Dutch electricity spot price forecasting

Two study cases using structured expert judgement

A.D.S. Bachasingh

TU Delft

Dutch electricity spot price forecasting

Two study cases using structured expert judgement

by

A.D.S. Bachasingh

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday March 17, 2022 at 9:00 AM.

Student number:	4221559	
Faculty:	EEMCS	
Master Programme:	Applied Mathematics	
Specialization:	Stochastics	
Date:	March 9, 2022	
Thesis committee:	Dr. ir. G.F. Nane,	TU Delft, Supervisor
	Dr. D. Kurowicka,	TU Delft, Chair
	Dr. P. Chen,	TU Delft
	M. Koenen	Greenchoice B.V.
	J. Duivenvoorden	Greenchoice B.V.

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Abstract

ELECTRICITY is a quite unique commodity. Due to the economically non-storable nature of the commodity that electricity is, the constant balance between consumption and production, weather effects, such as temperature, wind speed, solar intensity etc, and the intensity of everyday and business activities, e.g. holidays, weekends, on- and off-peak hours etc., the price dynamics of this commodity are quite unique as they are not observed in any other market. This extreme price volatility has forced market participants to hedge volumes as well as price risks. Naturally, electricity price forecast models are of great interest to portfolio managers.

Current data-driven models are combined with expertise of traders due to the considerable uncertainty of these models. We aim to subject trader expertise to a transparent methodology using structured expert judgement. During this study, we conduct two elicitations whose variables of interest concern average day-ahead spot prices for 2025, 2030 and 2035. We distinguish between baseload and peakload price. The first elicitation uses assessments regarding current Dutch day-ahead electricity spot prices to forecast forecast the variables of interest. We forecast the average day-ahead electricity spot price for 2025, 2030 and 2035. The second elicitation uses assessments regarding data on past, current and future developments in the Dutch electricity markets to forecast forecast the variables of interest. The second elicitation results in a decision maker with a higher price forecast than the decision maker of the first elicitation. Uncertainty regarding the forecast is higher for 2025 and 2030 during the second study. Uncertainty is equally high regarding the peakload price forecast for 2035 in both studies, however the range shifts.

Keywords: Electricity day-ahead spot prices, Baseload, Peakload, Cooke's Classical Model, Structured Expert Judgement, Experts, Electricity Price Forecast, EPEXSPOT, Dutch electricity market.

Preface

Dear reader,

THIS project has been so many things. It has been challenging, energizing, rewarding, educational, social, and so much more. Definitely a project worth working on. Unsurprisingly, I am very happy to present this report and hope you enjoy reading it as much as I enjoyed working on it.

There are so many people to thank. I would like to thank Maurice for inspiring me and giving me the space to create my own project. I would like to thank Jurgen for the brainstorm and feedback sessions. And I would like to thank my colleagues for their support and enthusiasm.

Words can not express how grateful I am for my supervisor Tina. She has been a source of inspiration since the Bachelor and has taught me so much. A simple thank you does not feel adequate. This project would not be what it is without you.

I would like to thank my friends and family for being supportive, patient and understanding throughout the Master. Especially, my dad, Monisha and Krishna. Kris for persuading me into the Master and supporting me through it. You saw my potential when I felt like I had none. And lets not forget the proofreading. My dad for always wishing for the moon for me. And Mo for the reassurance and space. You were always there, allowing me to vent. This is as much my accomplishment as yours.

Happy reading!

A.D.S. Bachasingh
The Hague, March 2022

Contents

1 Introduction

ELECTRICITY has become an essential commodity in our society. It provides light and warmth and plays a huge role in industry. Additionally, electricity is a quite unique commodity. Due to the economically non-storable nature of the commodity that is electricity, the constant balance between consumption and production, weather effects, e.g. temperature, wind speed, solar intensity etc, and the intensity of everyday and business activities, e.g. holidays, weekends, on- and off-peak hours etc., the price dynamics of this commodity are quite unique as they are not observed in any other market. This extreme price volatility has forced market participants to hedge volumes as well as price risks. Naturally, electricity price forecast models are of great interest to portfolio managers. This study proposes a new forecast method regarding electricity prices. That is, we use structured expert judgement, specifically Cooke's Classical Model, to forecast Dutch electricity spot prices.

There is much to say about Dutch electricity spot prices, structured expert judgement and this study. Therefore, this introduction goes through a few sections. The liberalization of the Dutch, but also the European, electricity market plays an prominent role in the volatility of Dutch electricity prices and the structure of the current Dutch electricity market. Therefore, we briefly discuss the liberalization of the Dutch electricity market. Additionally, we include background information about electricity trade on the supply level, e.g. the electricity spot market, as this study analyses and forecasts Dutch electricity spot prices. This section is followed by a brief introduction concerning Structured Expert Judgement. We discuss the Structured Expert Judgement method in depth in the following chapter, *Methodology*. The introduction is therefore kept rather short. Finally, we discuss the motivation behind this study. Moreover, we briefly discuss current studies regarding Dutch electricity spot price predictions and the need for these forecast models. We provide an overview of the study and explain why we have chosen this form. We end with an overview of this report.

1.1. The Dutch Electricity Market

THE gradual deregulation process of the European electricity market, starting in the early 1990s, resulted in electricity trade under market rules using spot and derivatives contracts. The first step towards the liberalization of the Dutch electricity market was taken in 1989 by the implementation of the so called *Elektriciteitswet 1989*. This law segregated the production and the distribution of electricity. The goal was to increase efficiency and trigger product development within companies as a result of competition. In 1998 the *Elektriciteitswet 1998* replaced the *Elektriciteitswet 1989* and thereby abolished the monopoly on electricity, with the exception of the grid operator. Following this law, liberalization of the Dutch electricity market took place in a few steps. During the first phase in 1998, the Dutch electricity market opened up for large consumers. That is, large consumers such as companies were free to choose their own electricity supplier. Starting 2001, the market opened up for all consumers regarding Green Energy. A year later, the market opened up for smaller businesses. Finally, on 1 July 2004, the market opened up for all consumers in the Netherlands.

The liberalization of the Dutch electricity market has resulted in the current Dutch electricity system which consists of four sectors, i.e. generation, transmission, distribution and supply. Electricity in the Netherlands is mainly generated by five large-scale companies, i.e. Essent, Vattenfall, Eneco, E.ON, Delta and Electabel. Furthermore, the Netherlands holds a large number of small-scale decentralised generators, i.e. co-generation plants in industry and horticulture, waste processing plants and sustainable energy, compared

to other European countries. The energy mix mainly consists of natural gas and coal, though it is the ambition of the Dutch government to increase the share of sustainable energy in the energy mix. The current government aims to increase the share of sustainable energy to 32% by 2030.

The transmission networks in the Netherlands are managed and owned by the Dutch transmission system operator Tennet [42]. Tennet is owned by the state. Tennet manages the transmission networks, monitors the electricity supply, balances supply and demand and resolves large-scale disruptions in electricity transmission. In addition, they are required to develop the electricity market and promote the integrated Central Western European market. Unsurprisingly, they installed multiple cross-border interconnection points in joint ventures with various European transmission system operators, e.g. the Nor-Nerd cable, the BritNed cable and the COBRAcable.

The distribution networks in the Netherlands are owned by the municipalities. However, the networks are operated by an independent network manager, appointed by the municipalities. This is due to the unbundeling of the Dutch electricity market to ensure operational and financial independence. Resulting in more than 27 regional network operators, e.g. Enexis Netbeheer B.V, Liander N.V., Stendin Netbeheer B.v. and Westland Infra Netbeheer B.V.. The distribution networks are the link between transmission and consumers. Simply put, the regional network operators manage the distribution networks.

The Dutch supply market counts over 45 energy supply companies, each offering different contracts and conditions. Unlike the previous three sectors, consumers are free to choose their own energy supplier since the liberalisation of the Dutch electricity market. Energy suppliers buy energy for their consumers and sell them to their consumers.

To buy energy, suppliers turn to the Dutch wholesale market. This can be divided into the bilateral market, the power exchange and the balancing market. EPEX SPOT [3] is an electricity exchange platform for European energy suppliers. Suppliers connect to this platform by becoming members. This enables them to submit orders concerning the purchase and/or sale of electricity. These order then reflect the supply and demand resulting in market price calculated by the Power Exchange.

The EPEX SPOT operates a Day-Ahead market and an Intraday market. The Day-Ahead market is based on a blind auction which is held once a day, every day. This auction trades electricity volumes of all hours of the following day. Market participants have to log in their orders for the next day before 12:00pm. The Power Exchange creates a supply curve based on the sell-orders and a demand curve based on the buy-orders for each hour of the following day. The intersection of both curves defines the market clearing price.

The Intraday market concerns continuous trade. The trade is executed as soon as a buy-order and a sell-order match. This can be done up to 5 minutes before delivery. Members use the Intraday market for last minute adjustments.

Lastly, energy suppliers trade in forward and futures contracts regarding electricity. Forward and futures contract are traded on exchanges world wide. With these contracts suppliers aim to hedge themselves against market risk. That is, futures and forward contracts are used to purchase base or sell loads of electricity volumes. Suppliers purchase or sell volumes with a delivery date in the future. If the spot market prices are expected to increase, the purchase of the futures contract manages the risk. However, it remains a difficult decision when to purchase these futures and forwards contracts.

1.2. Methodology

EXPERT judgement can be considered a rather simple concept. That is, we solicit an experts advice. Looking closely, we can conclude that any risk or decision analysis relies on expert judgement. It is used to select appropriate models, analytic methods, interpreting the output of an analysis and validate data. Experts evaluate whether data concerning current events are relevant to the prediction of future risk or opportunity. They even evaluate whether to implement risk management strategies or make a decision based on the output of an analysis. In this study, expert judgement, more specifically structured expert judgement, takes on another role.

It has never been easier to collect data. We share locations, consumer behaviour, interests, medical results, credit, debit, the list goes on. Nevertheless, we still come across situations where we lack data to assess all potential future events, risks or opportunities. The current COVID-19 pandemic is a strong example of this. At the beginning of the pandemic we lacked data to make useful quantitative assessments of risk. Thus we turned to experts. A lack of data is not the only possible motivation behind the use of Structured Expert Judgement. Forecast models often depend on historical data to forecast future values. However, historical data is not always sufficiently informative. Wind Energy costs for example, "Learning curves have been ap-

plied extensively to explain past wind cost trends, but when used to predict costs, they assume that future trends will follow the past." [46]. When historical data is not sufficiently informative, especially under highly volatile settings, expert judgement comes in. It is then only logical that expert judgement is also recognized as a type of scientific data. Methods have been developed to treat structured expert judgement as scientific data. Additionally, the heuristics and biases affecting subjective probability judgement have been analysed and reported as well to enhance and facilitate elicitation [30].

Structured expert judgement attempts to subject the process of soliciting expert advice to a transparent, traceable and validated methodology. The goal is to treat expert judgments as scientific data in a formal decision process. The scientific method is the process by which experts come to agree. Broadly speaking, Structured Expert Judgement is a tool for analysing and predicting certain variables of interest by aggregating and evaluating seed variables provided by a (manually selected) pool of experts. Cooke [12] defines seed variables as variables related to the experts' area of expertise with true values available post hoc. It is possible to have seed variables with true values determined before or during the elicitation. In this case it is crucial that the true values are not available to the experts until after the elicitation. The questions of interest drive the structured expert judgement elicitation. The values of the questions of interest are unknown. We are interested in these values. We use the Classical Model (CM), also known as Cooke's method [22], as a differential weighting scheme.

Cooke's method has been applied to studies in various sectors, e.g. Nuclear applications, Health, Banking and Aerospace [12][13]. It uses elicitation to asks experts to quantify uncertainty with regard to seed variables and the questions of interest. The experts are asked to give their point estimate and, in this case study, a 90% confidence interval for a specific uncertain quantity. The experts are treated as statistical hypotheses. They are combined in such a way that the statistical accuracy and informativeness of the decision maker is maximised. Finally, the decision maker is used to obtain the estimates and uncertainty intervals for the desired predictions.

1.3. The Project

OVER the last three decades, starting with the liberalization of the energy market, research in electricity price modeling and forecasting has propelled, resulting in various forecast models. We can broadly divide electricity price forecasting literature into six areas [29]: fundamental models, econometric models, reduced-form models, statistical models, game theory models and machine learning methods.

According to [29], statistical models and machine learning methods out perform the other methods. Moreover, [29] argues that machine learning methods such as deep neural networks, hybrid long-short term memory deep neural networks, hybrid gated recurrent unit deep neural networks and convolutional neural networks in turn out perform statistical models. The study uses the European power exchange Belgium, i.e. the Belgian electricity spot prices. According to [29], statistical models are usually linear forecasters. However, the data, hourly electricity prices, exhibit nonlinear behaviour. Resulting in possible poor performance.

A more direct approach is presented by [47]. They directly model the supply and demand curve, resulting in an estimate of the market clearing price. They use the German-Austrian day-ahead electricity market of EPEX for their model. Another approach are hybrid models. [15] proposes a hybrid model which consist of a cost-production optimisation model and a neural network model. The models are linked by using the output of the cost-production optimisation model as an input for the neural network.

These are just a few studies on electricity price modeling. Academic literature on electricity price modeling approaches is a very large collection. Reviewing these completely falls outside the scope of this paper. Nevertheless, the scientific uncertainty of these models to this date is substantial. Accurate electricity price forecasts result in the mitigation of negative effects of price uncertainty, a stabilized grid and thus economic profits. Decreasing the uncertainty in current models and developing new models with a lower uncertainty is an ongoing process. In the meantime, current models are combined with expertise of forecasters. Nonetheless, expert performance in quantifying uncertainty or validation have not been considered beforehand. We aim to subject this process to transparent methodology using structured expert judgement. Though we know that expert knowledge plays an important role in electricity price forecasting, there has not been an attempt yet, to the best of our knowledge, to apply the structured expert judgement method in this sector.

The questions of interest are constant throughout this project. That is, both elicitation studies have the exact same Questions of Interest. The experts are asked to answer the following questions,

- What will be the average electricity baseload spot price on the EPEX spot market in 2025?
- What will be the average electricity peakload spot price on the EPEX spot market in 2025?

- What will be the average electricity baseload spot price on the EPEX spot market in 2030?
- What will be the average electricity peakload spot price on the EPEX spot market in 2030?
- What will be the average electricity baseload spot price on the EPEX spot market in 2035?
- What will be the average electricity peakload spot price on the EPEX spot market in 2035?

Note that we are interested in forecasts rather far in the future, i.e. 2025, 2030 and 2035. These prices are of importance to electricity supplier due to their purchase strategy. Portfolio managers buy and sell electricity on the spot market, but they can also buy volumes of the commodity years in advance on the exchange or over the counter. Information about future electricity prices can influence the purchase strategy. Portfolio managers and electricity traders can choose to hedge their position to mitigate the negative effects of price uncertainty. In case of an increasing market price, suppliers can choose to purchase large volumes against low prices to sell in the future against high prices. Naturally, an accurate forecast is of great importance.

To obtain these forecasts we apply elicitation. Our first elicitation session concerned the forecast of the average day baseload Dutch electricity spot price and the average day peakload Dutch electricity spot price for the upcoming day. Here, peakloads refer to 08:00 - 20:00 and baseloads refer to a 24-hour time period. We aim to measure how well our experts can forecast electricity spot prices on short term. We use their performance to forecast future electricity spot prices. That is, the average electricity spot prices in 2025, 2030 and 2035.

However, future electricity spot prices do not necessarily depend on current electricity spot prices. Past, current and future developments in the Dutch electricity markets play a prominent role in future electricity market prices. Therefore, we are interested in measuring knowledge on past, current and projected developments concerning the Dutch electricity markets and possible corresponding data. We therefore preformed a second elicitation where the seed questions concern the Dutch electricity market. More details about these questions will be provided in Chapter 4.

The expert pool during the first and the second elicitation are not the same. The aim of the first elicitation is to measure how well experts can forecast electricity spot prices. Therefore, we created an expert pool consisting out of traders, electricity analysts, portfolio managers but also electricity business consultants and sales employees. Every expert is linked to the commodity, but not necessarily directly to the spot market. The aim of the second elicitation is to measure how knowledge concerning the Dutch electricity market. Past, current and future developments in the Dutch electricity markets play a prominent role in future electricity market prices. Therefore, we are interested in measuring knowledge on past, current and projected developments concerning the Dutch electricity markets and possible corresponding data. Hence, we restricted ourselves to electricity traders, electricity analysts, portfolio managers, Dutch electricity consultants and electricity price and load forecasters during the second elicitation.

We would like to note on the anonymity of the participating experts. During most Structured Expert Judgement studies the names and affiliations of the experts are mentioned in the reports. However, these are not linked with individual assessments. In this report, names and affiliations are left out. The experts are completely unanimous and only the facilitator of the elicitation knows their names and affiliations.

Before diving into the next chapter of this case study, we provide an overview of this thesis. In Chapter 2, we will discuss the scientific method that is Cooke's Classical Model, which is used to combine and assess expert uncertainty. We briefly discuss the elicitation procedure for Structured Expert Judgement. We end this chapter by discussing the software used during this case study, i.e. Excalibur. Chapters 3 and 4 have similar structures. We discuss the expert pool, elicitation procedure, seed variables, questions of interest and the data used to create the seed variables. Followed by the expert performance analysis and the decision maker performance. Based on the best performing decision maker, we present the results. That is, the answers to the Questions of Interest based on the decision maker. Finally, we discuss the analysis and results and formulate a conclusion. However, Chapter 3 concerns the first elicitation of this case study and Chapter 4 concerns the second elicitation of this case study. Chapter 5 is a brief comparison of both studies. Finally, Chapter 6 summarises our research finding. Additionally, we discuss the results and we provide recommendation for future research.

2

Methodology

We have already introduced Structured Expert Judgement and the Classical Model in the previous chapter. This chapter gives a more detailed description of the scientific method. Moreover, we discuss the Classical Model in depth. We also briefly discuss the elicitation process. The specific elicitation procedures for this study are discussed in chapters 3 and 4. This chapter only discusses the aim of the elicitation procedure and the overall structure of the elicitation procedure. Finally, we discuss the Excalibur software package. We restrict ourselves to the functions used during this case study. Additional information regarding the software package and its functions can be found in [7] and [31].

2.1. Structured Expert Judgement

Expert judgement can be defined as soliciting expert advice. We can broadly divide the context of expert consultation into three groups [22], i.e. the expert problem, the group decision problem and the textbook problem. All of these problems consult a group of experts. The difference lies in the responsibility and accountability. The responsibility and accountability lies with the problem owner in the case of an expert problem, giving the experts more freedom with their assessments. Whereas the responsibility and accountability lies with the expert group themselves in the case of a group decision problem. The textbook problem has no predefined risk or decision problem. Experts are asked for their judgement for others to use in currently undefined circumstances.

In all the contexts above, Cooke [12] distinguishes between three goals regarding the aspiration of structured expert judgement, i.e. census, political consensus, and rational consensus. Census refers to the process in which the totality of views are surveyed across an expert pool and expressed as a distribution. The process where experts are assigned weights according to their representations is considered political consensus. A group decision process is considered rational consensus. For this process, the method is agreed on by the group without knowing the results. The method generates a representation of uncertainty for the purposes for which the panel was convened.

Cooke formulates that structured expert judgement, like any other scientific method, should follow the general principles of rational consensus, i.e. reproducibility or accountability, empirical control, neutrality and fairness. That is, peer review has access to all data and all processing tools such that the results can be reproduced by competent reviewers (reproducibility or accountability), quantitative expert assessments are subjected to empirical quality controls (empirical control), the method used to combine/evaluate expert opinion should not bias results and encourage experts to state their true opinions (neutrality) and experts are not judged prior to processing the results of their assessments (Fairness). These principles are satisfied in the Classical Model which aims at rational consensus, which we will discuss in the next section.

We can define Structured expert judgement as an attempt to subject expert solicitation to transparent methodological rules. Our goal is to treat expert judgement as scientific data in a formal decision process. Structured expert Judgement is sought for multiple reasons. That is, a decision process is impacted by substantial scientific uncertainty, to build rational consensus or because we lack empirical data, i.e. data is unavailable, incomplete, uninformative or conflicting.

2.2. The Classical Model

THE Classical Model (CM) was developed with the main aim to quantify uncertainty. It measures the performance of experts as uncertainty assessors. CM uses elicitation to asks experts to quantify uncertainty with regard to calibration questions, or seed variables, and the questions of interest. Through the elicitation, experts provide the requested fixed and finite number of percentiles of a distribution. This results in a minimally informative non-parametric distribution.

We use seed variables with three objectives, i.e. as a measure for expert performance, to create performance-based weighted combination of experts' distributions and to analyse and validate the resulting combination. Cooke [12] defines seed variables as variables related to the experts' area of expertise with true values available post hoc. It is possible to have seed variables with true values determined before or during the elicitation. In this case it is crucial that the true values are not available to the experts until after the elicitation. The questions of interest drive the structured expert judgement elicitation. The values of the questions of interest are unknown. We are interested in these values.

Once we have obtained the assessed percentiles from various experts, we can use these assessments to construct expert distributions. However, we need to determine the support of the distribution before we can specify expert's distribution. Therefore, we need to define *range*.

Definition 2.2.0.1 (Range) *Assume N experts provide their assessments. Denote expert's e_i assessments for a given question as q_5^i, q_{50}^i and q_{95}^i for the 5th, 50th and 95th percentiles, respectively, and $i = 1,2,\ldots,N$. The range [L,U] is given by*

$$\begin{aligned} L &= \min_{1\le i\le N}\{q_5^i, \textit{realization}\}, \\ U &= \max_{1\le i\le N}\{q_{95}^i, \textit{realization}\}, \end{aligned}$$

for a given seed variable.

Naturally, for the questions of interest L and U become $\min_{1\le i\le N}\{q_5^i\}$, and $\max_{1\le i\le N}\{q_{95}^i\}$, respectively, for $i = 1,\ldots,N$.

Using the defined range, we determine the support of the experts' distribution by the so-called intrinsic range.

Definition 2.2.0.2 (Intrinsic range) *The intrinsic range is given by*

$$[L^*, U^*] = [L - k\cdot(U-L), U + k\cdot(U-L)],$$

where k denotes an overshoot and is chosen by the analyst.

Usually, $k = 10\%$. Note that for certain types of questions the intrinsic range can be specified a priori by the analyst, e.g. when eliciting percentages as the natural intrinsic range here is [0,100].

We construct each of the expert's distribution by interpolating between expert's percentiles, assigning the mass uniformly within the inter-percentile ranges. Using the uniform background measure, we define the distribution of expert e_i as,

Definition 2.2.0.3 (Expert distribution) *Assume a uniform background measure. The distribution of expert e_i is then given by,*

$$F_i(x) = \begin{cases} 0, & \textit{for } x < L^* \\ \frac{0.05}{q_5^i - L^*}\cdot(x - L^*), & \textit{for } L^* \le x < q_5^i \\ \frac{0.45}{q_{50}^i - q_5^i}\cdot(x - q_5^i) + 0.05, & \textit{for } q_5^i \le x < q_{50}^i \\ \frac{0.45}{q_{95}^i - q_{50}^i}\cdot(x - q_{50}^i) + 0.5, & \textit{for } q_{50}^i \le x < q_{95}^i \\ \frac{0.05}{U^* - q_{95}^i}\cdot(x - q_{95}^i) + 0.95, & \textit{for } q_{95}^i \le x < U^* \\ 1, & \textit{for } x \ge U^*. \end{cases}$$

Note that the cumulative distribution F_i is continuous.

We now have the subjective probability distributions per expert per question. Our aim is to measure expert performance and go as far as to create a decision maker, i.e. combine the distributions of expert opinions via linear pooling. Before we can create the decision maker, we need to calculate the weights. We start by calculating the calibration score and the information score.

