/05/COP21-Report-web.pdf`, Bonn, Germany, 2016. Last accessed 2018-01-03.

[234] G. K. Vallis. *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press, Cambridge, UK, 2006.

[235] G. J. van Oldenborgh, F. J. Doblas-Reyes, B. Wouters, and W. Hazeleger. Decadal prediction skill in a multi-model ensemble. *Climate Dynamics*, 38(7-8):1263–1280, 2012.

[236] A. G. R. Vaz, B. Elsinga, W. G. J. H. M. van Sark, and M. C. Brito. An artificial neural network to assess the impact of neighbouring photovoltaic systems in power forecasting in Utrecht, the Netherlands. *Renewable Energy*, 85:631–641, 2016.

[237] C. Voyant, G. Notton, S. Kalogirou, M. L. Nivet, C. Paoli, F. Motte, and A. Fouilloy. Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105:569–582, 2017.

[238] M. Wainberg, B. Alipanahi, and B. J. Frey. Are Random Forests Truly the Best Classifiers? *Journal of Machine Learning Research*, 17(110):1–5, 2016.

[239] C. Wan, Z. Xu, and P. Pinson. Direct interval forecasting of wind power. *IEEE Transactions on Power Systems*, 28(4):4877–4878, 2013.

[240] C. Wan, Z. Xu, P. Pinson, Z. Dong, and K. Wong. Optimal prediction intervals of wind power generation. *IEEE Transactions on Power Systems*, 29(3):1166–1174, 2013.

[241] X. Wang, P. Guo, and X. Huang. A review of wind power forecasting models. *Energy Procedia*, 12:770–778, 2011.

[242] G. I. Webb and Z. Zheng. Multistrategy Ensemble Learning : Reducing Error by Combining Ensemble Learning Techniques. *IEEE Transactions on Knowledge and Data Engineering*, 16(8):980–991, 2004.

[243] A. P. Weigel, M. A. Liniger, and C. Appenzeller. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, 134(630):241–260, jan 2008.

[244] A. P. Weigel, M. A. Liniger, and C. Appenzeller. Generalization of the Discrete Brier and Ranked Probability Skill Scores for Weighted Multimodel Ensemble Forecasts. *Monthly Weather Review*, 135(7):2778–2785, jul 2007.

[245] S. V. Weijs and N. van de Giesen. Accounting for Observational Uncertainty in Forecast Verification: An Information-Theoretical View on Forecasts, Observations, and Truth. *Monthly Weather Review*, 139(7):2156–2162, 2011.

[246] S. V. Weijs, R. van Nooijen, and N. van de Giesen. Kullback-Leibler Divergence as a Forecast Skill Score with Classic Reliability-Resolution-Uncertainty Decomposition. *Monthly Weather Review*, 138(9):3387–3399, 2010.

[247] R. L. Welch, S. M. Ruffing, and G. K. Venayagamoorthy. Comparison of Feedforward and Feedback Neural Network Architectures for Short Term Wind Speed Prediction. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN09)*, pages 3335–3340, Atlanta, USA, 2009.

[248] D. S. Wilks. Resampling hypothesis tests for autocorrelated fields. *Journal of Climate*, 10(1):65–82, 1997.

[249] D. S. Wilks. Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128(586):2821–2836, 2002.

[250] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, Elsevier Inc., Amsterdam, Netherlands, 3rd edition, 2011.

[251] World Meteorological Organization (WMO) Communications and Public Affairs Office. Guidelines on Ensemble Prediction Systems and Forecasting. Technical Report WMO-No. 1091, http://www.wmo.int/pages/prog/www/Documents/1091_en.pdf, Geneva, Switzerland, 2012. Last accessed 2018-01-03.

[252] Y.-K. Wu, C.-R. Chen, and H. Abdul Rahman. A Novel Hybrid Model for Short-Term Forecasting in PV Power Generation. *International Journal of Photoenergy*, 2014:1–9, 2014.

[253] J. Yan, Y. Liu, S. Han, Y. Wang, and S. Feng. Reviews on uncertainty analysis of wind power forecasting. *Renewable and Sustainable Energy Reviews*, 52:1322–1330, 2015.

[254] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya. A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010.

[255] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD02)*, pages 694–699, Edmonton, Canada, 2002.

[256] J. Zhang and L. Castillo. Multivariate and Multimodal Wind Distribution Model based on Kernel Density Estimation. In *Proceedings of the 5th ASME International Conference on Energy Sustainability & 9th Fuel Cell Science, Engineering and Technology Conference*, pages 2125–2135, Washington, USA, 2011.

[257] M. Zhang and Z. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819 – 1837, 2014.

[258] W. Zhang and W. Wang. Wind speed forecasting via ensemble Kalman Filter. In *IEEE International Conference on Advanced Computer Control (ICACC)*, volume 2, pages 73–77, 2010.

[259] Y. Zhang, J. Wang, and X. Wang. Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, 32:255–270, 2014.

[260] X. Zhao, S. Wang, and T. Li. Review of evaluation criteria and main methods of wind power forecasting. *Energy Procedia*, 12:761–769, 2011.

[261] Y. Zhao, J. Gao, and X. Yang. A survey of neural network ensembles. In *Proceedings of the IEEE International Conference on Neural Networks and Brain (ICNNB05)*, pages 438–442, Beijing, China, 2005.

[262] C. Ziehmann. Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 52(3):280–299, 2000.

[263] E. Zorita and H. Von Storch. The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of Climate*, 12(8):2474–2489, 1999.

ISBN 978-3-7376-0636-3