The calibration score of an expert reflects the statistical accuracy of that expert. To obtain the calibration score we first form the sample distribution of each experts inter-percentile intervals.

Definition 2.2.0.4 (Expert empirical distribution) *Assume there are N experts, $e_1, e_2, \ldots, e_N$ and M seed variables. Denote expert's e_i assessments on question j as $q_5^{i,j}$, $q_{50}^{i,j}$ and $q_{95}^{i,j}$ for the 5^{th}, 50^{th} and 95^{th} percentiles, respectively. Denote the realisations of the seed questions as $x_1, \ldots, x_M$. We define the empirical distribution for expert i as*

$$s(e_i) = (s_1(e_i), s_2(e_i), s_3(e_i), s_4(e_i)),$$

where

$$\begin{aligned} s_1(e_i) &= \frac{\sum_{k=1}^{M} \mathbb{1}_{\{x_k \le q_5^{i,k}\}}}{M}, \\ s_2(e_i) &= \frac{\sum_{k=1}^{M} \mathbb{1}_{\{q_5^{i,k} < x_k \le q_{50}^{i,k}\}}}{M}, \\ s_3(e_i) &= \frac{\sum_{k=1}^{M} \mathbb{1}_{\{q_{50}^{i,k} < x_k \le q_{95}^{i,k}\}}}{M}, \\ s_4(e_i) &= \frac{\sum_{k=1}^{M} \mathbb{1}_{\{q_{95}^{i,k} < x_k\}}}{M}, \end{aligned}$$

with

$$\mathbb{1}_{\{x \le a\}} = \begin{cases} 1, & \text{when } x < a \\ 0, & \text{otherwise} \end{cases}$$

the indicator function.

In other words, we count how many of the M realisations fall within each inter-percentile interval to form the sample distribution of expert e_i's inter-percentile intervals. Our goal is to measure how extreme expert's empirical distribution is from the expected distribution of the realizations in the inter-quantile intervals p. Therefore, we implement the Kullback-Leibler divergence measure, or KL divergence measure. The difference between two probability distributions is measured by the non-symmetric KL divergence measure. We use the KL divergence measure to measure the difference between inter-percentile probability vector and the empirical distribution obtained from the raw frequencies. Note that the inter-percentile probability vector is defined as the vector $p = (0.05, 0.45, 0.45, 0.05)$ in this study, corresponding with our predefined quantiles. Suffice to say, an alternative choice for quantiles results in a different probability vector, corresponding with the chosen quantiles.

Definition 2.2.0.5 (Kullback-Leibler divergence measure) *Assume $s(e_i)$ and p are two probability distributions of a discrete random variable. That is, both $s(e_i)$ and p sum up to 1, and $s_l(e_i) > 0$ and $p_l > 0$ for any $l \in \{1,2,3,4\}$. We define the KL divergence measure of $s(e_i)$ and p as*

$$I(s(e_i), p) = \sum_{l=1}^{4} s_l(e_i) \ln \frac{s_l(e_i)}{p_l},$$

The quantity $2MI(s(e_i), p)$ is asymptotically distributed as a chi-square random variable with 3 degrees of freedom, provided that the realisations are drawn independently from a distribution with percentiles given by the expert. Resulting in the calibration score of expert e_i defined as the statistical likelihood of the hypothesis

H_{e_i} : the inter-percentile interval containing the true value for each variable is drawn independently from probability vector p.

Formally,

Definition 2.2.0.6 (Calibration score) *The p-value of the hypothesis H_{e_i} is defined as the calibration score, or statistical accuracy,*

$$Cal(e_i) = \mathbb{P}\{2MI(s(e_i), p) > r | H_{ei}\},$$

where r is the value of $2MI(s(e_i), p)$ based on the observed values $x_1, \ldots, x_M$.

Note that a KL-divergence equal to zero, i.e. $I(s(e_i), p) = 0$, corresponds with the highest possible calibration score. That is, the expert's sample distribution equals the inter-percentile probability vector for the seed variables. Hence, an increasing KL-divergence, i.e. s diverging from p, results in a decreasing calibration score.

We are now able to measure statistical accuracy. But we would also like to measure the informativeness of the experts. An expert could ensure their statistical accuracy by providing very wide assessments. This is not a problem if this reflects the uncertainty of the expert. However, ideally we would like to have an expert pool where the experts are statistically accurate as well as highly informative.

We used the background measure to create the subjective probability distributions of the experts. The information score measures how informative expert's distributions are with respect to the background measure.

Definition 2.2.0.7 (Information score of expert e_i) *Assume $[L^*, U^*]$ is the intrinsic range. The information score of expert e_i for question j is determined by*

$$I_j(e_i) = \sum_{k=1}^{4} f_k \ln \frac{f_k}{r_k},$$

where, with respect to expert's distribution $F(\cdot)$,

$$\begin{aligned} f_1 &= F(q_5^i) - F(L^*) = 0.05, \\ f_2 &= F(q_{50}^i) - F(q_5^i) = 0.45, \\ f_3 &= F(q_{95}^i) - F(q_{50}^i) = 0.45, \\ f_4 &= F(U^*) - F(q_{95}^i) = 0.05, \end{aligned}$$

and the uniform background measure

$$\begin{aligned} r_1 &= U(q_5^i) - U(L^*) = \frac{q_5^i - L^*}{U^* - L^*}, \text{ for } x \in [L^*, q_5^i], \\ r_2 &= U(q_{50}^i) - U(q_5^i) = \frac{q_{50}^i - q_5^i}{U^* - L^*}, \text{ for } x \in (q_5^i, q_{50}^i], \\ r_3 &= U(q_{95}^i) - U(q_{50}^i) = \frac{q_{95}^i - q_{50}^i}{U^* - L^*}, \text{ for } x \in (q_{50}^i, q_{95}^i], \\ r_4 &= U(U^*) - U(q_{95}^i) = \frac{U_* - q_{95}^i}{U^* - L^*}, \text{ for } x \in (q_{95}^i, U^*], \end{aligned}$$

with,

$$U(x) = \frac{x - L^*}{U^* - L^*}, \text{ for } L^* \le x \le U^*$$

Definition 2.2.0.8 (Total Information score of expert e_i) *The information score of an expert over all seed questions is defined as the average of information scores*

$$I(e_i) = \frac{1}{M} \sum_{j=1}^{M} I_j(e_i).$$

The information score is a strictly positive function. Unlike the calibration score, which can take on a value between 0 and 1, the information score can take arbitrarily large values. Note that, if the intrinsic range spans multiple orders of magnitude, we apply the log-uniform measure to construct the distributions. Naturally, the informativeness of the constructed distribution is then also measured with respect to the log-uniform background measure.

Finally, statistical accuracy is more important than informativeness. That is, a high calibration score is preferred over an high information score. Assessments that are highly informative but statistically inaccurate are not useful. Non-informative but statistically accurate assessments are useful, as they teach us how large the uncertainties may be.

The calibration score and information score can be used to compare experts' performance. However, we can go a step further and use the scores to construct weights and mathematically aggregate distributions to create a decision maker. To construct weights we combine the calibration score and the information score into a combined score.

Definition 2.2.0.9 (Combined score global) *The combined score for expert i is given by*

$$CS(e_i) = Cal(e_i) \cdot I(e_i) \cdot \mathbb{1}_\alpha(Cal(e_i)), \text{ for } i = 1,\dots,N \text{ and } \alpha \geq 0.$$

We use a cutoff level α to only multiply the desired calibration scores and their corresponding information scores. For example, if we set $\alpha = 0.05$, all experts with a calibration score lower than 0.05 will receive a combined score equal to zero.

The weight of expert i will be proportional to their score,

Definition 2.2.0.10 (Weight global) *The weight for expert i is given by*

$$w_i = \frac{CS(e_i)}{\sum_{k=1}^{N} CS(e_k)}, \text{ for } i = 1,\dots,N.$$

Naturally, experts with a combined score equal to zero will receive a weight equal to zero. This does not mean that they have no contribution with respect to the decision maker. All experts' assessments determine the support of all variables, thus all experts contribute to the decision maker. Moreover, a weight equal to zero usually means that the experts' knowledge overlaps with other experts. Hence, the expert contributes via the other experts.

Note that the information score is calculated per question, or item, per expert and then averaged over all the questions. Alternatively, we can also calculate the combined score per item per expert.

Definition 2.2.0.11 (Combined score item) *The combined score for expert i and seed variable j is given by*

$$CS_j(e_i) = Cal(e_i) \cdot I_j(e_i) \cdot \mathbb{1}_\alpha(Cal(e_i)), \text{ for } i = 1,\dots,N, j = 1,\dots,M, \text{ and } \alpha \geq 0.$$

Resulting in a weight per item per expert.

Definition 2.2.0.12 (Weight item) *The weight for expert i and question j is given by*

$$w_i^j = \frac{CS_j(e_i)}{\sum_{k=1}^{N} CS_j(e_k)}, \text{ for } i = 1,\dots,N \text{ and } j = 1,\dots,M.$$

We obtain a large weight matrix where each row corresponds with an item and each column corresponds with an expert. Note that the calibration scores are the same for different questions of the same expert.

With the obtained weights we can now create decision makers. We use the performance-based weights to combine experts' judgements via linear pooling. Formally,

Definition 2.2.0.13 *Decision Maker distribution We define the combined distribution of the experts, or the distribution of the Decision Maker, as*

$$F_{DM} = \sum_{i=1}^{N} w_i F_i,$$

where N defines the expert group size, F_i are the experts' distributions indexed over the N experts and the numbers w_i are weights adding to 1.

We defined global weights and item weights, therefore we have the global weight decision maker (GWDM) and the item weight decision maker (IWDM). Moreover, these decision makers vary depending on the cutoff value α. Resulting in numerous possibilities for decision makers.

Furthermore, we can distinguish between the decision maker and the optimized decision maker. The decision maker sets $\alpha = 0$, thus assigning a weight to all experts with a calibration and or information score higher than 0. The optimized decision maker calculates the highest combined score for the decision maker and adjusts the cutoff value accordingly. This optimized decision maker can be calculated for the IWDM and the GWDM.

Finally, we have the equal weight decision maker (EWDM). The EWDM does not depend on the combined score. Assume N experts provide their assessments. Each expert is assigned a weight equal to $\frac{1}{N}$. The EWDM is then constructed using these equal weights.

Note that the decision maker becomes a virtual expert. With the above defined distribution we can calculate the calibration score and information score of the decision maker. The obtained combined score can be used to recalculate the weight, now with the decision maker as a virtual expert.

Moreover, we can compare the decision maker using the calibration score, information score and combined score of the decision makers. Hence, we choose the best performing decision maker.

2.3. Elicitation

THE elicitation is a process on its own. It involves more than collecting expert assessments. One or two facilitators typically conduct the elicitation, where at least one facilitator is necessary who is well-versed in elicitation practices. The second facilitator is then usually the person with extensive knowledge regarding the study.

There are multiple ways to conduct the elicitation. However, all elicitation procedures have the same basic structure and the same goal. That is, the procedure goes through the following steps:

- Introduction
- Training
- Dry-Run
- Elicitation
- Feedback

The structured expert judgement method is introduced during the introduction. Moreover, the motivation and background behind the study is explained. Important remarks are discussed, such as the fact that names and affiliations will or will not be published and if so they will not be linked with individual assessments. Moreover, the facilitator explains the scoring measures and stresses that statistical accuracy is valued above informativeness. Examples are discussed in detail to enhance understanding, bringing us to the training section of the elicitation procedure.

During the training the experts go through a few calibration questions and provide their specified percentiles for each question. These questions are usually connected to the study, but cannot be used in the study as the facilitator has to provide the true value during the training. This helps the experts understand the method and helps them understand their subjective probability. During the training the facilitator can provide the expert with feedback.

The elicitation ideally asks the experts for more than a quantification of their uncertainty. That is, the elicitation also asks for the qualitative reasoning behind the experts' assessments. These explanations are usually added to the appendix of the study and help the problem owner to understand expert assessments and the outcome of the study.

After all the expert data is collected and the analysis was performed, the experts are provided with feedback. This includes the final decision maker and the corresponding assessments of the Question(s) of Interest.

The procedure can be conducted in multiple ways. That is, facilitators can go through the steps in person, per video call or conduct the elicitation procedure remotely. Unsurprisingly, each form has its own advantages and disadvantages. Take for example remote elicitation procedures. This can be conducted by sending the experts the introduction per e-mail. Followed by an online training via for example Kahoot. The elicitation takes place via google docs and finally feedback is again provided via e-mail. The flexibility of this form is a big advantage here. Experts can plan their own moments to go through the procedure. However, it is argued that it is not as effective as in-person training as experts feel less invested. This results in less qualitative information. On the other hand, in-person elicitations can result in low response rates due to scheduling

problems and inflexibility. Quigley et al. [18] discuss advantages and disadvantages of various forms. They also discuss how to conduct in-person, video, plenary, and one-on-one sessions.

2.4. Excalibur

THE EXCALIBUR Software package is an implementation of Cooke's Classical Model. Aspinall [7] touches upon the various functions implemented in EXCALIBUR. We discuss the functionalities used during this study.

EXCALIBUR allows us to create cases. One can only analyse one case at a time. When we open a new case in EXCALIBUR, a window pops up called *Realisation data*. We are asked to put in the seed variables and the questions of interest. We also need to put in the realisations of the seed variables. Excalibur recognizes questions of interest by the lack of the realisation data in the *Realisation data* file. For user simplicity, EXCALIBUR allows us to tag the questions with an ID. For example, we can denote the first seed variable as Q1. Additionally, EXCALIBUR allows us to write out the full question under *Full name*.

Once the questions and realisations are put in EXCALIBUR, we add the experts. This can be done unanimously, e.g. numbering your experts, or by name. We press *Experts* and are asked to define the quantiles of the elicitation. After defining the quantiles we are free to add the experts.

We add expert assessments by selecting an expert and pressing *Assessments*. A window pops up containing the seed variables and the questions of interests together with the realisations. We are now able to put in the assessments of the expert. We do this for all the experts.

Finally, we can calculate the scores of the experts. Once all the assessments are added to the case, we can press *Calculate*. A parameter window pops up. We can choose the weights, i.e. global, equal or item. We can ask for the optimized decision maker or set a cutoff value α. We can even name the decision maker. After choosing the desired combination of parameters, we click *RUN*. EXCALIBUR returns the experts scores for the selected parameters. We are also able to export expert scores, expert and decision maker quantiles and expert and decision maker distributions. To export these we simply click *File* followed by *Export as text* for the expert scores or *Export as space delimited* for the quantiles of distributions. Clicking on *Solution* gives us the output of the decision maker.

The Classical Model allows us to apply the log-uniform measure to construct the distributions if the intrinsic range spans multiple order of magnitude. EXCALIBUR gives us this option as well. We can set the scale of the items to UNI or LOG. As a rule of thumb, we use the logarithmic if the experts' assessments for a question spans over four orders of magnitude.

The Intrinsic Range is set to 0.10 by default in EXCALIBUR. We have not adjusted this default during this study. Furthermore, EXCALIBUR allows us to adjust the Calibration Power. However, this option is redundant to us as all our expert answer the same questions and answer all questions.

Finally, EXCALIBUR has the option for a discrepancy analysis and robustness test. Both these options have not been used during this study. Aspinall [7] provides a detailed overview of these options.

3

Measuring expert electricity spot price prediction performance

THIS part of the case study focuses on the forecast of future Dutch electricity day-ahead spot prices by measuring expert performance based on the prediction of current Dutch electricity day-ahead spot prices. Experts were asked to predict day-average Dutch electricity day-ahead spot prices between 1 March 2021 and 1 May 2021. This period is rather interesting because of the increase in solar production. The end of the winter and the beginning of the spring causes a change in the weather. This results in an increase in solar production. However, forecasters and traders are left with historic solar generation information. In the Netherlands, households and companies tend to set up their solar panels and corresponding installation during the end of the summer and in the fall. Resulting in an increase in solar production compared to previous years. Hence, traders, forecasters and forecast models have a hard time forecasting the influence of this increase on the prices. Moreover, this increase causes high volatility in the price data. Figure 3.1 nicely depicts the volatility of the commodity price in March 2021 and April 2021.

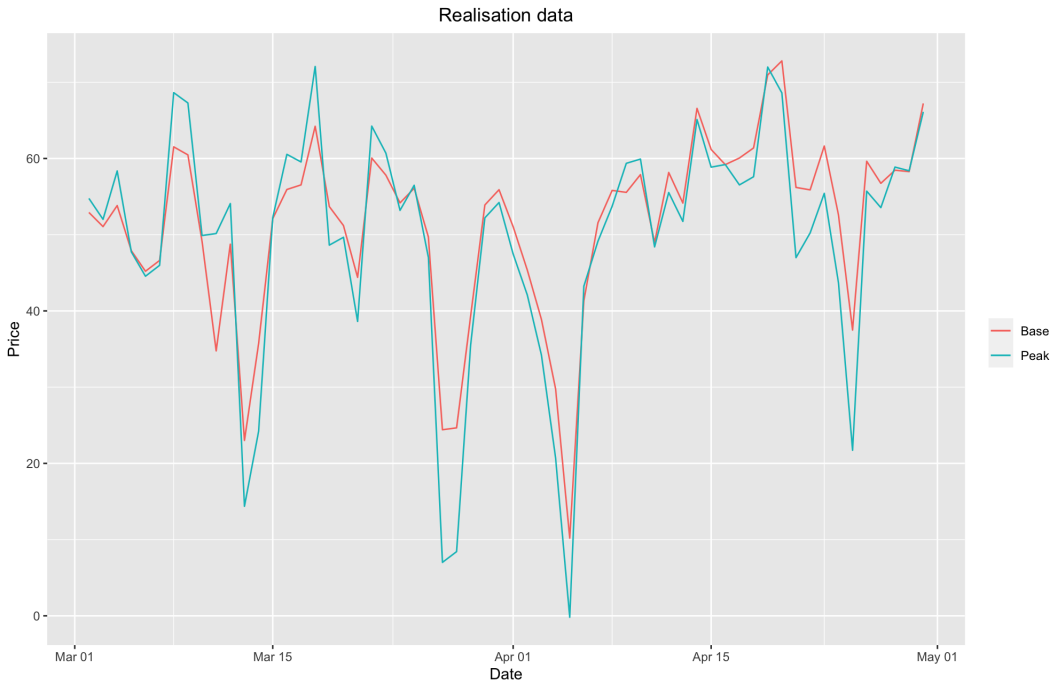

Figure 3.1: Average Dutch electricity day-ahead prices as published by EPEXSPOT between 2 March 2021 and 1 May 2021. Prices are given in euro's per MWh. The red line represents the day average Dutch electricity day-ahead baseload spot price on the EPEX spot market. The blue line represents the day average Dutch electricity day-ahead peakload spot price on the EPEX spot market.

Solar generation is not the only factor causing price volatility. Wind generation, holidays in April and influences from neighbouring countries play a role as well. All in all, we see that March and April are rather difficult months to forecast with the heavy price increases and decreases. This makes for a very interesting elicitation.

It is important to note that future spot prices do not necessarily depend on current spot prices. Therefore, our aim is not to forecast future spot prices based on current spot price prediction performance. Rather, the idea is that traders and forecasters who are able to forecast the electricity day-ahead spot prices accurately are aware of the important variables that play a prominent role in the market clearing price. The fact that the realisations are disclosed before the next assessment is provided gives the experts an opportunity to evaluate their previous assessment and adjust their next assessment. That is, experts are learning and adjusting accordingly as the elicitation proceeds.

This section discusses this first elicitation and the results in detail. We start by discussion our expert pool. Followed by the elicitation itself. We discuss the procedure, the Questions of Interest, the Calibration Questions and our data sources. Furthermore, we analyse the expert performance and the decision maker performance. Based on the best performing decision maker, we present the results. That is, the answers to the Questions of Interest based on the decision maker. Finally, we discuss the analysis and results and formulate a conclusion.

3.1. Experts

We defined a rather broad target group for the study. Naturally, we choose to invite forecasters, traders, portfolio managers and analysts originating from the Dutch electricity supplying companies. Additionally, we choose to include professionals linked to electricity trade within the companies, i.e. energy business consultants, sales employees and product owners. These professionals might not have experience in energy trading, but due to their daily tasks they are exposed to the energy (spot) prices. We assume them to provide assessments originating from an alternative point of view. For example, a trader would forecast electricity spot prices taking into account the weather forecast, neighbouring countries, generation versus demand, seasonality and so on. A sales employee might use the weather forecast, but additionally use their knowledge on pricing and try to find a pattern in historical data. Hence, we obtain assessments originating from various approaches. This in turn improves the model we are building.

Initially, 20 potential experts received an invitation to participate. Three experts did not respond to the invitation. After the first week of elicitation another expert dropped out. Furthermore, the assessments of three experts were deemed unfit at the end of the elicitation. These experts lacked 50% or more of the assessments, rendering the data rather useless. Therefore, they were not included in the study. This left us with an expert pool of 13 experts.

3.2. Elicitation

Elicitation always requires preparation. We had to decide on the expert pool, invite the pool and prepare them for the elicitation itself and of course conduct the elicitation. As mentioned before, the elicitation procedure goes through a hand-full of steps.

3.2.1. Elicitation procedure

Before we can actually collect expert assessments, we have to prepare the elicitation and the experts. Furthermore, we must decide on a few matters such as the elicitation platform. During this study we decided on an online and remote elicitation. Partly due to the COVID-19 measures in the Netherlands, but also to motivate the invited experts to participate. Experts had to provide two assessments on a daily basis for two months. Conducting the elicitation remotely seemed like the most beneficial procedure for all parties involved.

Experts received an invitation via Microsoft Teams to the introduction of the case study. During this meeting we discussed the motivation behind the study. Moreover, the structured expert judgement method was introduced including the scoring measures used in the study. Followed by a few plenary training questions where experts could test their understanding of the method. We asked the experts to provide assessments concerning the Dutch electricity spot prices of the previous week. These questions were very similar to the seed questions to create familiarity. Furthermore, we discussed that participation is anonymous and collaboration is not allowed. The experts were also given the chance to ask questions. Finally, the experts were told that they would all receive an elicitation form, Appendix A, containing all the information discussed the meeting. They would also receive an excel file for their assessments. Each expert received an individual excel

file, shared with the facilitator. Each day after the deadline, the facilitator would copy the assessments of each expert to ensure fairness.

As an extra motivation, we added a competing element to the elicitation. The best performing expert would receive a bar of chocolate.

One final remark. The elicitation period was rather lengthy. Experts would naturally take holidays and therefore would not be able to provide assessments. Moreover, busy schedules would result in missing assessments. Experts were given instructions regarding these situations. They were told to provide the assessments before going on vacation. It is important to note that this could increase uncertainty for the expert or make for less accurate assessments. Furthermore, the facilitator would check the assessments at least two hours before the deadline. Experts would receive a reminder via e-mail asking them to provide their assessments before the deadline to prevent missing assessments.

3.2.2. Questions of Interest

Through the elicitation we obtain experts assessments concerning seed variables and questions of interest. In this study, the Questions of Interest aim to predict future Dutch electricity spot prices on the EPEX spot market. We are interested in the average price on a year level. Particularly for 2025, 2030 and 2035. Resulting in the following Questions of Interest.

- What will be the average electricity baseload spot price on the EPEX spot market in 2025?
- What will be the average electricity peakload spot price on the EPEX spot market in 2025?
- What will be the average electricity baseload spot price on the EPEX spot market in 2030?
- What will be the average electricity peakload spot price on the EPEX spot market in 2030?
- What will be the average electricity baseload spot price on the EPEX spot market in 2035?
- What will be the average electricity peakload spot price on the EPEX spot market in 2035?

The prices are given un euros per MWh. We distinguish between the baseload prices and the peakload prices. That is, peakloads refer to 08:00 - 20:00. Baseloads refer to a 24-hour time period.

3.2.3. Calibration questions

We measure expert performance and quantify their uncertainty using seed variables. We use calibration questions to objectively evaluate the uncertainty assessments of the experts. There were 120 calibration questions in total.

Between 1 March and 1 May we asked the same two questions each day. The date changes with one day each day.For example, on the first of March the expert was presented two calibration questions:

- What will be the average electricity day-ahead baseload spot price on the EPEX spot market on 2 March 2021?
- What will be the average electricity day-ahead peakload spot price on the EPEX spot market on 2 March 2021?

On the second of March we changed 2 March to 3 March in the question etc. The prices were given in euros per MWh. Peakloads refer to 08:00 - 20:00. Baseloads refer to a 24-hour time period.

The assessments are given each day before 10:00. These assessments concern the prices of the next day. That is, the assessments provided on the first day of March before 10:00 concern the average electricity day-ahead baseload/peakload spot price on the EPEX spot market on 2 March 2021. The assessments given between 1-3-2021 10:00 and 2-3-2021 10:00 concern the average electricity day-ahead baseload/peakload spot price on the EPEX spot market on 3 March 2021.

3.2.4. Data

The EPEX SPOT [3] publishes electricity spot prices each day around 12:00. They distinguish between the baseload spot price and the peakload spot price. Moreover, they publish the prices for all European countries. We use the prices published for the Netherlands. They provide hourly prices and the average day price regarding the day-ahead market and the intraday market. This study focuses on the day-ahead market. We kept track of the average baseload day price and the average peakload day price regarding Dutch electricity during this study.

3.3. Expert Performance Analysis

THE elicitation has resulted in 13 × 126 assessments. That is, each expert has given us 60 assessments regarding the baseload electricity prices, 60 assessments regarding the peakload electricity prices and 6 assessments regarding the questions of interest. With these assessments and the corresponding realisations we can calculate the calibration score, information score and combined score for each expert. The information score can be calculated including assessments regarding the questions of interest or excluding the assessments regarding the questions of interest. We start by discussing these scores first.

3.3.1. General Performance

Table 3.1 presents us with the calibration score, information score all questions, Information score seed questions score and the combined score of each participating expert.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score
Expert 1	$1.14 \cdot 10^{-13}$	1.064	1.058	$1.20 \cdot 10^{-13}$
Expert 2	$3.46 \cdot 10^{-16}$	0.658	0.674	$2.33 \cdot 10^{-16}$
Expert 3	$4.96 \cdot 10^{-10}$	1.123	1.127	$5.59 \cdot 10^{-10}$
Expert 4	$3.05 \cdot 10^{-11}$	0.502	0.489	$1.49 \cdot 10^{-11}$
Expert 5	$5.75 \cdot 10^{-18}$	1.420	1.413	$8.12 \cdot 10^{-18}$
Expert 6	$1.22 \cdot 10^{-12}$	0.763	0.788	$9.62 \cdot 10^{-13}$
Expert 7	0	0.717	0.707	0
Expert 8	$4.48 \cdot 10^{-16}$	0.655	0.654	$2.93 \cdot 10^{-16}$
Expert 9	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$
Expert 10	$4.51 \cdot 10^{-09}$	0.590	0.591	$2.67 \cdot 10^{-09}$
Expert 11	$8.14 \cdot 10^{-02}$	0.490	0.466	$3.80 \cdot 10^{-02}$
Expert 12	0	1.212	1.201	0
Expert 13	$2.13 \cdot 10^{-12}$	0.919	0.901	$1.92 \cdot 10^{-12}$

Table 3.1: General Performance

We see that two experts, expert 7 and expert 12, will have no weight in decision maker. Their calibration score equals zero. Hence, their combined score will always equal zero and result in a weight equal to zero. This does not mean that the experts do not contribute. As mentioned before, all expert's assessments determine the support of all variables. Moreover, experts' knowledge most likely overlaps with other experts.

Expert	Total	Baseload	Peakload
Exp1	77	41	36
Exp2	72	37	35
Exp3	81	46	35
Exp4	80	36	44
Exp5	74	42	32
Exp6	77	43	34
Exp7	56	30	26
Exp8	75	40	35
Exp9	100	52	48
Exp10	84	41	43
Exp11	99	54	45
Exp12	56	29	27
Exp13	78	39	39

Table 3.2: Number of times each expert has captured the realisation within their 5th and 95th percentile. Each expert could capture 120 realisations in total, 60 related to the baseload and 60 related to the peakload.

Table 3.2 shows us that expert 7 and expert 12 have captured the least realisations in their provided intervals. The corresponding information scores do not necessarily stand out. That is, compared to the information scores of the other experts, the scores are neither high nor low. However, the experts might have similar information scores, they provided very different percentiles. In Appendix D figure C.7 we observe that expert 7 provides the same 5th percentile and 95th percentile for multiple days consecutively. Expert 7 does not adjust the percentiles when the prices start increasing. Resulting in missed realisations. We can clearly observe the constant underestimating from expert 7, resulting in the incredibly low calibration score. Expert 12 continuously updates the percentiles. Moreover, the provided intervals are rather narrow. Resulting in missed realisation. Figure C.12 suggests overconfidence.

The best performing experts are expert 9 and expert 11. Expert 9 has the highest calibration score, almost twice as high as the calibration score of expert 11. Expert 9 has also captured the most realisations, missing only 20 realisations of the 120. However, expert 11 has captured 99 out of the 120 realisations. Hence, a difference of one realisation captured. Clearly, capturing realisations does not result in a high calibration score. Expert 9 has provided assessments such that their expert empirical distribution is closer to the inter-percentile probability vector when compared to the empirical distribution based on expert 12's assessments. Expert 12's realizations which were captured were not distributed evenly within the second and third inter-quantile range. We turn to figure C.12 and observe quite a lot of overestimation for expert 12. Resulting in a much higher calibration score for expert 9.

Nevertheless, the difference between the combined score of expert 9 and expert 11 is not very big. The relatively low information score of expert 9 results in a combined score lower than the combined score of expert 11. Expert 11 has a relatively low information score, but in expert 11's case their information score increases their combined score.

Expert 1 has a very low calibration score, but a very high information score. From figure C.1 we can conclude that expert 1 was rather overconfident during the first weeks of the elicitation. This resulted in the low calibration score and high information score.

The calibration score and information score of expert 2 would suggest a rather poor performance. Figure C.2 displays a rather large uncertainty for expert 2, especially compared to the other experts. That is, expert 2 provided quite a few wide intervals. Moreover, we see that expert 2 managed to roughly model the volatility shape of the prices. Unfortunately, expert 2 is penalised by the expert empirical distribution construction.

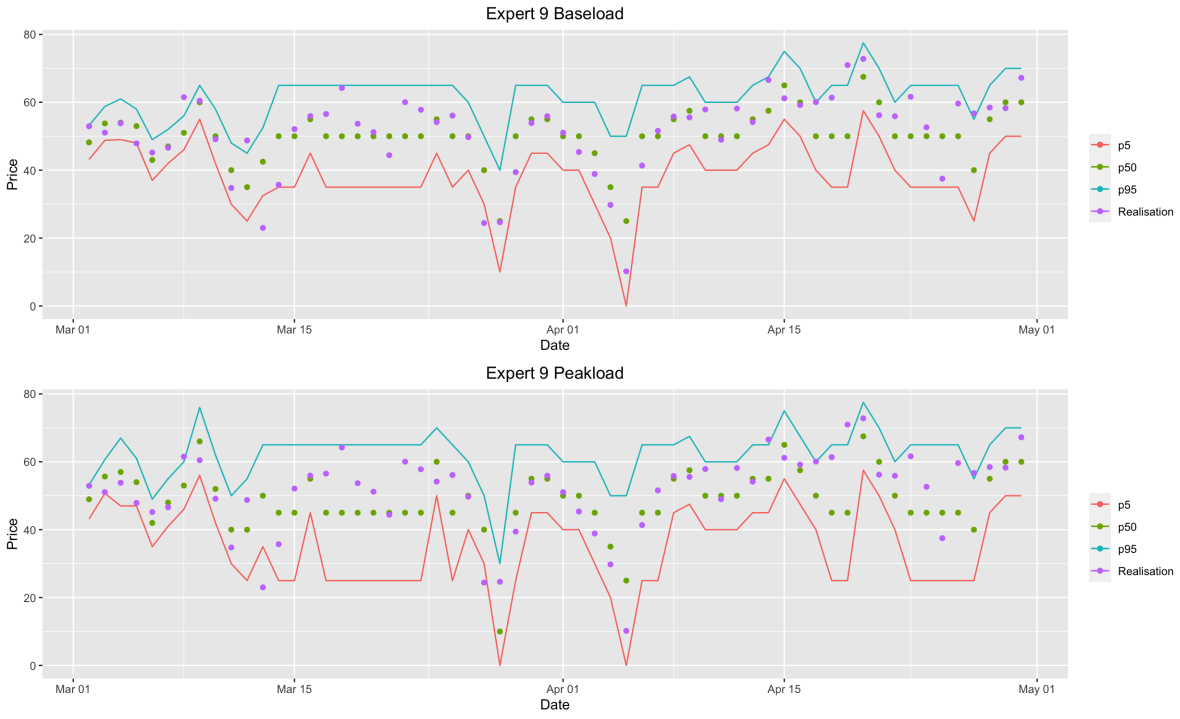

Figure 3.2: Assessments of expert 9. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asesements and the peakload assessments.

Expert 3 has a very high information score due to the narrow intervals provided by them. However, we see in figure C.3 that the experts low calibration score comes from the way the expert empirical distribution is constructed. Expert 4 provides rather wide consistent assessments. Resulting in the capture of 80 assessments. However, these method does not reward expert 4 with high scores.

Figure C.5 displays very narrow intervals. Expert 5 is therefore rewarded with the highest information score. Moreover, a large amount of realisations seem to lie very close to the p50 values. Quite a few uncaptured realisations lie very close the p95 or p5 values. However, this overconfidence results in a very low calibration score for expert 5.

Experts 6,8,10 and 13 have rather *jumpy* intervals. The intervals attempt to follow the volatility of the realisations. Around 60% of the realisations are captured. Resulting in a rather low calibration score compared to for example expert 9. Information scores are neither high nor low and do not imply overconfidence. The experts seem to display a level of uncertainty. This is not surprising considering the commodity we are trying to forecast.

The experts have given assessments for the Dutch day-ahead electricity peakload prices and for the Dutch day-ahead electricity baseload prices. Hence, we can distinguish between the two and look at the expert performances regarding each separately. Table 3.3 displays the expert performance regarding the forecast of the Dutch day-ahead electricity baseload prices. That is, expert's calibration score, information score all questions, information score seed questions and combined score. Table 3.4 presents us the expert performance regarding the forecast of the Dutch day-head electricity peakload prices.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score
Expert 1	$2.90 \cdot 10^{-05}$	0.998	0.980	$2.84 \cdot 10^{-05}$
Expert 2	$5.40 \cdot 10^{-08}$	0.720	0.757	$4.09 \cdot 10^{-08}$
Expert 3	$8.77 \cdot 10^{-03}$	1.136	1.144	$1.00 \cdot 10^{-02}$
Expert 4	$1.40 \cdot 10^{-11}$	0.460	0.430	$6.04 \cdot 10^{-12}$
Expert 5	$6.65 \cdot 10^{-06}$	1.386	1.369	$9.10 \cdot 10^{-06}$
Expert 6	$7.30 \cdot 10^{-04}$	0.701	0.745	$5.44 \cdot 10^{-04}$
Expert 7	$3.24 \cdot 10^{-16}$	0.668	0.643	$2.08 \cdot 10^{-16}$
Expert 8	$6.56 \cdot 10^{-07}$	0.620	0.615	$4.03 \cdot 10^{-07}$
Expert 9	$7.40 \cdot 10^{-01}$	0.644	0.678	$5.00 \cdot 10^{-01}$
Expert 10	$2.90 \cdot 10^{-05}$	0.636	0.642	$1.86 \cdot 10^{-05}$
Expert 11	$3.94 \cdot 10^{-01}$	0.468	0.419	$1.65 \cdot 10^{-01}$
Expert 12	$5.39 \cdot 10^{-16}$	1.205	1.182	$6.37 \cdot 10^{-16}$
Expert 13	$1.13 \cdot 10^{-06}$	0.910	0.872	$9.82 \cdot 10^{-07}$

Table 3.3: General Performance Baseload Prices

Clearly, most calibration scores improve in both tables compared to table 3.1. Unsurprisingly, the statement can be made regarding the combined scores of the experts. However, we can not make such a general statement about the information scores. These increase for some experts and decrease for others. The differences between the information scores are at most a few tenths. When we compare table 3.3 with table 3.4 we can make similar statements. Baseload price calibration scores mostly improve compared to peakload price calibration scores. Again, combined scores are higher as well when these two tables are compared. Finally, information scores increase for some experts and decrease for others when comparing table 3.3 with table 3.4.

We see that most experts have a harder time capturing peakload price realisations compared to baseload price realisations, table 3.2. Note that the both prices are very volatile, figure 3.1. However, the peaks and drops regarding the peakload prices are heavier compared to the baseload prices. If experts assume a certain constant discrepancy between the average baseload price and peakload price, they could fail to capture the peakload price realisation due to the assumption. That is, they could account for the volatility in the data and decrease or increase their assessments accordingly, but not account for the increase in discrepancy between the average baseload price and average peakload price.

We already stated that the difference in the information scores between table 3.3 and table 3.4 is rather small. By distinguishing between baseload prices and peakload prices experts do not becomes significantly

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score
Expert 1	$7.74 \cdot 10^{-09}$	1.140	1.136	$8.78 \cdot 10^{-09}$
Expert 2	$2.76 \cdot 10^{-09}$	0.568	0.591	$1.63 \cdot 10^{-09}$
Expert 3	$4.51 \cdot 10^{-09}$	1.105	1.110	$5.01 \cdot 10^{-09}$
Expert 4	$9.01 \cdot 10^{-04}$	0.567	0.547	$4.93 \cdot 10^{-04}$
Expert 5	$7.14 \cdot 10^{-13}$	1.466	1.456	$1.04 \cdot 10^{-12}$
Expert 6	$9.70 \cdot 10^{-10}$	0.779	0.830	$8.05 \cdot 10^{-10}$
Expert 7	$3.63 \cdot 10^{-18}$	0.785	0.772	$2.80 \cdot 10^{-18}$
Expert 8	$1.27 \cdot 10^{-09}$	0.691	0.693	$8.80 \cdot 10^{-10}$
Expert 9	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-02}$
Expert 10	$3.90 \cdot 10^{-04}$	0.542	0.540	$2.11 \cdot 10^{-04}$
Expert 11	$9.97 \cdot 10^{-03}$	0.554	0.514	$5.12 \cdot 10^{-03}$
Expert 12	$4.67 \cdot 10^{-17}$	1.239	1.219	$5.69 \cdot 10^{-17}$
Expert 13	$3.81 \cdot 10^{-06}$	0.963	0.930	$3.55 \cdot 10^{-06}$

Table 3.4: General Performance Peakload Prices

more informative or less informative. Moreover, we reward statistical accuracy over informativeness. Therefore, these slight increases and decreases have no significant contribution to the combined scores.

The increase in the calibration scores, thus the statistical accuracy, is important. Naturally, this increased the combined scores as well. Notice that expert 9 has a calibration score much closer to 1 regarding the baseload prices. The score is roughly 5 times as high as the previous calibration score. Hence, expert 9 has an expert empirical distribution very close to the inter-percentile probability vector. However, expert 9 has a very low calibration score regarding the peakload prices. We can trace this back to their expert empirical distribution. That is, we see that expert 9 has captured only 4 more baseload realisation compared to the peakload realisation. From figure C.9, we can then conclude that the expert empirical distribution regarding the peakload prices most likely differs a lot from the inter-percentile probability vector. Resulting in a very low calibration score. This in turn causes the lower calibration score display in table 3.1. Similar explanations hold for the other experts as well.

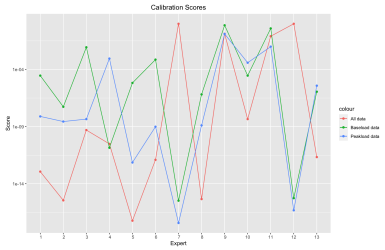

Figure 3.3: Calibration scores of the experts. We compare the scores computed using all seed variables (red), the baseload seed variables (green) and the peakload seed variables (blue).

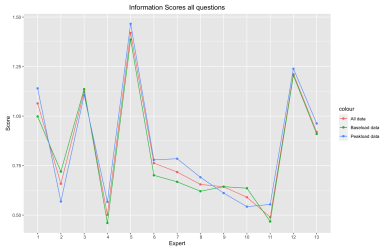

Figure 3.4: Information scores all questions of the experts. We compare the scores computed using all seed variables (red), the baseload seed variables (green) and the peakload seed variables (blue).

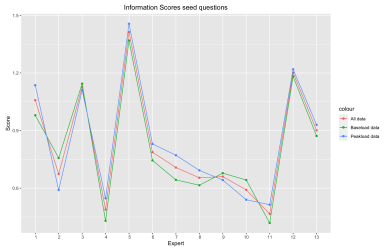

Figure 3.5: Information scores seed questions of the experts. We compare the scores computed using all seed variables (red), the baseload seed variables (green) and the peakload seed variables (blue).

We can zoom in a bit further. Next to the baseload-peakload division, we can also bin the dates. That is, we can distinguish between weeks, work-weeks and weekends. We define weeks as Monday-Sundays, work-weeks as Monday-Friday and weekends as Saturday-Sundays. Note that the scores regarding weekends are based on a very small data set. This in turn results in very different scores compared to week and work-week data.

First, we use the complete data set, i.e. we do not distinguish between baseload and peakload prices, and calculate the scores on a week-level. We aggregate the weeks, i.e. we start with week 1 and add weeks to the data. Resulting in figure D.7.

Experts all start with a calibration score between 0.000042 and 0.7. Through the week most expert calibra-

tion scores drop gradually. Two experts remain on roughly the same level throughout the weeks, i.e. experts 9 and 11. Experts 7 and 10 decrease rather quickly. Within the first half of the elicitation expert 10 drops to a calibration score equal to zero. Expert follows after week 6. We see the same pattern in the combined scores. Furthermore, we see the overconfidence of expert 1 reflected in the information scores. During the second half of this study expert 1 corrects this. Most experts have constant information score throughout the weeks. A few experts, i.e. expert 3,7,9,13, start with very high information scores but reach a constant value around week 4. This is understandable. Experts becomes more familiar with the method, the prices and the variables behind the market price. Reflection can then result in a higher uncertainty, thus in a different information score.

Aggregating the data can impact expert scores negatively. That is, if an expert performance poorly during one week, that performance influences the next calculated scores. Therefore, we also consider the previous situation without aggregating the weeks. We split the data the same way, but now consider each week separately. This results in figure 3.6.

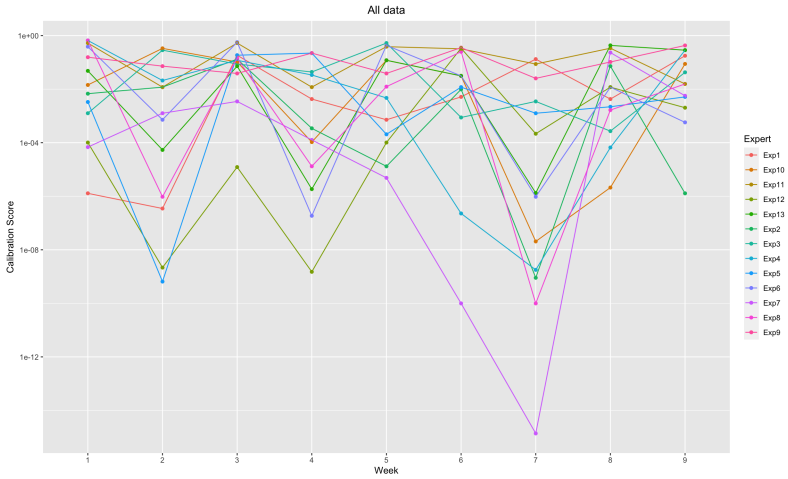

Figure 3.6: Expert performance on a weekly basis. Presented are the calibration scores. Weeks are not aggregated.

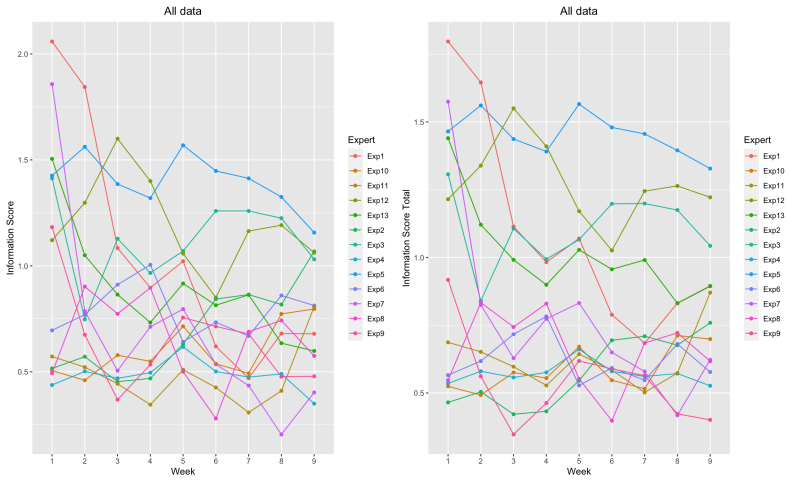

Figure 3.7: Expert performance on a weekly basis. Presented are the information score all questions and information score seed questions. Weeks are not aggregated.

Clearly, the calibration scores follow a very different pattern here. We can observe very fluctuating calibration scores for each expert. Figure D.7 presents a gradual decrease of the scores throughout the weeks. Figure 3.6 shows no gradual decrease of the scores nor a constant level for most experts. Experts 9 and 11 are the exception. They roughly remain on the same level. Other experts, for example expert 12, display heavy peaks and drops in their calibration scores. Moreover, experts do not seem to increase and decrease as a group.

Each week seems to have a different impact on each expert calibration score. The information scores display more fluctuation as well.

However, the differences between figure D.7 and figure 3.7 are not as prominent as the calibration scores. We still see that the information scores of a subset of experts remain around the same level. Furthermore, just as in D.7, experts 1,3,7,9 and 13 start with a high information score but adjust the information scores throughout the weeks and end up with a rather constant value.

As said before, we can also distinguish between work-weeks and weekends. During work-weeks energy consumption and production is rather high, higher than during the weekends. Offices and industry demand energy, resulting in an generation increase as well. Naturally, this impacts the prices. During the weekends, consumption can mostly be traced back to consumers. This target group demands a smaller volume of energy, which again impact the prices.

Again, we distinguish between aggregating the weeks and not aggregated the weeks. Figures D.9 and D.10 present us with much better results for expert 12. However, expert 7 seems to perform worse due to the exclusion of the weekends. Overall, experts seems to perform better without the weekend data. We can observe the same increase in calibration scores in figure 3.8. Excluding weekend data does not result in significant changes regarding the information scores. That is, compared to figures D.7 and 3.7.

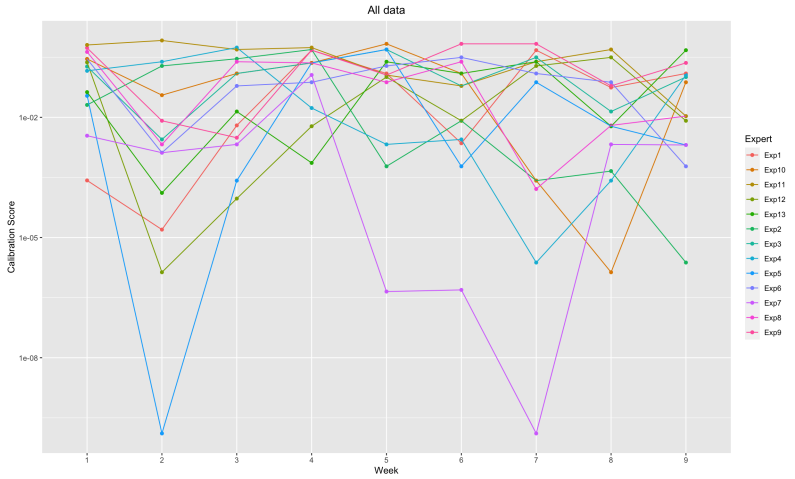

Figure 3.8: Expert performance on a work-week basis. Presented are the calibration scores. Weeks are not aggregated.

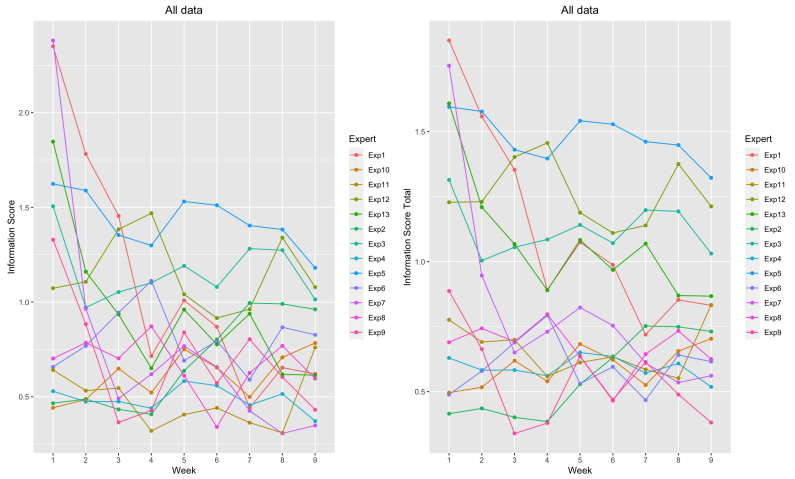

Figure 3.9: Expert performance on a work-week basis. Presented are the information score all questions and information score seed questions. Weeks are not aggregated.

We mention again that the scores in figures D.11, 3.10, 3.11 and D.12 are based on only 4 data points per week. We therefore interpret the scores lightly. We can observe that expert 12 performs rather poorly

when it comes to weekends. In the second half of the study performance improves. However, only two out of six weeks result in a high calibration score. We see that all experts have fluctuating calibration score in figure 3.10. Some experts have extreme high peaks and drops. This is reflected in figure D.11 where we see that these expert's calibration scores decrease sooner and faster. Furthermore, the information scores of the experts are not as high as during the weekdays. We can observe a lot of fluctuating values regarding the information scores in figure 3.11. However, when we aggregate the weekend, we see a rather constant line in D.11.

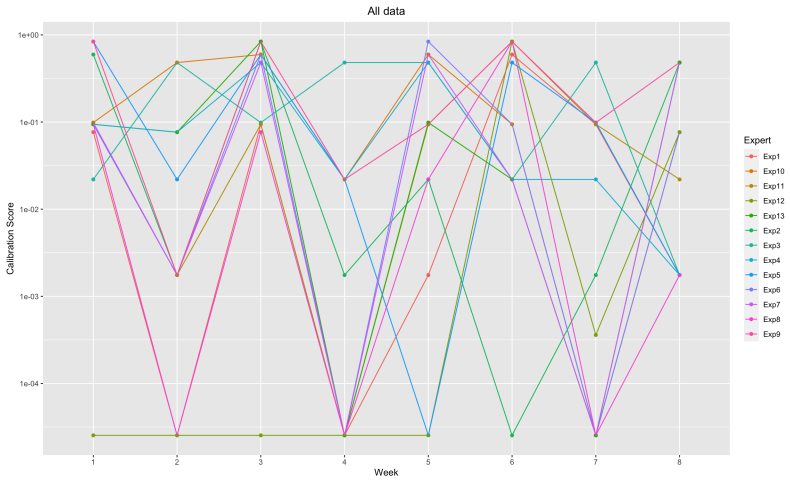

Figure 3.10: Expert performance on a weekend basis. Presented are the calibration scores. Weeks are not aggregated.

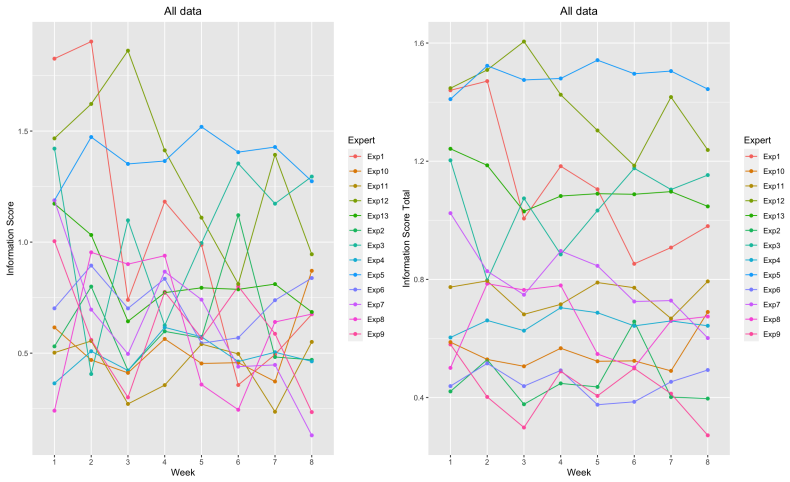

Figure 3.11: Expert performance on a weekend basis. Presented are the information score all questions and information score seed questions. Weeks are not aggregated.

We saw that distinguishing between baseload prices and peakload prices improved the calibration scores of the experts and thus the combined score. We also saw that most experts performed better when it came to baseload prices. This can clearly be observed in figure D.13. We see that almost all expert calibration scores decrease throughout the week. However, we observe no jumps to zero. Experts 5, 7 and 12 have the lowest calibration scores on the long term due to their performance throughout the weeks. That is, in figure 3.12 we can observe quite a few heavy drops in the data regarding the calibration scores for these experts. Furthermore, there are a few minor differences in the information scores compared to figure D.7. Resulting in the slight differences we have observed in tables 3.1 and 3.3.

Expert 12 performance much better on a work-week level. Weekends seem to lower their calibration score significantly.

Zooming in on just the weekend, we observe that most experts perform rather well throughout the elicitation. Expert 12 performs rather poorly. In figure 3.16 we can observe that expert 12 has a very low calibration

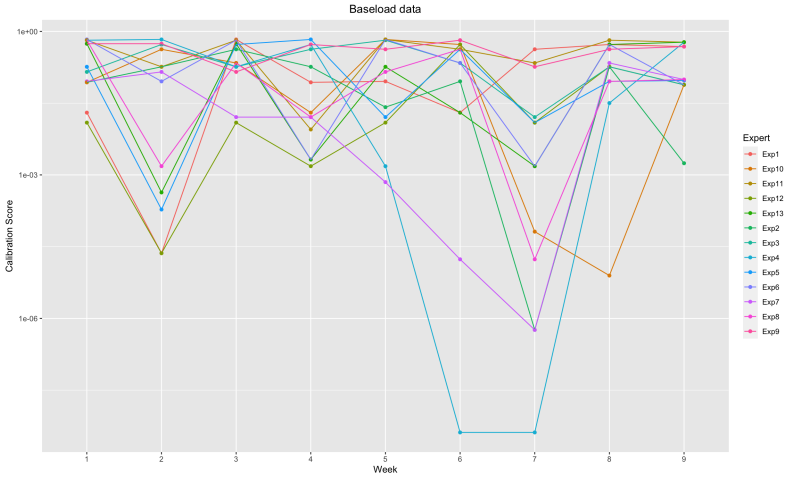

Figure 3.12: Expert performance on a week basis. Presented are the calibration scores. The scores are calculated for the baseload prices. Weeks are not aggregated.

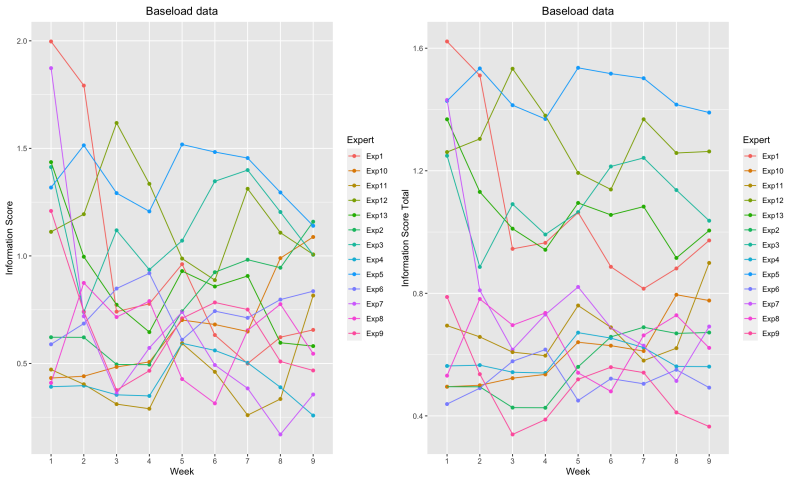

Figure 3.13: Expert performance on a week basis. Presented are the information score all questions and information score seed questions. The scores are calculated for the baseload prices. Weeks are not aggregated.

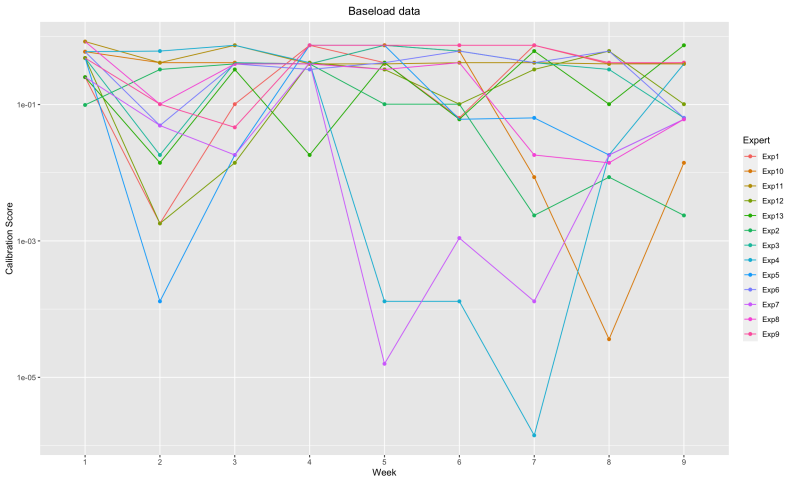

Figure 3.14: Expert performance on a work-week basis. Presented are the calibration scores. The scores are calculated for the baseload prices. Weeks are not aggregated.

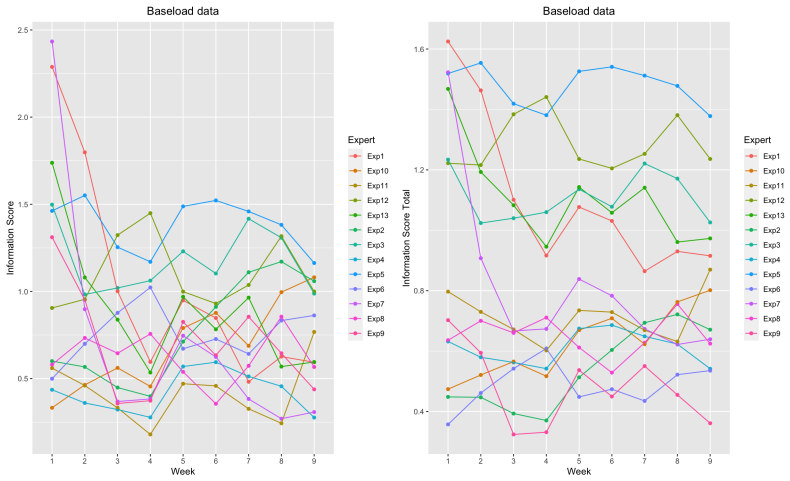

Figure 3.15: Expert performance on a work-week basis. Presented are the information score all questions and information score seed questions. The scores are calculated for the baseload prices. Weeks are not aggregated.

score during the first five weeks and drops again during week 7, resulting in the performance depicted in D.17. We see that expert 1 and 8 have significantly lower calibration scores compared to the other experts throughout the weeks. Both experts have rather heavy fluctuating calibration scores throughout the weeks when we calculate the scores without aggregating data. The heavy drops results in the decrease we observe in figure D.17.

Furthermore, the information scores are lower compared to the weekday data and full-week data. In figure D.17 these scores seem rather constant throughout the weeks, with a few experts as exception. However, figure 3.17 depicts heavy fluctuation in the information scores. The small size of the dataset, 2 assessments each week, could be the reason behind this.

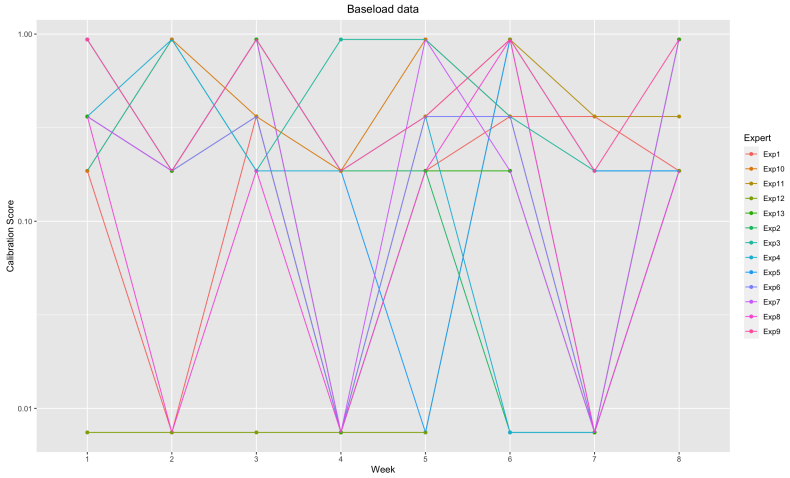

Figure 3.16: Expert performance on a weekend basis. Presented are the calibration scores. The scores are calculated for the baseload prices. Weeks are not aggregated.

Finally, we take a look at the scores based on the peakload prices. Again, we look at the scores for each week and compare by aggregating the weeks. We can see that the calibration scores are higher compared to the calibration scores of the complete data set, but lower compared to the calibration scores based on the baseload price set. Overall, we roughly see the same pattern as in the baseload price based figures. The differences are displayed in figure 3.18. We see different drops at different moments for different experts. For example, in figure 3.12 we see a prominent drop in the calibration score of expert 5 during weeks six and seven. Contrary to figure 3.18 where we see only a slight drop in week 2 for expert 5, but a heavy drop during week 7 for expert 7.

Again, just as with the baseload price based data, we see that the calibration score of expert 12 increases

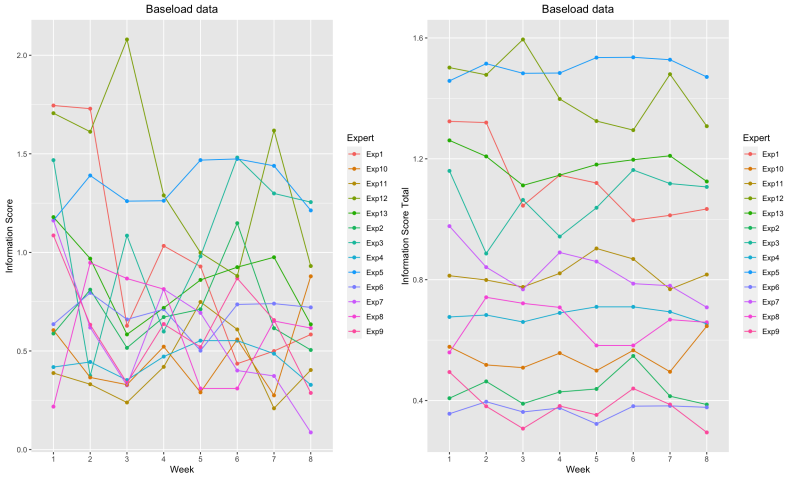

Figure 3.17: Expert performance on a weekend basis. Presented are the information score all questions and information score seed questions. The scores are calculated for the baseload prices. Weeks are not aggregated.

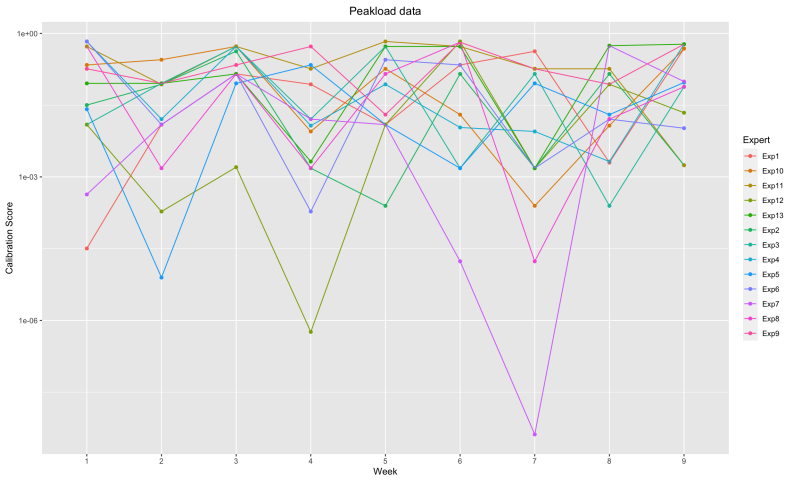

Figure 3.18: Expert performance on a week basis. Presented are the calibration scores. The scores are calculated for the peakload prices. Weeks are not aggregated.

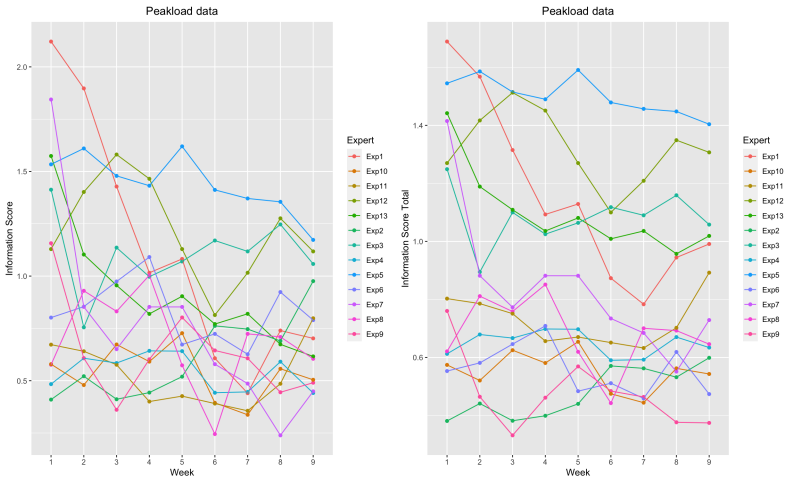

Figure 3.19: Expert performance on a week basis. Presented are the information score all questions and information score seed questions. The scores are calculated for the peakload prices. Weeks are not aggregated.

when we consider work-weeks. Experts 5 and 7 remain the expert with the lowest calibration scores and the heaviest decreases. In figure 3.20 we see that these experts experience a heavy drop in their calibration score during week 2 and 7, respectively.

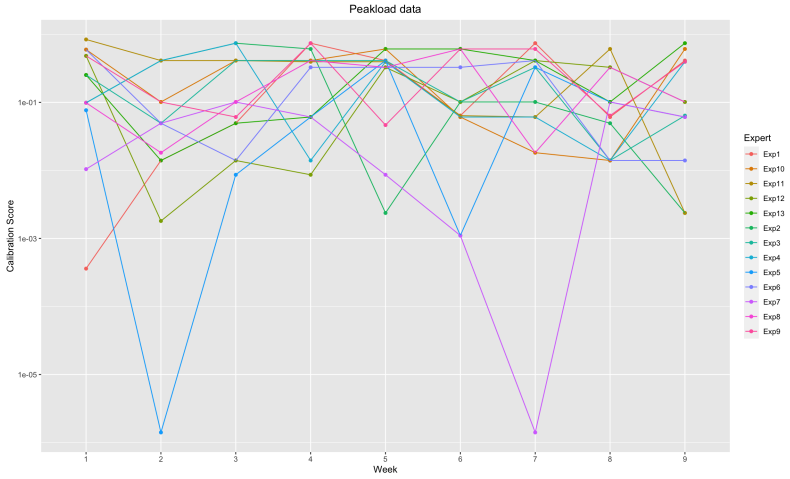

Figure 3.20: Expert performance on a work-week basis. Presented are the calibration scores. The scores are calculated for the peakload prices. Weeks are not aggregated.

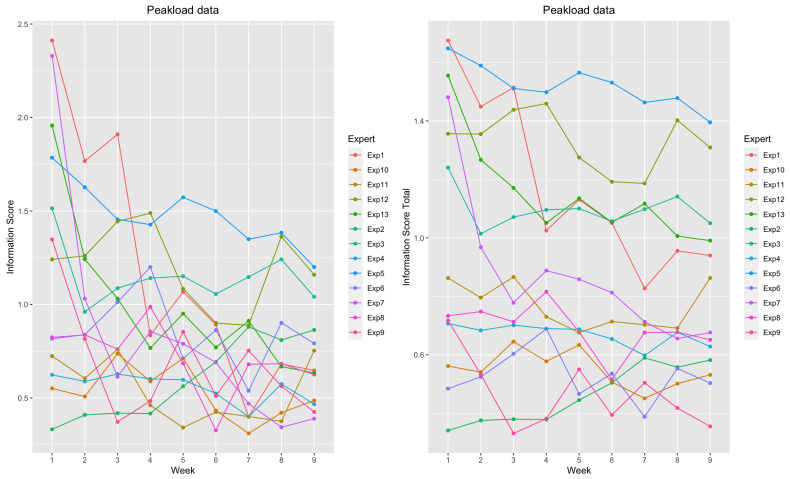

Figure 3.21: Expert performance on a work-week basis. Presented are the information score all questions and information score seed questions. The scores are calculated for the peakload prices. Weeks are not aggregated.

Lastly, we look at the weekend data regarding the peakload prices. Expert 12 performs rather poorly here as well. In figure 3.22 we can observe that expert 12 has a very low calibration score during the first five weeks and drops again during week 7, resulting in the performance depicted in D.23. We see that experts 1,8 and 13 have significantly lower calibration scores compared to the other experts throughout the weeks. These experts have rather heavy fluctuating calibration scores throughout the weeks when we calculate the scores without aggregating data. The heavy drops results in the decrease we observe in figure D.23.

We have seen that expert calibration scores do not seem to increase and decrease as a group. Each week seems to have a different impact on each expert calibration score. Distinguishing between full weeks, work-weeks and weekends has a different impact on each expert. Some experts seem to perform better with week-end data and other with work-week data.

Distinguishing between baseload data and peakload data has a clear impact on the performance of the experts. The scores of all experts improve by this division. However, experts with a constant performance remain the best.

Finally, note that current structured expert judgement studies, according to Roger Cooke's database, use 21 calibration questions at the most. Therefore, our dataset is rather unique. Allowing for this extensive

analysis of the expert scores.

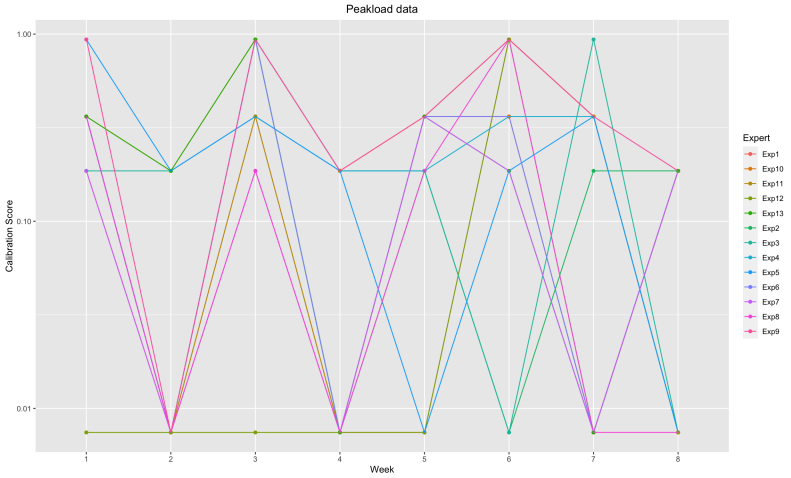

Figure 3.22: Expert performance on a weekend basis. Presented are the calibration scores. The scores are calculated for the peakload prices. Weeks are not aggregated.

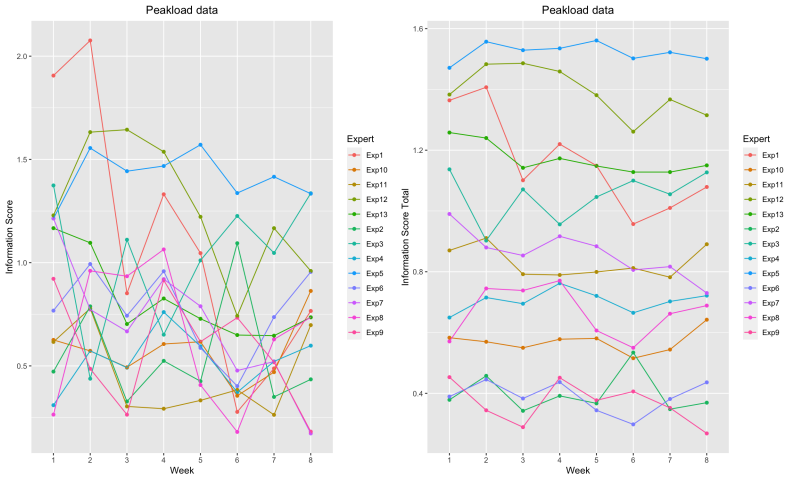

Figure 3.23: Expert performance on a weekend basis. Presented are the information score all questions and information score seed questions. The scores are calculated for the peakload prices. Weeks are not aggregated.

3.3.2. Decision Makers

As discussed in the Methodology of this report, expert scores lead to weights which are in turn used to create a decision maker. There are multiple possibilities for the weight, i.e. global weights, equal weight and item weights. Moreover, the cutoff value α increases this pool. Resulting in numerous possibilities for decision makers, which we will discuss here.

Table 3.5 summerizes the calculated decision makers and the corresponding scores regarding performance. To compare performance of the decision makers, we compare the combined scores of the decision makers. That is, we say that decision maker X outperforms decision maker Y if decision maker X has an higher combined score than decision maker Y. Weighting schemes of each of the decision makers can be found in Appendix E.

The Global Weight Decision Maker, calculated via Excalibur, has the cutoff value α set to 0. That is, each expert is a assigned a weight to create the decision maker. Coincidentally, in this case the optimized Global Weight Decision Maker equals the Global Weight Decision Maker with cutoff value $\alpha = 0$.

The optimized GWDM is the best performing decision maker. That is, it already applies the optimal cutoff value α. However, table E.1 show us that most experts have a small part in the decision maker. Therefore,

ID	DM	Dataset	Calibration Score	Information score all questions	Information score seed questions	Combined Score
DM1	GWDM $\alpha = 0$	Baseload & Peakload	0.311	0.340	0.342	0.106
DM2	GWDM optimized	Baseload & Peakload	0.311	0.340	0.342	0.106
DM3	GWDM $\alpha = 0.08138$	Baseload & Peakload	0.311	0.340	0.342	0.106
DM4	GWDM $\alpha = 0.1332$	Baseload & Peakload	0.133	0.643	0.660	0.088
DM5	EWDM	Baseload & Peakload	0.062	0.207	0.206	0.013
DM6	IWDM $\alpha = 0$	Baseload & Peakload	0.564	0.376	0.376	0.212
DM7	IWDM optimized	Baseload & Peakload	0.564	0.376	0.376	0.212
DM8	IWDM $\alpha = 0.081138$	Baseload & Peakload	0.564	0.376	0.376	0.212
DM9	IWDM $\alpha = 0.1332$	Baseload & Peakload	0.133	0.643	0.660	0.088
DM10	GWDM $\alpha = 0$	Baseload	0.038	0.346	0.350	0.013
DM11	GWDM optimized	Baseload	0.739	0.644	0.678	0.501
DM12	GWDM $\alpha = 0.3$	Baseload	0.038	0.347	0.351	0.014
DM13	EWDM	Baseload	0.094	0.196	0.193	0.018
DM14	IWDM $\alpha = 0$	Baseload	0.063	0.383	0.383	0.023
DM15	IWDM optimized	Baseload	0.739	0.644	0.678	0.501
DM16	IWDM $\alpha = 0.3$	Baseload	0.063	0.382	0.383	0.024
DM17	GWDM $\alpha = 0$	Peakload	0.625	0.425	0.438	0.274
DM18	GWDM optimized	Peakload	0.625	0.440	0.454	0.285
DM19	GWDM $\alpha = 0.1292$	Peakload	0.129	0.611	0.642	0.083
DM20	EWDM	Peakload	0.338	0.220	0.219	0.074
DM21	IWDM $\alpha = 0$	Peakload	0.557	0.459	0.475	0.264
DM22	IWDM optimized	Peakload	0.557	0.475	0.492	0.274
DM23	IWDM $\alpha = 0.1292$	Peakload	0.129	0.611	0.643	0.083

Table 3.5: Overview of the decision makers.

we do not expect the decision makers scores to decrease significantly when we consider only the two best performing experts. That is, set the cutoff value to $\alpha = 0.08138$.

As expected, the decision maker performs equally well. This is caused by the very low weights assigned to the other 11 experts. The decision maker changes significantly when we exclude the second best expert.

Naturally, the decision maker has the same scores as the expert when we cutoff all experts except for the best performing expert. We also see that the performance of the decision maker decreases. The combined score of the first three decision makers is higher than the combined score of this last decision maker.

We have seen that we can also obtain the best performance by including all the experts. Each expert has a different weight, thus a different contribution to the decision maker. We would also like to analyse what happens if we include all experts but apply equal weights.

From table 3.5 we can conclude that the EWDM does not result in a better performing decision maker when compared with the GWDM. That is, the GWDM has a combined score of 0.1064. A score higher than the combined score of the EWDM which equals 0.01271. We consider another option. The Item Weight Decision Maker. We start by setting the cutoff value $\alpha = 0$.

Remember that the calibration scores of the experts remain the same when applying Item Weights. The difference lies in aggregation of the information scores. However, we know that statistical accuracy is more important than informativeness. Hence, the calibration score has more influence on the combined score than the information score. Therefore, we do not expect the Item Weight Decision Maker to have much higher combined score compared to the Global Weight Decision Maker. However, we see that the Item Weight Decision Maker has a better performance than all the Global Weight Decision Makers.

Just like with the Global Weight Decision Maker, we analyse the Item Weight Decision Maker by calculating the optimized decision maker and setting the cutoff value α such that we are left with the two best performing expert and with the best performing expert. Hence, we set $\alpha = 0.08138$ and $\alpha = 0.1332$ respectively. The optimized decision maker again equals the IWDM with cutoff value $\alpha = 0$.

Yet again, we see that the performance of the IWDM with cutoff value $\alpha = 0.08798$ performs equally well

as the optimized IWDM, which coincidentally equals the IWDM with cutoff value $\alpha = 0$. However, setting the cutoff value to $\alpha = 0.1332$, thus assigning a weight to the best performing expert only, results in a decision maker with scores equal to that of the expert. These scores are lower than scores of the previous three decision makers. Hence, we see again that the decision maker based on only the best performing expert does not result in the best performing decision maker.

We analysed the general performance of the expert by distinguishing between scores calculated based on baseload prices and scores calculated based on peakload prices. This increased expert performance in both cases. Naturally, we are interested in the decision makers calculated by distinguishing between baseload prices and peakload prices as well. We start by considering decision makers calculated based on the baseload prices.

We set the cutoff value to $\alpha = 0$ to obtain the decision maker in table E.8. This is the Global Weight Decision Maker based on scores determined using the baseload price assessments. We compare this decision maker with the optimized Global Weight Decision Maker, obtained using Excalibur. The optimized decision maker assigns full weight to the best performing expert. Hence, the decision maker scores equal the expert scores. Notice that the combined score of the optimized GWDM is higher than the combined score of the GWDM with cutoff value $\alpha = 0$.

Theoretically, the optimized decision maker has the best performance compared to the decision makers with other values for α. For validations sake, we consider the Global Weight Decision Maker based on the two best performing experts. That is, the GWDM with $\alpha = 0.3$.

It is clear that the optimized GWDM has the best performance. However, we see a slight change in the information scores when we compare the decision makers with cutoff values $\alpha = 0$ and $\alpha = 0.3$. The information scores of the decision maker with cutoff value $\alpha = 0.3$ are slightly higher compared the information scores of the decision maker with $\alpha = 0$. This results in a slightly higher combined score for the decision maker, thus a better performance.

Naturally, we do not expect the Equal Weight Decision Maker to have a better performance compared to the optimized GWDM. However, for a complete analysis we do consider the decision maker.

We see that the optimized Global Weight Decision Maker has a better performance compared to the Equal Weight Decision Maker. However, The Equal Weight Decision Maker has a better performance when compared to all other GWDM.

The Item Weight Decision Maker with $\alpha = 0$ has a better performance compared to the GWDM and the EWDM, with the exception of the optimized GWDM. However, the optimized IWDM displayed in tabel E.13 has a performance equal to the performance of the optimized GWDM. Hence, both decision makers have the best performance. Setting the cutoff value to $\alpha = 0.3$ increases the combined score of the IWDM compared to the IWDM with cutoff value $\alpha = 0$. However, the optimized IWDM still has a better performance.

Finally, we consider decision makers with scores based on the peakload prices. Our first decision maker is the Global Weight Decision Maker with the cutoff value set to $\alpha = 0$. We compare the Global Weight Decision Maker with cutoff value set to $\alpha = 0$, table E.15, with the optimized Global Weight Decision Maker, table E.16.

We see that the optimized GWDM assigns weights to 2 experts. This results in the same calibration score as the GWDM with cutoff value $\alpha = 0$. However, the information score improves slightly, resulting in a slightly higher combined score. Assigning full weight to best performing expert decreases the combined score of the GWDM due to the lower combined score of the best performing expert.

Table E.18 presents us the decision maker when we assign equal weights to the experts. We see that the EWDM has the worst performance compared to all GWDM.

Finally, we apply the item weights. The Item Weights Decision Maker with cutoff value $\alpha = 0$ is presented in table E.19. We compare this with the optimized Item Weight Decision Maker as presented in table E.20

We see that both decision makers have the same calibration score. The difference in calibration scores is caused by the difference in information scores. We see that the optimized IWDM has a slightly higher information score, resulting in a slightly combined score. Assigning full weight to the best performing expert decreases the performance of the decision maker.

We distinguished between the full data set, i.e. scores based on all assessments, the baseload data set, i.e. scores based on the baseload price assessments, and the peakload data set, i.e. scores based on the peakload price assessments. When we consider the full data set we saw that the Item Weight Decision Maker had the best performance, i.e. the highest combined score. The optimized IWDM, the IWDM with $\alpha = 0$ and the IWDM with $\alpha = 0.08138$ all have the same combined score, thus the best performance. The optimized IWDM and the IWDM with $\alpha = 0$ assign weights to all expert. Experts 9 and 11 have the highest weights. Experts 7 and 12 are assigned a weight equal to 0 due to their calibration score which also equals 0. The IWDM with

$\alpha = 0.1332$, where we assign full weight to expert 9, has third best performance with a combined score equal to 0.08798. This equals the combined score of the GWDM with $\alpha = 0.1332$ where we also assign full weight to expert 9. The optimized GWDM, the GWDM with $\alpha = 0$ and the GWDM with $\alpha = 0.08138$ have the second best performance with a combined score equal to 0.1064. Just as with the IWDM, the optimized GWDM and the GWDM with $\alpha = 0$ assign weights to all expert. Experts 9 and 11 have the highest weights. Experts 7 and 12 are assigned a weight equal to 0 due to their calibration score which also equals 0. The EWDM has the lowest performance when we consider the full data set with a combined score equal to 0.01271.

The optimized decision makers have the best performance when we consider the baseload data set. We see that the optimized GWDM and the optimized IWDM have the best performance with a combined score equal to 0.5009. Both these decision makers assign full weight to expert 9. The IWDM with $\alpha = 0.3$ has the second best performance with a combined score equal to $2.428 \cdot 10^{-2}$. This decision maker assigns weight to expert 9 and 11, where expert 9 is assigned significantly more weight compared to the weight assigned to expert 11. The IWDM with $\alpha = 0$ has a slightly lower weight. The weights are assigned to all experts for this decision maker. The EWDM has the fourth best performance with a combined score equal to $1.81 \cdot 10^{-2}$. The GWDM with $\alpha = 0.3$ outperforms the GWDM with $\alpha = 0$ with only a slight difference. The GWDM with $\alpha = 0.3$ assigns weight to expert 9 and 11, where expert 9 is assigned significantly more weight compared to the weight assigned to expert 11. The GWDM with $\alpha = 0$ assigns weight to all experts, where expert 9 is assigned significantly more weight compared to the weight assigned to expert 11 and expert 11 is assigned significantly more weight compared to the weight assigned to all other experts.

The optimized GWDM has the best performance when we consider the peakload data set. The optimized GWDM assigns weight experts 9 and 11, where expert 9 is assigned significantly more weight compared to the weight assigned to expert 11. The second best performing decision makers are GWDM with $\alpha = 0$ and the IWDM optimized. The GWDM with $\alpha = 0$ assigns weights to all experts, where expert 9 is assigned significantly more weight compared to the weight assigned to expert 11 and expert 11 is assigned significantly more weight compared to the weight assigned to all other experts. The optimized IWDM assigns weight to experts 9 and 11. Where expert 9 is assigned significantly more weight compared to the weight assigned to expert 11. The GWDM and IWDM with $\alpha = 0.1292$, where full weight is assigned to expert 9, has the fourth best performance. Leaving the EWDM with the lowest combined score compared to the other decision makers.

Note that the three highest combined scores are obtained by the optimized GWDM and the optimized IWDM based on the baseload data set, the optimized GWDM based on the peakload data set and the GWDM with $\alpha = 0$ and the optimized IWDM based on the peakload data set.

Finally, consider DM11, DM15 and DM18. These are the best performing decision makers with a difference between baseload prices and peakload prices. Plotting the calibration and combination scores of the decision makers against the weeks gives us the following,

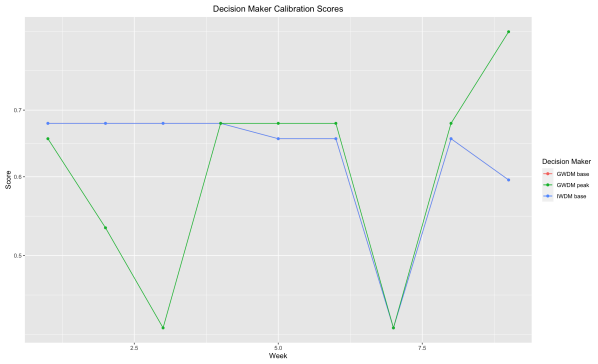

Figure 3.24: Calibration scores of the decision makers throughout the weeks. We consider the full week here.

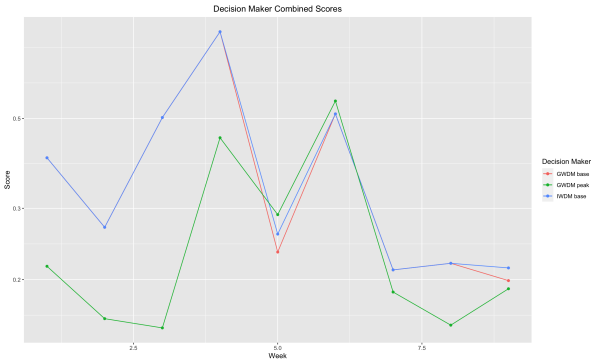

Figure 3.25: Combined scores of the decision makers throughout the weeks. We consider the full week here.

Note that the scores of the GWDM base (red) equal the scores of the IWDM base (blue) in figure 3.24. Hence, the blue line lies on top of the red line. We see that the calibration scores of the baseload decision makers are rather constant during the first half of the study. We see a heavy drop in week 7. The peakload decision maker, DM18, fluctuates more. We see two prominent drops in week 3 and week 7. We also see a prominent increase in the last week. Note that the baseload decision makers are identical.

The combined scores of the baseload decision makers exhibit some differences. However, these differences are marginal compared to the differences between the scores of the baseload decision makers and the

peakload decision maker. The decision makers seem to follow the same pattern. That is, the combined scores of all the decision makers seem to increase and decrease at the same moments in time.

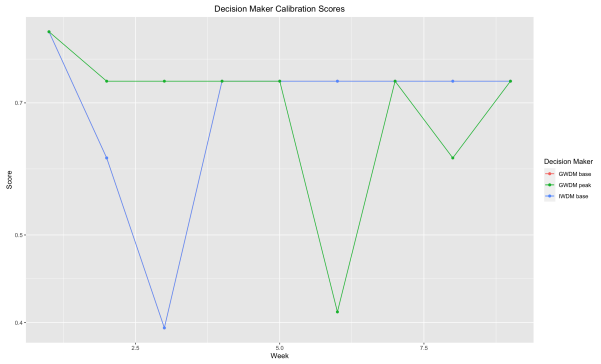

Figure 3.26: Calibration scores of the decision makers throughout the weeks. We consider work-weeks here.

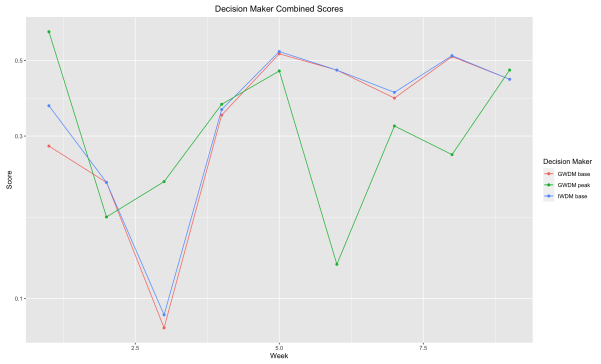

Figure 3.27: Combined scores of the decision makers throughout the weeks. We consider work-weeks here.

Excluding the weekends from the data causes an earlier drop in the baseload decision maker calibration score. We see that the score drops in week 3 instead of week 7. Furthermore, the peakload decision maker has one heavy drop instead of two. Again, note that the scores of the GWDM base (red) equal the scores of the IWDM base (blue) in figure 3.26. Hence, the blue line lies on top of the red line.

The combined scores of the peakload decision maker are higher in the first few weeks compared to the baseload decision makers. However, the combined scores of the baseload decision makers improve significantly during the last half.

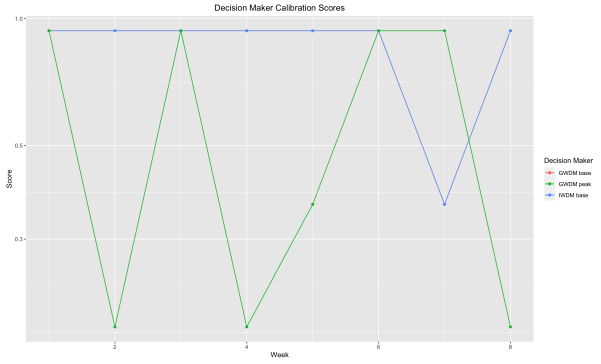

Figure 3.28: Calibration scores of the decision makers throughout the weeks. We consider weekends here.

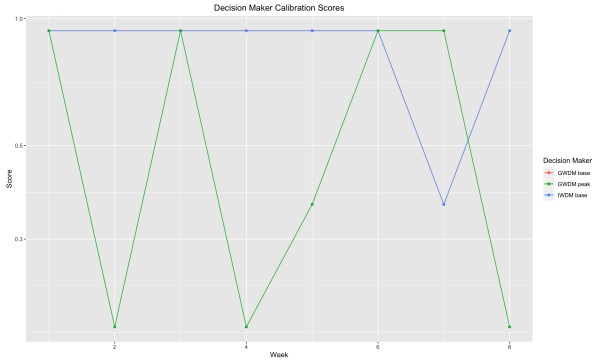

Figure 3.29: Combined scores of the decision makers throughout the weeks. We consider weekends here.

The weekend dataset contains a low amount of data points. This needs to be kept in mind while evaluating the scores. Moreover, the scores of the GWDM base (red) equal the scores of the IWDM base (blue) in figures 3.28 and figure 3.29. Hence, the blue line lies on top of the red line. We see that the baseload decision makers perform rather well throughout the weeks with regard to the weekend data if we consider the calibration scores. The peakload decision makers fluctuates a lot and the scores drop heavily from very high values to very low values.

The combined scores of the peakload decision maker are overall lower than the combined scores of the baseload decision makers. Again, we see the scores of the peakload decision maker fluctuate heavily.

3.3.3. Comparison with a data-driven model

We were able to obtain the point-wise forecast of a fundamental spot price forecast model. The model uses historical prices and weather information to forecast the next days spot prices. Energy companies currently use the model to obtain spot price forecasts.

We compared the performance of the model with the performance of our decision makers. That is, we calculated the calibration score, information scores and combined score of the model by interpreting the point-wise forecast of the model as the 50th percentile. This meant that we still needed a 5th and 95th percentile.

Therefore, we created a 90 percent confidence interval. We calculated the empirical standard deviation σ of the data and obtained a margin error by $\pm 1.64 \cdot \frac{\sigma}{\sqrt{n}}$. This margin error is then added to each data point for the upperbound of the confidence interval and deducted from each data point for the lowerbound of the confidence interval. The data and the created CI are presented in figure 3.30. Note that this is the approach used in constructing confidence intervals around the mean.

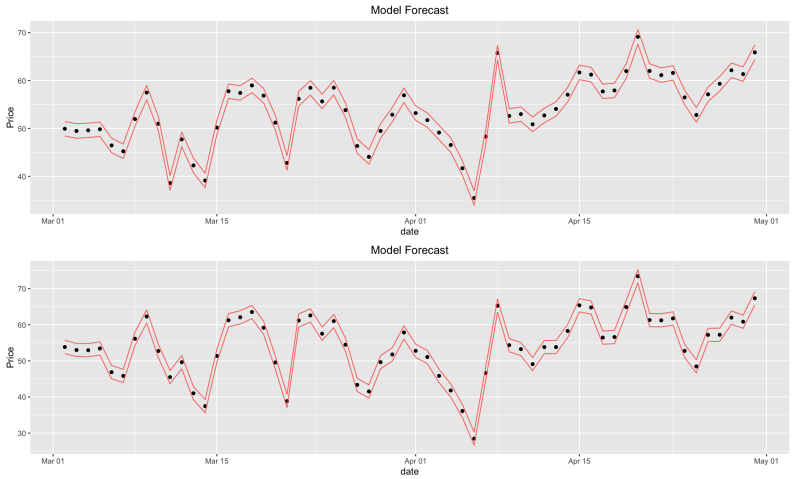

Figure 3.30: Point-wise forecast of the model (black dots) and the constructed 90 percent CI (red lines).

Clearly, the interval is extremely narrow. Unsurprisingly, the model has calibration scores equal to zero. Resulting in combined scores equal to zero. Using all the data, i.e. baseload data and peakload data, we obtained a calibration score equal to zero, an information score equal to 2,362 and a combined score equal to zero. Taking the baseload data subset, we obtained a calibration score equal to zero, an information score equal to 2.322 and a combined score equal to zero. The peakload data subset resulted in a calibration score equal to zero, an information score equal to 2,402 and a combination score equal to zero.

The information scores clearly indicates overconfidence and underestimation. This comes as no surprise, considering the narrow intervals depicted in figure 3.30. Naturally, we would like to stress that these scores are rather biased and should not lead to a conclusion regarding the performance of the model compared to the decision makers. More research is definitely needed into how to construct confidence intervals that could appropriately depict model uncertainty.

3.4. Results

THE previous section presented 23 decision makers. We use these decision makers to answer the questions of interest of this study. That is, we use these decision makers to forecast the future day average day-ahead Dutch electricity spot price on the EPEX spot market for 2025, 2030 and 2035. We distinguish between the baseload spot price and the peakload spot price. The forecasted prices are given in euro's per MWh. For convenience, we have labeled each decision maker and displayed the labels in tabel 3.5.

Some of the decision makers performed equally well. That is, these decision makers had the same scores. Coincidentally in this study, the question of interest assessments of these decision makers are the same.

We see that the 95th-percentile for 2035 is very similar for the decision makers. This holds for the baseload prices and the peakload prices. The biggest discrepancy can be found in the 5th-percentile for 2025 in table 3.6. We see that DM3, the GWDM with $\alpha = 0.08138$ based on the full data set, provides us with 5.11 euro's per MWh compared to the average value of the other decision makers, 35 euro's per MWh.

On average, we see that we obtain a 5-th percentile value for 2025 of 31.55 euro's per MWh regarding the baseload prices. An average 50-th percentile value for 2025 of 50.36 euro's per MWh regarding the baseload prices. An average 95-th percentile value for 2025 of 64.98 euro's per MWh regarding the baseload prices. The average 5-th percentile value for 2030 is lower compared to 2025 with an average value equal to 24.28 euro's per MWh. The same hold for the average 50-th percentile value for 2030 equal to 47.04 euro's per MWh. The average 95-th percentile value for 2030 is slightly higher than the 2025 average value with an average value equal to 65.31 euro's per MWh. Finally, we have an average 5-th percentile value for 2035 of 25.33 euro's per MWh. An average 50-th percentile value for 2035 of 45.70 euro's per MWh. And an average 95-th percentile

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
DM1	35.11	50.74	64.89	25.33	47.14	64.78	25.50	49.18	65.00
DM2	35.11	50.74	64.89	25.33	47.14	64.78	25.50	49.18	65.00
DM3	5.11	50.74	64.89	25.33	47.14	64.78	25.50	49.18	65.00
DM4	35.00	50.00	65.00	25.00	45.00	65.00	25.00	45.00	65.00
DM5	21.57	48.40	65.59	15.89	47.78	69.68	21.42	48.16	64.06
DM6	35.24	51.12	64.78	26.10	48.42	64.28	27.11	52.70	65.00
DM7	35.24	51.12	64.78	26.10	48.42	64.28	27.11	52.70	65.00
DM8	35.24	51.12	64.78	26.10	48.42	64.28	27.11	52.70	65.00
DM9	35.00	50.00	65.00	25.00	45.00	65.00	25.00	45.00	65.00
DM10	35.09	50.66	64.92	25.27	46.90	64.84	25.41	48.59	65.00
DM11	35.00	50.00	65.00	25.00	45.00	65.00	25.00	45.00	65.00
DM12	35.09	50.62	64.92	25.25	46.84	64.83	25.38	48.54	65.00
DM13	21.57	48.40	65.59	15.89	47.78	69.68	21.42	48.16	64.06
DM14	35.21	51.07	64.81	25.99	48.32	64.37	26.91	52.39	64.99
DM15	35.00	50.00	65.00	25.00	45.00	65.00	25.00	45.00	65.00
DM16	35.21	51.05	64.81	25.97	48.31	64.37	26.87	52.44	65.00

Table 3.6: Question of Interest 1, 3 and 5. What will be the average electricity baseload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

value for 2035 of 64.88 euro's per MWh. Therefore, we see that on average uncertainty grows as we move on from 2025 to 2030 and 2035. The uncertainty range remains roughly the same for 2030 and 2035. Moreover, we expect the baseload prices to decrease with roughly five euro's over 10 years.

We obtain an average 5-th percentile value for 2025 of 17.67 euro's per MWh regarding the peakload prices. An average 50-th percentile value for 2025 of 44.31 euro's per MWh. An average of 95-th percentile value for 2025 of 64.04 euro's per MWh. The average 5-th percentile value for 2030 is lower than for 2030 with an average value equal to 16.09 euro's per MWh. The average 50-th percentile value for 2030 is slightly higher than for 2030 with an average value equal to 45.82 euro's per MWh. The average 95-th percentile value for 2030 equals 69.70 euro's per MWh. Furthermore, we have an average 5-th percentile value for 2035 of 7.29 euro's per MWh. An average 50-th percentile value for 2035 of 39.65 euro's per MWh. And an average 95-th percentile value for 2035 of 65.71 euro's per MWh. Therefore, we see that on average uncertainty grows as we move on from 2025 and 2030 to 2035. Moreover, we expect the peakload prices to decrease with roughly five euro's over 10 years. Finally, note that we expect a bigger spread, on average, for the peakload prices than for the baseload prices.

Finally, we observed in the previous section that the three best performing decision makers are DM11 and DM15, DM18 and DM17 and DM22. If we distinguish between the baseload prices and the peakload prices, we are left with DM11 and DM15 for the baseload prices and DM18 for the peakload prices. DM11 and DM15 have identical scores. Therefore, we see identical forecasts. This leads to the forecasts presented in table 3.8. These are the forecasts based on the best performing decision makers.

Note that uncertainty is greater for the peakload prices. For 2025 and 2030 we see that the average price has a range of 15-65 euro's per MWh. For 2035 this range is even bigger. The uncertainty regarding the baseload prices grows after 2025. We can see that the range expands from 35-65 euro's per MWh to 25-65 euro's per MWh. Moreover, the average Dutch electricity day-ahead baseload price, on a year level, is expected to drop from 50 euro's per MWh in 2025 to 45 euro's per MWh for 2030 and 2035 on the EPEX spot market. The average Dutch electritiy day-ahead peakload price, on a year level, is expected to drop from 41.55 euro's per MWh in 2025 and 2030 to 36.87 euro's per MWh for 2035.

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
DM1	15.84	45.14	64.63	15.83	48.27	70.26	6.02	41.49	65.80
DM2	15.84	45.14	64.63	15.83	48.27	70.26	6.02	41.49	65.80
DM3	15.84	45.14	64.63	15.83	48.27	70.26	6.02	41.49	65.80
DM4	15.00	40.00	65.00	15.00	40.00	65.00	5.00	35.00	65.00
DM5	18.59	47.20	61.89	16.54	44.72	70.14	9.14	42.30	66.89
DM6	24.71	47.76	62.86	18.53	54.87	70.83	12.53	42.80	65.97
DM7	24.71	47.76	62.86	18.53	54.87	70.83	12.53	42.80	65.97
DM8	24.71	47.76	62.86	18.53	54.87	70.83	12.53	42.80	65.97
DM9	15.00	40.00	65.00	15.00	40.00	65.00	5.00	35.00	65.00
DM17	15.13	41.55	64.93	15.13	41.44	67.98	5.16	36.84	65.30
DM18	15.12	41.55	64.94	15.12	41.56	68.03	5.15	36.87	65.36
DM19	15.00	40.00	65.00	15.00	40.00	65.00	5.00	35.00	65.00
DM20	18.59	47.20	61.89	16.54	44.72	70.14	9.14	42.30	66.89
DM21	16.78	46.34	64.21	15.53	45.27	69.74	6.21	41.55	65.73
DM22	16.79	46.40	64.29	15.50	45.94	69.85	6.17	41.66	65.82
DM23	15.00	40.00	65.00	15.00	40.00	65.00	5.00	35.00	65.00

Table 3.7: Question of Interest 2, 4 and 6. What will be the average electricity peakload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
Baseload	35.00	50.00	65.00	25.00	45.00	65.00	25.00	45.00	65.00
Peakload	15.12	41.55	64.94	15.12	41.56	68.03	5.15	36.87	65.36

Table 3.8: Questions of Interest answers, i.e. forecasts, based on the best performing decision makers. What will be the average electricity baseload/peakload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

3.5. Conclusion

THis sub-study aimed to forecast future Dutch electricity day-ahead spot prices on a year level using structured expert judgement. We applied Cooke's Classical Model during this structured expert judgement study. Furthermore, we distinguished between baseload prices and peakload prices. Through elicitation we obtained assessments regarding the day average Dutch electricity day-ahead baseload and peakload spot prices.

We used the assessments to measure expert performance. We saw that expert performance increased significantly when we distinguish between baseload and peakload prices. Moreover, experts performed better with the baseload prices compared to the peakload prices. This is probably due to the higher price volatility among the peakload prices. Experts 9 and 11 have the best performance in all situations. Experts 7 and 12 have the lowest combined scores in all situations. We saw that expert 9 obtained roughly the same calibration score through the weeks. The same holds for expert 11. Contrary to expert 7, where we saw that the calibration scores fluctuated a lot during the weeks. This resulted in a steady decrease of the calibration score and a low calibration score in total. Statistical accuracy is more important than informativeness. Hence, the high, low calibration score results in a high, low combined score. We also saw that the information scores remained rather constant through the weeks for the most experts. There were a few exceptions where experts had to correct for overconfidence or absence due to vacation. Generally, fluctuations in the information scores are smaller than in the calibration scores.

The expert scores lead to decision makers. We computed 23 decision makers, i.e. 9 decision makers based on all data, 7 decision makers based on baseload data and 7 decision makers based on peakload data. Note that we cannot use the decision makers based on the baseload data to forecast future peakload prices and vice versa. We saw that the optimized Global Weight Decision Maker and the optimized Item Weight Decision Maker based on baseload data and the optimized Global Weight Decision Maker based on peakload data performed best. Their combined scores were the highest among the 23 decision makers. The optimized Global Weight Decision Maker and the optimized Item Weight Decision Maker based on baseload data also have the highest information scores.

Based on these decision makers we obtain our desired forecasts. These forecasts are given in euro's per MWh on a year level, i.e. the year average. Uncertainty is greater for the peakload prices. For 2025 and 2030 we see that the average peakload price has a range of 15-65 euro's per MWh. For 2035 this range is even bigger. The uncertainty regarding the baseload prices grows after 2025. We can see that the range expands from 35-65 euro's per MWh to 25-65 euro's per MWh. Moreover, the average Dutch electricity day-ahead baseload price, on a year level, is expected to drop from 50 euro's per MWh in 2025 to 45 euro's per MWh for 2030 and 2035 on the EPEX spot market. The average Dutch electriticy day-ahead peakload price, on a year level, is expected to drop from 41.55 euro's per MWh in 2025 and 2030 to 36.87 euro's per MWh for 2035.

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
Baseload	35.00	50.00	65.00	25.00	45.00	65.00	25.00	45.00	65.00
Peakload	15.12	41.55	64.94	15.12	41.56	68.03	5.15	36.87	65.36

Table 3.9: Questions of Interest answers, i.e. forecasts, based on the best performing decision makers. What will be the average electricity baseload/peakload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

4

Measuring underlying market developments

CHAPTER 3 discusses an elicitation, and the corresponding results, where the seed variables are dutch electricity spot prices. The focus of this part of the study lies in knowledge. Specifically, knowledge concerning the Dutch electricity market. Past, current and future developments in the Dutch electricity markets play a prominent role in future electricity market prices. Therefore, we are interested in measuring knowledge on past, current and projected developments concerning the Dutch electricity markets and possible corresponding data. Hence, the elicitation in this chapter contains seed questions concerning the Dutch electricity market. That is, questions concerning energy conservation, energy demand developments, the development of the emission prices, capacity cross-boarder, the energy mix and its development, flow based market coupling etc.

This section discusses this second elicitation and the results in detail. We start by discussion our expert pool. Followed by the elicitation itself. We discuss the procedure, the Questions of Interest, the Calibration Questions and our data sources. Additionally, we briefly discuss our elicitation platform. We analyse the expert performance and the decision maker performance. Based on the best performing decision maker, we present the results. That is, the answers to the Questions of Interest based on the decision maker. Finally, we discuss the analysis and results and formulate a conclusion.

4.1. Experts

THE expert pool used for this part of the Structured Expert Judgement case study differs from the expert pool used for the first part of the Structured Expert Judgement case study. The aim of this part of the case study is to measure how well informed experts are on past, current and future data and events regarding electricity demand, production, consumption and prices, past, current and future electricity market developments and commodities with influences on the electricity prices. Therefore, we are interested in experts with expertise concerning the Dutch electricity market. Resulting in a target group set on traders, Dutch electricity consultants, electricity price and load forecasters, electricity portfolio managers and electricity portfolio analysts. Additionally, members of the Faculty Mechanical, Maritime and Materials Engineering and the Faculty Technology, Policy and Management at Delft University of Technology were added to the target group. Specifically the departments Energy Technology and Engineering Systems and Services, group Energy and Industry, respectively. Members of these departments are familiar with the Dutch electricity market and its developments and prices. Moreover, members of these departments have contributed to analysis regarding the Dutch electricity market and Dutch electricity prices [27].

Initially, 60 potential experts received an invitation to participate in the study. We received 21 responses. These 21 experts gave their consent to participate anonymously in the study. However, only 9 experts filled in the questionnaire before the deadline. Thus resulting in an expert pool of size 9.

The decrease in the expert pool, i.e. expected expert pool size based on the response rate versus the actual expert pool size, is subject to speculation. Experts received an invitation in May 2021 and were able to answer the calibration questions and questions of interest between 15 September 2021 and 1 October 2021. A few plausible causes for the decrease in the expert pool are the holidays, the heavy price fluctuations in quarter 3

and 4 of 2021 and content of the questionnaire. We can define the period June-August as the summer holiday period. Most people take leave from work during this period, resulting in a bigger work load at the beginning of September. Participation in this study case then moves down on most to-do lists or even disappears from the list. Furthermore, people unable to take leave in June-August take leave in September-October. Thus missing the elicitation period. Heavy price fluctuations and increases can be another reason for the decrease. Energy spot prices started increases during the summer, causing various employees a bigger work load and less time for sideline activities. This could have resulted in less participating experts than expected. Lastly, the questionnaire could have been either daunting to some potential experts or time consuming. Experts received an invitation stating that the questions concern the Dutch electricity market. That is, the questions concern energy conservation, energy demand developments, the development of the emission prices, capacity cross-boarder, the energy mix and its development, flow based market coupling etc. However, the detailed nature of the questions could have been daunting for some of the potential experts, causing them to not participate. The time indication for the questionnaire turned out to be accurate according to multiple participating experts. However, it is possible that potential experts scrolled through the questionnaire after receiving it and estimated that it would take them longer to answer the questions. Thus choosing to not participate in the case study.

Finally, the realised expert pool consists of electricity traders, consultants, electricity market and price analysts, electricity price and load forecasters and electricity portfolio managers. Our current expert pool is positively diverse. About a third of the pool is female. Experts age ranges from 25 to 50. Ideally, we would have liked to have a bigger expert pool. An increase participants is desired here. None of the TU Delft members participated in the study. Naturally, their participation would have been preferable. However, due to the time limitation, the corona crisis and the energy crisis of 2021, our nine experts are acceptable for the case study.

4.2. Elicitation

THE elicitation requires preparation. Decisions had to be made before reaching out to the experts. That is, how will the facilitator carry out the elicitation, how are we going to train the experts, but also, where do we find the required data. The elicitation procedure discusses these subjects.

4.2.1. Elicitation procedure

The elicitation of this part of the cast study was carried out online and remote using a platform called Minerva, partly due to the COVID-19 measures in the Netherlands. But also to motivate the invited experts to participate.

Possible experts received an invitation to participate per e-mail, Appendix B. The invitation briefly discussed the motivation behind the study. Moreover, the structured expert judgement method was introduced. We stressed that participation is anonymous and asked for the written consent of the participating experts.

The participating experts received a second e-mail directing them to the elicitation platform, Minerva. By clicking the link in the e-mail, the experts were directed to the platform. The experts were required to go through the training first. Introducing them to structured expert judgement, the scoring measures and familiarizing them with the assessment formats. Afterwards, the platform welcomed the experts, introducing them, again, to the structured expert judgement method. Providing them, again, with the motivation behind the study and navigating them through the platform.

4.2.2. Questions of Interest

Through the elicitation we obtain experts assessments concerning seed variables and questions of interest. In this study, the Questions of Interest aim to predict future Dutch electricity spot prices on the EPEX spot market. We are interested in the average price on a year level. Particularly for 2025, 2030 and 2035. Resulting in the following Questions of Interest.

- What will be the average electricity baseload spot price on the EPEX spot market in 2025?
- What will be the average electricity peakload spot price on the EPEX spot market in 2025?
- What will be the average electricity baseload spot price on the EPEX spot market in 2030?
- What will be the average electricity peakload spot price on the EPEX spot market in 2030?
- What will be the average electricity baseload spot price on the EPEX spot market in 2035?

- What will be the average electricity peakload spot price on the EPEX spot market in 2035?

The prices are given in euros per MWh. We distinguish between the baseload prices and the peakload prices. That is, peakloads refer to 08:00 - 20:00. Baseloads refer to a 24-hour time period.

4.2.3. Calibration Questions

We measure expert performance and quantify their uncertainty using seed variables. We use calibration questions to objectively evaluate the uncertainty assessments of the experts. There were 21 calibration questions in total, which can be found in Appendix F.

The data is extracted from various sources. Each question includes their source. Most of the data is freely accessible to the public.

The International Energy Agency [25] publishes energy statistics for countless countries. The Netherlands is one of those countries. On the IAE's website we can find historical and current information about the electricity generation, supply, demand and consumption in the Netherlands and much more. Their website also contains the Dutch energy mix over various year.

The website of Centraal Bureau Statistiek (CBS) provides us with similar statistics as the IAE does. However, navigation on the CBS website is not as easy as the IAE website. Nevertheless, the information found is very usefull.

On the website of the Government of the Netherlands we can find information regarding the National Climate Agreement and sustainability goals of the Dutch Government. Naturally, these influence the future electricity prices in the Netherlands. Hence, they are added as seed variables in this study.

Montel is known to all eletricity traders in the Neterlands. Formally Montel Online, Monel provides news and data to the European energy market. They provide transparency to the second since 1997. Most traders are subscribed to their news e-mail.

DNV GL is an international accredited registrar and classification society. They provide their costumers with facts and reliable insights regarding multiple energy sources and energy sector developments. The publish a lot of papers. Part of these papers discuss past, present and future electricity developments. Resulting in a quite a few interesting seed questions. Particularly those regarding projections.

PBL Netherlands Environmental Assessment Agency, Planbureau voor de Leefomgeving, is the national institute for strategic policy analysis in the fields of the environment, nature and spatial planning. They conduct research regarding the environment, nature and spatial planning. Resulting in papers containing useful information for our study.

Finally, the EPEXSPOT [3] provides us with Dutch electricity day-ahead spot prices on the EPEX market. These prices are given in euros per MWh for each hour of the day. Hence, we have to calculate averages to obtain the realisation of certain seed variables.

4.2.4. Minerva

The questions were asked via the platform Minerva, created by Excogent, a company focusing on uncertainty quantification using expert opinion. The platform follows the standards of data privacy (GDPR) and protection. To receive the questions, i.e. use the platform, the expert's e-mail address was necessary. This was the only piece of personal information needed for this research. The platform is available through Amazon Web Services cloud provider. At the end of the project, after downloading the data, experts' assessments were deleted from the platform.

4.3. Performance Analysis

ELICITATION has resulted in 9×27 assessments. Here, we distinguish between the 21 calibration questions and 6 questions of interest presented in the previous sections of this chapter. With these assessments and the corresponding realisations we can calculate the calibration score, information score and combined score for each expert. The information score can be calculated including assessments regarding the questions of interest or excluding the assessments regarding the questions of interest. We start by discussing these scores first.

4.3.1. Expert Performance

Table 4.1 presents us with the calibration score, information score all questions, information score seed questions and the combined score of each participating expert. Note that the experts labeling does not coincide with the previous study labeling.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$

Table 4.1: Expert performance in terms of calibration score, information score for seed questions and all questions, and combined score.

We can clearly see that expert 2 has the best performance, i.e. the highest combined score given by the highest calibration score. Expert 9 has the second best performance, however the combined scores are one order of magnitude apart. Hence, the scores are very different for both experts. The same statement holds for experts 4 and 6 respectively.

Expert 3 is the most informative, followed swiftly by expert 4. However, the low calibration score of expert 3 results in a low combined score and thus not the best performance. Expert 4 has a better calibration score, resulting in the third best performance.

Experts 5 and 8 have the lowest calibration scores, resulting in the lowest combined score. They do not have the lowest information scores. Their information scores are the third and fourth highest among the experts. However, statistical accuracy is more important than informativeness. Therefore, they obtain low combined scores.

Experts managed to capture 6-18 realisation regarding the seed variables. Table 4.2 present how many realisations each expert has captured during this study.

	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Exp9
#captures	11	18	10	14	6	14	10	6	15

Table 4.2: Number of realisations captured by experts 90% CI.

We see that expert 2, the expert with the best performance, has captured the most realisations. Expert 2 is rewarded with a high calibration score, due to their expert empirical distribution. Expert 8 has captured the lowest amount of realisations and also has the lowest calibration score. The expert empirical distribution of expert 8 is probably not optimal, resulting in the lowest score. The same holds for experts 4 and 9.

When we zoom in and look at the range plots of the experts in Appendix G, we see that expert 8 constantly overestimates. Expert 8 captures most of the realization in the second inter-quantile range. Resulting in low calibration score. Expert 2 roughly captures an even amount of realizations in the second inter-quantile range and the third inter-quantile range. We can not conclude overconfidence based on the intervals of expert 2.

Expert 9 seems to exhibit overestimating behaviour. That is, of the captured realizations, most seem to be captured in the second inter-quantile range. Suggesting overestimation. The same can be said about expert 4.

Expert 5 seems to capture an equal amount of realizations in the first and fourth inter-quantile range. Resulting in the lowest calibration score and thus in the lowest combined score.

Moreover, we notice that expert 1 provides rather wide assessments. Implying a lot of uncertainty. Expert 3 and expert 7 provide rather narrow assessments. Implying more certainty regarding their assessments.

Question 1 was very hard to answer for all experts. We see in figure G.1 that the realisation is not captured in any of the expert assessments. We see a high uncertainty among expert 8. Question 2 has three expert with a high uncertainty, i.e. expert 1, expert 3 and expert 9. Experts perform better with question 3. Expert 7 manages to capture the realisation with a very narrow interval. In figure G.2 we see that experts 1 and 3 are rather overconfident. Expert 2 misses the realisation by a small margin. Question 5 makes for relatively

wide interval among the experts. Question 6 is a hard question to answer for most experts, only two experts manage to capture the realisation. Moreover, experts tend to provide a very narrow interval. In figure G.3 we see that nearly half of the experts manage to capture the realisations for questions 7 - 9. Most of these experts tend to provide a rather wide interval compared to those missing the realisation. Question 11 displays a big uncertainty for one expert. Expert 3 provides a rather big uncertainty, but misses the realisation. Questions 19 - 21 in figure G.7 are captured by nearly all experts.

4.3.2. Decision Makers

The expert scores lead to weights which are in turn used to create a decision maker. There are multiple possibilities for the weight, i.e. global weights, equal weight and item weights. Moreover, the cutoff value α increases this pool. Resulting in numerous possibilities for decision makers, which we present here in table 4.3 with the corresponding scores regarding performance.

ID	DM	Calibration Score	Information score all questions	Information score seed questions	Combined Score
DM1	GWDM	$5.38 \cdot 10^{-01}$	0.555	0.628	$3.38 \cdot 10^{-01}$
DM2	GWDM optimized	$6.55 \cdot 10^{-01}$	0.936	1.090	$7.10 \cdot 10^{-01}$
DM3	GWDM $\alpha = 0.09331$	$9.23 \cdot 10^{-01}$	0.578	0.652	$6.02 \cdot 10^{-01}$
DM4	EWDM	$3.96 \cdot 10^{-01}$	0.312	0.351	$1.39 \cdot 10^{-01}$
DM5	IWDM	$9.23 \cdot 10^{-01}$	0.694	0.799	$7.37 \cdot 10^{-01}$
DM6	IWDM optimized	$9.23 \cdot 10^{-01}$	0.694	0.800	$7.38 \cdot 10^{-01}$
DM7	IWDM $\alpha = 0.6546$	$6.55 \cdot 10^{-01}$	0.936	1.090	$7.10 \cdot 10^{-01}$
DM8	IWDM $\alpha = 0.09331$	$7.07 \cdot 10^{-01}$	0.707	0.813	$5.86 \cdot 10^{-01}$

Table 4.3: Decision Maker performance in terms of calibration score, information score for seed questions and all questions, and combined score.

The Global Weight Decision Maker, calculated via Excalibur, has the cutoff value α set to 0. That is, each expert is assigned a weight to create the decision maker. Unsurprisingly, we see in Appendix H that expert 2 is assigned the highest weight. Expert 9 is assigned a weight 8 times as low compared to expert 2. Expert 4 is again assigned a weight one order lower than expert 9. The other experts are all assigned weights orders much lower. The GWDM has a combined score lower than the combined score of expert 2. Hence, expert 2 outperforms our decision maker. This implies that the optimized GWDM will probably assign a higher weight to expert 2 for a better result.

We see that the optimized GWDM assigns full weight to expert 2, the best performing expert. Moreover, we see that the optimized GWDM has the best performance compared to the GWDM with $\alpha = 0$ in table H.1 and the GWDM with $\alpha = 0.09331$ in table H.3. The performance of GWDM decreases with roughly $\frac{1}{10}$ when we assign weight to expert 9 next to expert 2. We assign much more weight to expert 9 compared to expert 2. However, we see that the small weight of expert 2 results in a lower combined score for the GWDM. Naturally, assigning weight to more experts on the long run leads to the performance of the GWDM with $\alpha = 0$.

We compare decision maker performance by comparing the combined scores of the decision. However, the calibration score of the GWDM with $\alpha = 0.09331$ is higher than the calibration score of the optimized GWDM. The information score of the optimized GWDM is higher in this situation. Apparently, the optimized GWDM is rewarded the highest combined score due to the preferred calibration score - information score ratio.

Another possible decision maker is the Equal Weight Decision Maker. We assign the same weight to each expert. The scores of the obtained EWDM can be found in table H.4. We see that we have obtained lowest score yet. Next to the lowest combined score yet, we see that the calibration score and the information score are both lower compared to the Global Weight Decision Makers.

The final group of decision makers we consider is the Item Weight Decision Maker. We immediately see that the IWDM with cutoff value set to $\alpha = 0$ has a better performance than the optimized GWDM. That is, we see that the combined score of the IWDM with $\alpha = 0$ is higher than the combined score of expert 2. The IWDM with $\alpha = 0$ has a lower information score compared to the optimized GWDM, but it has a much higher calibration score. The calibration score of the IWDM with $\alpha = 0$ is as high as the calibration score of the GWDM with $\alpha = 0.09331$.

The optimized IWDM has an even better performance. The combined score is the highest yet. It assigns weight to the three best performing experts. We see that a different aggregation of the information scores results in a better performance of the decision maker. The calibration score of the optimized IWDM is one of the highest, but the optimized GWDM has the better information score. All in all, the optimized IWDM has the best performance.

Theoretically, adjusting α will not result in a better performing IWDM. However, we consider three more Item Weight Decision Makers to complete the analysis. We have the IWDM with $\alpha = 0.6546$, i.e. assigning full weight to expert 2. As expected, the performance of this decision maker is not better than the performance of the optimized IWDM. Naturally, the decision maker's scores are identical to the scores of expert 2. We already concluded that the calibration score here is lower than the calibration score of the optimized IWDM, but the information score is higher. Assigning weight to the second best performing expert results in the IWDM with $\alpha = 0.09331$ presented in table H.8. We see that the calibration score of this decision maker increases compared to the previous decision maker. The information score and combined score decrease. Assigning weight to expert 4 results in the optimized IWDM. Hence, we see that the calibration score increase with the addition of an expert, but the information score decrease. Once expert 4 is assigned a weight the combination of this lower information score but much higher combination score results in the highest combined score. Assigning a weight to the next best expert slightly decrease the information score, resulting in a slightly lower combined score.

It is clear that the optimized IWDM is the best performing decision maker. This is the decision maker with the highest score. The EWDM has the lowest combined score. We see that the optimized decision makers both have the highest information scores. However, the optimized IWDM has the higher calibration score, resulting in the highest combined score.

There are three questions concerning baseload and peakload prices in the second study, questions 13-15. If we delete those questions and calculate the decision makers again, we obtain the optimized IWDM as the best performing decision maker with,

DM	Calibration Score	Information score all questions	Information score seed questions	Combined Score
IWDM new	0.637	0.922	1.091	0.695
IWDM previous	0.923	0.694	0.800	0.738

Table 4.4: Best performing decision makers based on the second study. The first decision maker (new) is based on 18 calibration question, excluding the three questions concerning baseload and peakload prices. The second decision maker (previous) is based on all 21 calibration questions.

The optimized IWDM remains the decision maker with the best performance. However, weight is now assigned to one expert instead of three. Moreover, the combined score of this decision maker is lower than the combined score of the best performing decision maker based on all 21 questions.

4.4. Results

THE previous section presented 8 decision makers. We use these decision makers to answer the questions of interest of this study. That is, we use these decision makers to forecast the future day average day-ahead Dutch electricity spot price on the EPEX spot market for 2025, 2030 and 2035. We distinguish between the baseload spot price and the peakload spot price. The forecasted prices are given in euro's per MWh. For convenience, we have labeled each decision maker and displayed the labels in tabel 4.3.

We see that uncertainty grows as we move further into the future. That is, the uncertainty range becomes bigger and is biggest in 2035 for most decision makers. On average, we see that we obtain a 5-th percentile value for 2025 of 38.64 euro's per MWh regarding the baseload prices. An average 50-th percentile value for 2025 of 59.54 euro's per MWh regarding the baseload prices. An average 95-th percentile value for 2025 of 81.12 euro's per MWh regarding the baseload prices. The average 5-th percentile value for 2030 is lower compared to 2025 with an average value equal to 29.33 euro's per MWh. The same hold for the average 50-th percentile value for 2030 equal to 50.58 euro's per MWh. The average 95-th percentile value for 2030 is also lower than the 2025 average value with an average value equal to 74.16 euro's per MWh. Finally, we have an average 5-th percentile value for 2035 of 24.85 euro's per MWh. An average 50-th percentile value for 2035 of

46.37 euro's per MWh. And an average 95-th percentile value for 2035 of 77.77 euro's per MWh. Therefore, we see that on average uncertainty grows as we move on from 2025 to 2030 and 2035. We also see that the range shifts over the years. We expect the baseload prices to decrease with roughly fifteen euro's over 10 years.

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
DM1	35.91	59.75	82.42	30.04	50.94	76.22	25.12	47.15	78.14
DM2	45.00	60.00	80.00	30.00	50.00	70.00	25.00	45.00	75.00
DM3	35.69	59.40	80.00	30.00	50.54	75.39	25.09	46.97	78.26
DM4	29.37	57.51	83.49	24.54	51.72	78.66	23.25	45.93	81.46
DM5	39.44	59.95	81.52	30.03	50.56	74.68	25.12	47.02	78.04
DM6	39.44	59.95	81.52	30.03	50.57	74.68	25.12	47.02	78.05
DM7	45.00	60.00	80.00	30.00	50.00	70.00	25.00	45.00	75.00
DM8	39.30	59.77	80.00	30.00	50.28	73.62	25.09	46.84	78.17

Table 4.5: Question of Interest 1, 3 and 5. What will be the average electricity baseload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

We obtain an average 5-th percentile value for 2025 of 33.76 euro's per MWh regarding the peakload prices. An average 50-th percentile value for 2025 of 63.81 euro's per MWh. An average of 95-th percentile value for 2025 of 86.43 euro's per MWh. The average 5-th percentile value for 2030 is lower than for 2030 with an average value equal to 25.13 euro's per MWh. The average 50-th percentile value for 2030 is much lower than for 2030 with an average value equal to 48.27 euro's per MWh. The average 95-th percentile value for 2030 equals 80.17 euro's per MWh. Furthermore, we have an average 5-th percentile value for 2035 of 20.18 euro's per MWh. An average 50-th percentile value for 2035 of 44.89 euro's per MWh. And an average 95-th percentile value for 2035 of 82.38 euro's per MWh. Therefore, we see that on average uncertainty grows as we move on from 2025 and 2030 to 2035. Moreover, we expect the peakload prices to decrease with roughly twenty euro's over 10 years. Note that we expect a bigger spread, on average, for the peakload prices than for the baseload prices.

Finally, we have coloured the forecast of the decision maker with the best performance red in table 4.5 and table 4.6. These are our best forecasts. Again, we see a bigger spread for the peakload prices than for the baseload prices. This spread grows as time passes. Hence, the 2035 forecast comes with a higher uncertainty. Moreover, we see a slight shift in the assessments when we are comparing the years. We expect the baseload prices to fall in the range of 39.44-81.52 euro's per MWh in 2025. By 2035 this becomes 25.12-78.05 euro's per MWh. We see a similar shift in the forecast of the peakload prices. Here we expect prices to fall in the range of 32.97-86.19 euro's per MWh in 2025. By 2035 this becomes 20.22-80 euro's per MWh.

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
DM1	30.56	64.24	86.91	25.17	48.35	79.69	20.02	45.02	79.99
DM2	40.00	65.00	85.00	25.00	45.00	75.00	20.00	40.00	80.00
DM3	30.25	64.21	84.94	25.18	48.07	79.00	20.19	45.03	80.00
DM4	30.51	58.51	92.22	24.96	53.00	94.59	20.43	45.89	99.09
DM5	32.97	64.51	86.19	25.23	48.95	79.43	20.22	47.68	79.99
DM6	32.97	64.51	86.19	25.23	48.94	79.43	20.22	47.69	80.00
DM7	40.00	65.00	85.00	25.00	45.00	75.00	20.00	40.00	80.00
DM8	32.79	64.50	84.97	25.23	48.85	79.20	20.32	47.77	80.00

Table 4.6: Question of Interest 2, 4 and 6. What will be the average electricity peakload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

4.5. Conclusion

THis sub-study aimed to forecast future dutch electricity day-ahead spot prices on a year level using structured expert judgement. We applied Cooke's Classical Model to determine the decision makers and obtain the desired forecasts during this structured expert judgement study. Contrary to the previous chapter, we measured expert performance using data on past, current and future developments in the Dutch electricity markets. However, the forecast data still distinguished between baseload and peakload prices.

We used assessments to measure expert performance. We saw that the best performing expert did not capture the most, or even a high number of, realisations. However, their fortunate expert empirical distribution resulted in the highest calibration score. This in turn resulted in the highest combined score. The expert with the most realisation captures obtained one of the lowest calibration scores due to their unfortunate expert empirical distribution. The big discrepancies between the calibration scores result in low contributions of the information scores.

The expert scores lead to decision makers. We computed 8 decision makers. The optimized Item Weight Decision Maker had the best performance. However, the Global Weight Decision Maker with $\alpha = 0.09331$ has a combined score not much lower. We saw that both decision makers have the same combined score. The information score became the tie breaker.

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
Baseload	39.44	59.95	81.52	30.03	50.57	74.68	25.12	47.02	78.05
Peakload	2.97	64.51	86.19	25.23	48.94	79.43	20.22	47.69	80.00

Table 4.7: Questions of Interest answers, i.e. forecasts, based on the best performing decision makers. What will be the average electricity baseload/peakload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

Using the optimized Item Weight Decision Maker we obtained our desired forecasts, table 4.7. We saw a bigger spread for the peakload prices than for the baseload prices. This spread grows as time passes. Hence, the 2035 forecast comes with a higher uncertainty. Moreover, we saw a slight shift in the assessments when we were comparing the years. We expected the baseload prices to fall in the range of 39.44-81.52 euro's per MWh in 2025. By 2035 this became 25.12-78.05 euro's per MWh. We saw a similar shift in the forecast of the peakload prices. Here we expected prices to fall in the range of 32.97-86.19 euro's per MWh in 2025. By 2035 this became 20.22-80 euro's per MWh. The average baseload price in 2020 was 32.24 euro's per MWh, where the average peakload price was 35.27 euro's per MWh. We see that experts expect the baseload prices to increase with roughly 20 euro's per MWh in upcoming years compared to 2020. The peakload prices are expected to increase with roughly 25 euro's per MWh by 2025 compared to 2020 and with roughly 15 euro's per MWh compared to 2020. The differences with the 2021 prices are much larger. The average bsaeload price in 2021 was 102.96 euro's per MWh, where the average peakload price was 111.03 euro's per MWh. Compared to 2021 expert expect a price drop in the upcoming years. The baseload prices and the peakload price for 2025 are expected to drop with 50% compared to 2021. The peakload prices for 2030 and 2035 are expected to drop with roughly 60 euro's per MWh compared to 2021.

5

Comparison of the two studies

We have conducted two structured expert judgement studies to forecast future Dutch electricity spot prices. During the first study, see Chapter 3, our aim was to measure how well our experts can forecast electricity spot prices. Experts had to forecast the average day baseload dutch electricity spot price and the average day peakload dutch electricity spot price for the upcoming day. We used their performance to forecast future electricity spot prices.

As mentioned before, future electricity spot prices do not necessarily depend on current electricity spot prices. Past, current and future developments in the Dutch electricity markets play a prominent role in future electricity market prices. This resulted in the second study, i.e. chapter 4, where the seed questions concern the Dutch electricity market. That is, questions concerning energy conservation, energy demand developments, the development of the emission prices, capacity cross-boarder, the energy mix and its development, flow based market coupling etc. Our aim was to measure knowledge on past, current and projected developments concerning the Dutch electricity markets and possible corresponding data.

This chapter compares the forecasts resulting from each of the studies. We dedicate a section to each forecast year, i.e. a section for 2025, a section for 2030 and a section for 2035. We start with a brief outline of both studies. We close this chapter with a conclusion.

5.1. Remarks

We obtain forecasts and uncertainty intervals for the average Dutch electricity baseload/peakload spot price on the EPEX spot market from both studies, i.e. forecasts for 2025, 2030 and 2035. However, there are differences between the studies. First, the expert pool during the first and the second elicitation are not the same. The aim of the first elicitation is to measure how well experts can forecast electricity spot prices. Therefore, we created an expert pool consisting out of traders, electricity analysts, portfolio managers but also electricity business consultants and sales employees. Every expert is linked to the commodity, but not necessarily directly to the spot market. Resulting in an expert pool of size 13.

The aim of the second elicitation is to measure how knowledge concerning the Dutch electricity market. Past, current and future developments in the Dutch electricity markets play a prominent role in future electricity market prices. Therefore, we are interested in measuring knowledge on past, current and projected developments concerning the Dutch electricity markets and possible corresponding data. Hence, we restricted ourselves to electricity traders, electricity analysts, portfolio managers, Dutch electricity consultants and electricity price and load forecasters during the second elicitation. Resulting in an expert pool of size 9. There are some overlaps between the two expert pools. That is, 4 experts have participated in both studies.

Next to the different expert pools and different calibration questions, the forecasts are based on different decision makers. The first study uses the optimized GWDM or the optimized IWDM for the baseload prices. That is, all weight is assigned to one expert. The optimized GWMD is used for the peakload prices. That is, all weight is assigned to the two best performing experts. The second study uses the optimized IWDM for the baseload and the peakload prices. That is, all weight is assigned to the three best performing experts.

A final note is on the duration of both studies. The first elicitation took place from 1 March 2021 - 1 May 2021. That is, each expert had to provide an assessment each day for 60 days. The second elicitation took

place between 15 September 2021 and 1 October 2021. Experts had to provide 21 assessments. However, the second study did not ask daily participation from the experts. Experts could provide the 21 assessments all at once. The first study can therefore be considered somewhat exhausting for the experts. Resulting in less accurate assessments towards the end.

5.2. 2025 forecast

TABLE 5.1 presents the forecasts of the baseload and peakload prices for 2025. We distinguish between the first study and the second study, i.e. the study described in Chapter 3 and the study described in Chapter 4 respectively. We see that a bigger uncertainty is quantified during the second study with respect to both the baseload price and the peakload price.

Moreover, we see that the 50th percentile of both studies differs significantly. The first study forecast an average baseload price of 50 euro's per MWh for 2025. Where the second study forecasts an average peakload price of 59.95 euro's per MWh, roughly 10 euro's per MWh more. The average peakload price is forecasted at 41.55 euro's per MWh according to the first study. The second study forecasts roughly 25 euro's per MWh more with 64.51 euro's per MWh.

	First study			Second study		
	5%	50%	95%	5%	50%	95%
Baseload	35.00	50.00	65.00	39.44	59.95	81.52
Peakload	15.12	41.55	64.94	2.97	64.51	86.19

Table 5.1: Questions of Interest answers, i.e. forecasts, for 2025. Hence the forecast for the average electricity baseload/peakload spot price on the EPEX spot market in 2025. We distinguish between the forecast obtained via the first study and the forecast obtained via the second study.

5.3. 2030 forecast

TABLE 5.2 presents the forecasts of the baseload and peakload prices for 2030. We have already observed a drop in the 50th percentile values compared to 2025. The uncertainty intervals seem to be of the same size in both studies. However, the 5th and 9th percentile are forecasted higher during the second study compared to the first study for both the baseload prices and the peakload prices.

Moreover, the differences between the 50th percentiles seem to be smaller when we compare the first and the second study. The first study forecast an average baseload price of 45 euro's per MWh for 2030. Where the second study forecasts an average peakload price of 50.57 euro's per MWh, roughly 5 euro's per MWh more. The average peakload price is forecasted at 41.56 euro's per MWh according to the first study. The second study forecasts roughly 7.50 euro's per MWh more with 48.94 euro's per MWh.

	First study			Second study		
	5%	50%	95%	5%	50%	95%
Baseload	25.00	45.00	65.00	30.03	50.57	74.68
Peakload	15.12	41.56	68.03	25.23	48.94	79.43

Table 5.2: Questions of Interest answers, i.e. forecasts, for 2030. Hence the forecast for the average electricity baseload/peakload spot price on the EPEX spot market in 2030. We distinguish between the forecast obtained via the first study and the forecast obtained via the second study.

5.4. 2035 forecast

LASTLY, we have the 2035 forecast. Table 5.3 presents the forecasts of the baseload and peakload prices for 2035. As mentioned in the previous chapter, the 50th percentile values are lower compared to 2025 and 2030. The uncertainty interval for the baseload prices during the second study are slightly larger compared to the first study. The 95th percentile lies roughly 13 euro's per MWh higher. For the peakload prices the uncertainty interval seems to be of the same size in both studies. However, the 5th and 9th percentile are forecasted higher during the second study compared to the first study for both the baseload prices and the peakload prices.

Moreover, the differences between the 50th percentiles for the baseload prices seem to be quite small when we compare the first and the second study. The first study forecast an average baseload price of 45 euro's per MWh for 2035. Where the second study forecasts an average peakload price of 47.02 euro's per MWh, roughly 2 euro's per MWh more.

The difference is bigger when we consider the peakload prices. The average peakload price is forecasted at 36.87 euro's per MWh according to the first study. The second study forecasts roughly 11 euro's per MWh more with 47.69 euro's per MWh.

	First study			Second study		
	5%	50%	95%	5%	50%	95%
Baseload	25.00	45.00	65.00	25.12	47.02	78.05
Peakload	5.15	36.87	65.36	20.22	47.69	80.00

Table 5.3: Questions of Interest answers, i.e. forecasts, for 2035. Hence the forecast for the average electricity baseload/peakload spot price on the EPEX spot market in 2035. We distinguish between the forecast obtained via the first study and the forecast obtained via the second study.

5.5. Conclusion

CHAPTER 3 and 4 discussed the differences between the forecasts for 2025, 2030 and 2035. We saw that expert expect prices to drop throughout the years. In some cases we saw uncertainty decrease as well. This chapter compared the forecasts of the first study with the forecast of the second study and discussed these for each forecast year, i.e. 2025, 2030 and 2035.

We can conclude that the forecasts originating from the second study are consistently higher compared to the forecasts originating from the second study. However, the uncertainty intervals in 2030 are roughly of the same size for both studies. In 2025, these are clearly bigger for the second study. The uncertainty interval in 2035 for the baseload prices during the second study is slightly larger compared to the first study. For the peakload prices the uncertainty interval seems to be of the same size in both studies.

Moreover, we see differences in the 50th percentile of both studies. However, these differences seem to become smaller throughout the years. The first study forecast an average baseload price of 50 euro's per MWh for 2025. Where the second study forecasts an average peakload price of 59.95 euro's per MWh, roughly 10 euro's per MWh more. The average peakload price is forecasted at 41.55 euro's per MWh according to the first study. The second study forecasts roughly 25 euro's per MWh more with 64.51 euro's per MWh.

These prices all decrease for 2030 according to the experts. The first study forecast an average baseload price of 45 euro's per MWh for 2030. Where the second study forecasts an average peakload price of 50.57 euro's per MWh, roughly 5 euro's per MWh more. The average peakload price is forecasted at 41.56 euro's per MWh according to the first study. The second study forecasts roughly 7.50 euro's per MWh more with 48.94 euro's per MWh.

Finally, the prices decrease again for 2035 according to all the experts. However, this decrease is smaller. The first study forecast an average baseload price of 45 euro's per MWh for 2035. Where the second study forecasts an average peakload price of 47.02 euro's per MWh, roughly 2 euro's per MWh more. The average peakload price is forecasted at 36.87 euro's per MWh according to the first study. The second study forecasts roughly 11 euro's per MWh more with 47.69 euro's per MWh.

6

Discussion and conclusion

THis project aimed to forecast future dutch electricity day-ahead spot prices on a year level using structured expert judgement. We obtained two different forecasts using two different kinds of seed variables. The seed variables during the first elicitation regarded the day average dutch electricity day-ahead baseload and peakload spot prices. During the second elicitation we measured expert performance using data on past, current and future developments in the Dutch electricity markets. We applied Cooke's Classical Model during this structured expert judgement study.

We distinguished between baseload prices and peakload prices. Recall that this resulted in an increase in expert performance based on the first elicitation. Moreover, experts performed better with the baseload prices compared to the peakload prices. This is probably due to the higher price volatility among the peakload prices. Experts 9 and 11 have the best performance in all situations. Experts 7 and 12 have the lowest combined scores in all situations. We saw that expert 9 obtained roughly the same calibration score through the weeks. The same holds for expert 11. Contrary to expert 7, where we saw that the calibration scores fluctuated a lot during the weeks. This resulted in a steady decrease of the calibration score and a low calibration score in total. Statistical accuracy is more important than informativeness. Hence, the high, low calibration score results in a high, low combined score. We also saw that the information scores remained rather constant through the weeks for the most experts. There were a few exceptions where experts had to correct for overconfidence or absence due to vacation. Generally, fluctuations in the information scores are smaller than in the calibration scores.

We computed 23 decision makers based on the first elicitation, i.e. 9 decision makers based on all data, 9 decision makers based on baseload data and 9 decision makers based on peakload data. We saw that the optimized Global Weight Decision Maker and the optimized Item Weight Decision Maker based on baseload data and the optimized Global Weight Decision Maker based on peakload data performed best. Their combined scores were the highest among the 23 decision makers. The optimized Global Weight Decision Maker and the optimized Item Weight Decision Maker based on baseload data also have the highest information scores.

Based on these decision makers we obtain our first set of desired forecasts. These forecasts are given in euro's per MWh on a year level, i.e. the year average. Uncertainty is greater for the peakload prices. For 2025 and 2030 we see that the average peakload price has a range of 15-65 euro's per MWh. For 2035 this range is even bigger. The uncertainty regarding the baseload prices grows after 2025. We can see that the range expands from 35-65 euro's per MWh to 25-65 euro's per MWh. Moreover, the average dutch electricity day-ahead baseload price, on a year level, is expected to drop from 50 euro's per MWh in 2025 to 45 euro's per MWh for 2030 and 2035 on the EPEX spot market. The average dutch electriticy day-ahead peakload price, on a year level, is expected to drop from 41.55 euro's per MWh in 2025 and 2030 to 36.87 euro's per MWh for 2035.

In the second study we saw that the best performing expert did not capture the most, or even a high number of, realisations. However, their fortunate expert empirical distribution resulted in the highest calibration score. That is, the realizations were captured mostly and evenly in the second and third inter-quantile range. This in turn resulted in the highest combined score. The expert with the most realisation captures obtained one of the lowest calibration scores due to their unfortunate expert empirical distribution. The big discrepancies between the calibration scores result in low contributions of the information scores.

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
Baseload	35.00	50.00	65.00	25.00	45.00	65.00	25.00	45.00	65.00
Peakload	15.12	41.55	64.94	15.12	41.56	68.03	5.15	36.87	65.36

Table 6.1: Questions of Interest answers, i.e. forecasts, based on the best performing decision makers of the first elicitation. What will be the average electricity baseload/peakload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

We computed 8 decision makers based on the second elicitation. The optimized Item Weight Decision Maker had the best performance. However, the Global Weight Decision Maker with $\alpha = 0.09331$ has a combined score not much lower. We saw that both decision makers have the same combined score. The information score became the tie breaker.

Using the optimized Item Weight Decision Maker we obtained our second set of desired forecasts. We saw a bigger spread for the peakload prices than for the baseload prices. This spread grows as time passes. Hence, the 2035 forecast comes with a higher uncertainty. Moreover, we saw a slight shift in the assessments when we were comparing the years. We expected the baseload prices to fall in the range of 39.44-81.52 euro's per MWh in 2025. By 2035 this became 25.12-78.05 euro's per MWh. We saw a similar shift in the forecast of the peakload prices. Here we expected prices to fall in the range of 32.97-86.19 euro's per MWh in 2025. By 2035 this became 20.22-80 euro's per MWh.

	2025			2030			2035		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
Baseload	39.44	59.95	81.52	30.03	50.57	74.68	25.12	47.02	78.05
Peakload	2.97	64.51	86.19	25.23	48.94	79.43	20.22	47.69	80.00

Table 6.2: Questions of Interest answers, i.e. forecasts, based on the best performing decision makers. What will be the average electricity baseload/peakload spot price on the EPEX spot market in 2025, 2030 and 2035 respectively?

The second elicitation results in a decision maker with a higher price forecast compared to the decision maker of the first elicitation. Uncertainty regarding the forecast is higher for 2025 and 2030 during the second study. Uncertainty is equally high regarding the peakload price forecast for 2035 in both studies, however the range shifts.

We can conclude that the forecasts originating from the second study are consistently higher compared to the forecasts originating from the second study. However, the uncertainty intervals in 2030 are roughly of the same size for both studies. In 2025, these are clearly bigger for the second study. The uncertainty interval in 2035 for the baseload prices during the second study is slightly larger compared to the first study. For the peakload prices the uncertainty interval seems to be of the same size in both studies.

Moreover, we see differences in the 50th percentile of both studies. However, these differences seem to become smaller throughout the years. The first study forecast an average baseload price of 50 euro's per MWh for 2025. Where the second study forecasts an average peakload price of 59.95 euro's per MWh, roughly 10 euro's per MWh more. The average peakload price is forecasted at 41.55 euro's per MWh according to the first study. The second study forecasts roughly 25 euro's per MWh more with 64.51 euro's per MWh.

These prices all decrease for 2030 according to the experts. The first study forecast an average baseload price of 45 euro's per MWh for 2030. Where the second study forecasts an average peakload price of 50.57 euro's per MWh, roughly 5 euro's per MWh more. The average peakload price is forecasted at 41.56 euro's per MWh according to the first study. The second study forecasts roughly 7.50 euro's per MWh more with 48.94 euro's per MWh.

Finally, the prices decrease again for 2035 according to all the experts. However, this decrease is smaller. The first study forecast an average baseload price of 45 euro's per MWh for 2035. Where the second study forecasts an average peakload price of 47.02 euro's per MWh, roughly 2 euro's per MWh more. The average peakload price is forecasted at 36.87 euro's per MWh according to the first study. The second study forecasts roughly 11 euro's per MWh more with 47.69 euro's per MWh.

We recommend a comparison study of current forecast models with the structured expert judgement model. It is important to use forecast models that provide confidence intervals. The forecasts provided by the various forecast models can then be used to calculate performance scores, i.e. calibration score, information score and combined score. These measures are in turn used to compare performance of the different models and the experts.

Finally, we recommend combined forecasting. That is, we suggest to combine data driven forecast method with structured expert judgement. According to [6] and [9], combined forecasts outperform forecasts obtained by single models. It would therefore be interesting to apply this in the field of structured expert judgement.

Bibliography

[1] Elektriciteitsproductie stijgt in 2020 naar recordhoogte, 2021. URL https://www.cbs.nl/nl-nl/nieuws/2021/09/elektriciteitsproductie-stijgt-in-2020-naar-recordhoogte.

[2] Energieverbruik met 3 procent gedaald in 2020, 2021. URL https://www.cbs.nl/nl-nl/nieuws/2021/14/energieverbruik-met-3-procent-gedaald-in-2020.

[3] Epexspot, 2021. URL https://www.epexspot.com/en.

[4] M. Afman, S. Hers, and T. Scholten. Energy and electricity price scenarios 2020-2023-2030. Technical report, CE Delft, 2017.

[5] International Energy Agency. The netherlands 2020. Technical report, International Energy Agency, 2020.

[6] R.R. Andrawis, A.F. Atiya, and H. El-Shishiny. Combination of long term and short term forecasts, with application to tourism demand forecasting. *International Journal of Forecasting*, 27:870–886, 2011. URL https://doi.org/10.1016/j.ijforecast.2010.05.019.

[7] W. Aspinall. Expert judgement elicitation using the classical model and excalibur. 2008. URL http://dutiosc.twi.tudelft.nl/~risk/extrafiles/EJcourse/Sheets/Aspinall%20Briefing%20Notes.pdf.

[8] H. Bahar, Y. Abdelilah, T. Criswell, P. Bojek, F. Briens, P. Le Feuvre, G. Rodrigeuz Jimenez, and et all. Renewables 2020. Technical report, International Energy Agency, 2020.

[9] S. Bordignon, D.W. Bunn, F. Lisi, and F. Nan. Combining day-ahead forecasts for british electricity prices. *Energy Economics*, 35:88–103, 2013. URL https://doi.org/10.1016/j.eneco.2011.12.001.

[10] CMS. Electricity law and regulation in the netherlands, 2021. URL https://cms.law/en/int/expert-guides/cms-expert-guide-to-electricity/netherlands.

[11] R.M. Cooke and L.H.J. Goossens. Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry*, 90:303–309, 2000.

[12] R.M. Cooke and L.H.J. Goossens. Tu delft expert judgment data base. *Reliability Engineering and System Safety*, 93:657–674, 2008. URL https://doi.org/10.1016/j.ress.2007.03.005.

[13] R.M. Cooke, L.H.J. Goossens, A.R. Hale, and Lj. Rodić-Wiersma. Fifteen years of expert judgement at tudelft. *Safety Science*, 46:234–244, 2008. URL https://doi.org/10.1016/j.ssci.2007.03.002.

[14] J. Crespo Cuaresma, J. Hlouskova, S. Kossmeier, and M. Obersteiner. Forecasting electricity spot-prices using linear univariate time-series models. *Applied Energy*, 77:87–106, 2004. URL https://doi.org/10.1016/S0306-2619(03)00096-5.

[15] R.A. de Marcos, A. Bello, and J. Reneses. Electricity price forecasting in the short term hybridising fundamental and econometric modelling. *Electric Power Systems Research*, 167:240–251, 2019. URL http://doi.org/10.1016/j.epsr.2018.10.034.

[16] R.A. de Marcos, A. Bello, and J. Reneses. Short-term electricity price forecasting with a composite fundamental-econometric hybrid methodology. *Energies*, 2019.

[17] K. Van den Bergh, J. Boury, and E. Delarue. The flow-based market coupling in central western europe. *The Electricity Journal*, 29:24–29, 2016. URL http://dx.doi.org/10.1016/j.tej.2015.12.004.

[18] L.C. Dias, A. Morton, and J. Quigley. *Elicitation*, volume 261. Springer, 2018.

[19] S. Fleten and J. Lemming. Constructing forward price curves in electricity markets. *Energy Economics*, 25:409–424, 2003. URL www.elsevier.com/locate/eneco.

[20] DNV GL. Energy transition outlook 2020. Technical report, DNV GL, 2020.

[21] DNV GL. Transition faster together. Technical report, DNV GL, 2020.

[22] A.M. Hanea, G.F. Nane, T. Bedford, and S. French. *Expert Judgement in Risk and Decision Analysis*, volume 1. Springer Cham, 2021.

[23] D. Hartley and S. French. A bayesian method for calibration and aggregation of expert judgement. *International Journal of Approximate Reasoning*, 130:192–225, 2021. URL http://doi.org/10.1016/j.ijar.2020.12.007.

[24] A. Heydari, M.M. Nezhad, E. Pirshayan, D.A. Garcia, F. Keynia, and L. De Santoli. Short-term electricit price and load forecasting in isolated power grids based on composite neural network and gravitational search optimization algorithm. *Applied Energy*, 277, 2020. URL http://doi.org/10.1016/j.apenergy.2020.115503.

[25] The International Energy Agency (IAE). Countries regions, 2021. URL https://www.iea.org/countries/the-netherlands.

[26] N.V. Karakatsani and D.W. Bunn. Intra-day and regime-switching dynamics in electricity price formation. *Energy Economics*, 30:1776–1797, 2008. URL http://doi.org/10.1016/j.eneco.2008.02.004.

[27] J. Klooster, R. Schillemans, and G. Warringa. Vrije stroom, vieze stroom, weg stroom? Technical report, CE Delft, 2005. URL https://ce.nl/publicaties/vrije-stroom-vieze-stroom-weg-stroom/.

[28] T. Kristiansen. The flow based market coupling arrangement in europe:. *Energy Strategy Reviews*, 27, 2019. URL https://doi.org/10.1016/j.esr.2019.100444.

[29] J. Lago, F. De Ridder, and B. De Schutter. Forecasting spot electricity prices:. *Applied Energy*, 221:386–405, 2018. URL https://doi.org/10.1016/j.apenergy.2018.02.069.

[30] K. Leung, S. Verga, and CSS OR Team. Expert judgement in risk assessment. 2007. URL https://cradpdf.drdc-rddc.gc.ca/PDFS/unc88/p529083.pdf.

[31] Lighttwist. Excalibur, 1989-2013. URL https://www.expertsinuncertainty.net/Publicationscasestudies/Excalibur/tabid/4386/Default.aspx.

[32] K. Maciejowska, J. Nowotarski, and R. Weron. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *International Journal of Forecasting*, 32:957–965, 2016. URL http://dx.doi.org/10.1016/j.ijforecast.2014.12.004.

[33] T.A. Mazzuchi and J.R. van Dorp. A bayesian expert judgement model to determine lifetime distributions for maintenance optimisation. *Structure and Infrastructure Engineering*, 8:307–315, 2012. URL https://doi.org/10.1080/15732479.2011.563084.

[34] A. Misiorek, S. Trueck, and R. Weron. Point and interval forecasting of spot electricity prices. *Studies in Nonlinear Dynamics & Econometrics*, 10, 2006.

[35] J. Moccia, A. Arapogianni, Wind Energy Association, J. Wilkes, C. Kjaer, and R. Gruet. Pure power. Technical report, European Wind Energy Association, 2011.

[36] MOOC. Decision making under uncertainty, 2020. URL https://ocw.tudelft.nl/courses/decision-making-under-uncertainty-introduction-to-structured-expert-judgment/.

[37] J. Munkhammar, D. van der Meer, and J. Widén. Very short term load forecasting of residential electricity consumption using the markov-chain mixture distribution (mcm) model. *Applied Energy*, 282, 2021. URL http://doi.org/10.1016/j.apenergy.2020.116180.

[38] J. Nowotarski and R. Weron. Recent advances in electricity price forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018. URL http://doi.org/10.1016/j.rser.2017.05.234.

[39] Ministry of Economic Affairs of the Netherlands. Energy report. Technical report, Ministry of Economic Affairs of the Netherlands, 2020.

[40] O. Ozdemir, M. Scheepers, and A. Seebregts. Future electricity prices. Technical report, ECN, 2008.

[41] H.P. Stumpf and B. Hu. Offshore wind access 2018. Technical report, ECN, 2018.

[42] Tennet. Tennet. URL https://www.tennet.eu/nl/#&panel1-2.

[43] Tennet. Infrastructure outlook 2050. Technical report, Gasunie, 2020.

[44] Tennet. Visie2030. Technical report, Tennet, 2020.

[45] C. Werner, T Bedford, R.M. Cooke, and O. Morales-Nápoles A.M. Hanea. Expert judgement for dependence in probabilistic modelling. *European Journal of Operational Research*, 258:801–819, 2017. URL http://dx.doi.org/10.1016/j.ejor.2016.10.018.

[46] R. Wiser, J. Rand, J. Seel, P. Beiter, E. Baker, E. Lantz, and P. Gilman. Expert elicitation survey predicts 37% to 49% declines in wind energy costs by 2050. *Nature Energy*, 6:555–565, 2021. URL https://doi.org/10.1038/s41560-021-00810-z.

[47] F. Ziel and R. Steinert. Electricity price forecasting using sale and purchase curves. *Energy Economics*, 59:435–454, 2016. URL http://doi.org/10.1016/j.eneco.2016.08.008.

A

Elicitation document

Motivation

The gradual deregulation process of the European electricity market, starting in the early 1990, resulted in electricity trade under market rules using spot and derivatives contracts. Due to the economically non-storable nature of the commodity that is electricity, the constant balance between consumption and production, weather effects, e.g. temperature, wind speed, solar intensity etc, and the intensity of everyday and business activities, e.g. holidays, weekends, on- and off-peak hours etc., the price dynamics of this commodity are quite unique as they are not observed in any other market. This extreme price volatility has forced market participants to hedge volumes as well as price risks. Naturally, electricity price forecast models are of great interest to portfolio managers.

Over the last three decades, research in electricity price modeling and forecasting has propelled, resulting in various forecast models. These can broadly be divided into six classes, i.e. cost-based models, game theoretic approaches, fundamental methods, econometric models, statistical approaches and artificial intelligence-based techniques.

However, the scientific uncertainty of these models to this date is substantial. Moreover, the slightest decrease in the mean absolute percentage error of the price forecasts can result in a saving of hundreds of thousands of euros. Decreasing the uncertainty in current models and developing new models with a lower uncertainty is an ongoing process. In the meantime, current models are combined with expertise of forecasters. We aim to subject this process to transparent methodology.

Structured Expert Judgement

Structured expert judgement attempts to subject the process of soliciting expert advice to transparent methodological rules. The goal is to treat expert judgments as scientific data in a formal decision process. The scientific method is the process by which experts come to agree. Broadly speaking, Structured Expert Judgement is a tool for analysing and predicting certain variables of interest by aggregating and evaluating seed variables provided by a (manually selected) pool of experts.

A lot of work has gone into translating the scientific method into a workable procedure which gives good results in practice. The *Classical Model* by Cooke, therefore sometimes mentioned as Cooke's method, embodies this procedure. The goal of the model is the sensible aggregation of experts' assessments of uncertainty. The model uses two scores to asses experts' performance in quantifying uncertainty, i.e. the calibration score and the information score. These scores are combined and in turn yield weights for a performance-based aggregation of experts' opinion, i.e. the decision maker.

The calibration score measures the statistical likelihood that a set of experimental results correspond with the expert's assessments. That is, it measures an experts statistical accuracy. The information score measures the degree to which an expert's uncertainty distribution is concentrated. A good probability assessor is informative, thus has a high information score, by providing narrow assessments and at the same time is statistically accurate, thus having a high calibration score, by capturing the true values with the long run correct relative frequencies. That is, 5% of the answers should be near the lower bound of the assessments, 5% of the answers should be near the upper bound of the assessments and 90% of the answers should fall in between these. The next part will clarify this concept.

Finally, statistical accuracy is more important than informativeness. That is, a high calibration score is preferred over an high information score. Assessments that are highly informative but statically inaccurate are not useful. Non-informative but statistically accurate assessments are useful. They teach us how large the uncertainties may be.

Task description

Suppose you are presented with an uncertainty quantity:

What will be the average electricity day-ahead baseload spot price on the EPEX spot market on 23 February 2021?

5%	50%	95%

You are asked to quantify your uncertainty by specifying percentiles of your subjective uncertainty. That is, you provide your **best guest** as the 50%-tile. The 5%-tile represents your **lower bound**, according to you, the true value will be lower than your lower bound with a probability of 5%. The 95%-tile represents your **upper bound**, according to you, the true value will be larger than your upper bound with a probability of 5%. You are also asked to provide rationales, e.g. data, assumptions, scenarios, that informed your uncertainty for each question.

Suppose you give the following assessment:

What will be the average electricity day-ahead baseload spot price on the EPEX spot market on 23 February 2021?

5%	50%	95%
44.32	65.4	73.81

Then you've stated that your *best guess* equals 65.40 euros per MWh and the price, according to you, is very unlikely to exceed 73.81 euros per MWh or drop down 44.31 euros per MWh or lower.

Remarks

Statements such as *I am not the best expert for that.* are not uncommon. However, often turn to be untrue. The people with the most detailed knowledge are not always the best at quantifying their uncertainty. *Knowing little* about the subject results in a broader assessment, that is a broader uncertainty interval. This is still more informative than a narrow uncertainty interval with no statistical accuracy. Experts with (high) statistical accuracy and more informative assessments, will automatically exert more influence on the decision maker. However, in cases of experts with no statistical accuracy and more informative assessments, uninformative assessments, i.e. that of those who *know little*, will accurately depict the uncertainty.

Spot prices are determined in weekends as well. If you are unable to work on your assessment during the weekend, we advise you to assess Saturday and Sunday on the Friday in advance. The calibration question needs to be answered before the realisation is published. However, it is strongly advised to forecast on a daily basis.

Collaboration is not allowed during the elicitation. Your assessments should reflect your own expertise and uncertainty. Discussing with other experts can influence your assessment.

The qualitative rationales, the information such as data, assumptions scenarios, mentioned in the task description will become part of the published record of the study as they help us understand the results of the study. They might also be compared with the qualitative rationales of the other experts to illuminate differences in the assessments. However,the rationales, as well as the assessments will not be linked with the experts. This ensures the anonymity of expert opinion.

The realisations will be published in the group app on a daily basis, on request of the group. This way the experts will be able to evaluate their previous assessment(s). The expert with the highest combination of the calibration score and the information score receives a small prize. And all questions concern the Netherlands.

Calibration Questions

Between 1 March and 1 May we will ask the same two questions each day. The date changes with one day each day. So on the first of March you are presented two calibration questions:

- What will be the average electricity day-ahead baseload spot price on the EPEX spot market on 2 March 2021?
- What will be the average electricity day-ahead peakload spot price on the EPEX spot market on 2 March 2021?

On the second of March we change 2 March to 3 March in the question etc. The prices are given in euros per Mwh. Peakloads refer to 08:00 - 20:00. Baseloads refer to a 24-hour time period.

The assessments are given each day before *10:00*. These assessments concern the prices of the next day. That is, the assessments provided on the first day of March before 10:00 concern the average electricity day-ahead baseload/peakload spot price on the EPEX spot market on 2 March 2021. The assessments given between 1-3-2021 10:00 and 2-3-2021 10:00 concern the average electricity day-ahead baseload/peakload spot price on the EPEX spot market on 3 March 2021.

Questions of Interest

The questions of interest will somewhat resemble the calibration questions. The prices are given in euros per MWh. Note that, instead of the average price on a day-level, we are interested in the average price on a year level. **The questions of interest need to be answered before *1 May 2021 23:59***. Please follow the same conventions and sources of data as for the calibration questions.

- What will be the average electricity baseload spot price on the EPEX spot market in 2025?
- What will be the average electricity peakload spot price on the EPEX spot market in 2025?
- What will be the average electricity baseload spot price on the EPEX spot market in 2030?
- What will be the average electricity peakload spot price on the EPEX spot market in 2030?
- What will be the average electricity baseload spot price on the EPEX spot market in 2035?
- What will be the average electricity peakload spot price on the EPEX spot market in 2035?

B

Case study participation invitation

Dear potential study participant,

We invite you to participate in a research study conducted by Ashni Bachasingh, MSc student in TU Delft's Mathematics program. The research focuses on electricity spot prices and your domain expertise is highly valuable. The research is part of the MSc project: A Structured Expert Judgment study to forecast electricity spot prices.

Due to the economically non-storable nature of the commodity that is electricity, electricity price forecast models are of great interest to portfolio managers and traders. Unsurprisingly, due to the (gradual) deregulation process of the European electricity market starting in the early 1990s, research in electricity price modelling and forecasting has propelled. Resulting in various forecast models. However, the scientific uncertainty of these models to this date is substantial. Moreover, the slightest decrease in the mean absolute percentage error of the price forecasts can result in a saving of hundreds of thousands euros. Decreasing the uncertainty in current models and developing new, more accurate, models is an ongoing process. In the meantime, current models are combined with expertise of energy market professionals. We aim to subject this process to a transparent methodology.

Structured expert judgement attempts to subject the process of soliciting expert advice to a transparent, traceable and validated methodology. The goal is to treat expert judgments as scientific data in a formal decision process. The scientific method is the process by which experts come to agree. Broadly speaking, Structured Expert Judgement is a tool for analysing and predicting certain variables of interest by aggregating and evaluating seed variables provided by a (manually selected) pool of experts.

During the study, you are asked to answer questions concerning the Dutch electricity market. That is, questions concerning energy conservation, energy demand developments, the development of the emission prices, capacity cross-boarder, the energy mix and its development, flow based market coupling etc. It will take about 30 minutes to answers the questions.

The questions are asked via the platform Minerva, created by Excogent. To receive the questions, i.e. use the platform, we need to use your e-mail address. This is the only piece of personal information needed for this research. The platform is available through Amazon Web Services cloud provider. The platform follows the standards of data privacy (GDPR) and protection. At the end of the project, after downloading the data, all the answers will be deleted from the platform.

Any information you provide will be kept anonymous. The researcher will not use your personal information for any purposes outside of this research project. Also, the researcher will not include your name or anything else that could identify you in the study reports or the platform.

Your participation will be a valuable addition to our research and findings could lead to (new) models with lower scientific uncertainty. If interested, the research and findings will be send to you in the form of the

final master project report. Additionally, you will be informed of any manuscript that will be submitted for publication.

If you would like to participate in the study please reply to this e-mail providing us with a written consent for taking part in the study and the e-mail address we can use for the study. If you have any questions please do not hesitate to ask.

Please feel free to forward this invitation to other experts in your network. Your help is highly appreciated.

Thank you for your time and participation,

Sincerily,

Ashni Bachasingh
Master Student, Technical University Delft

C

Assessment graphs first elicitation

Figure C.1: Assessments of expert 1. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asesements and the peakload assessments.

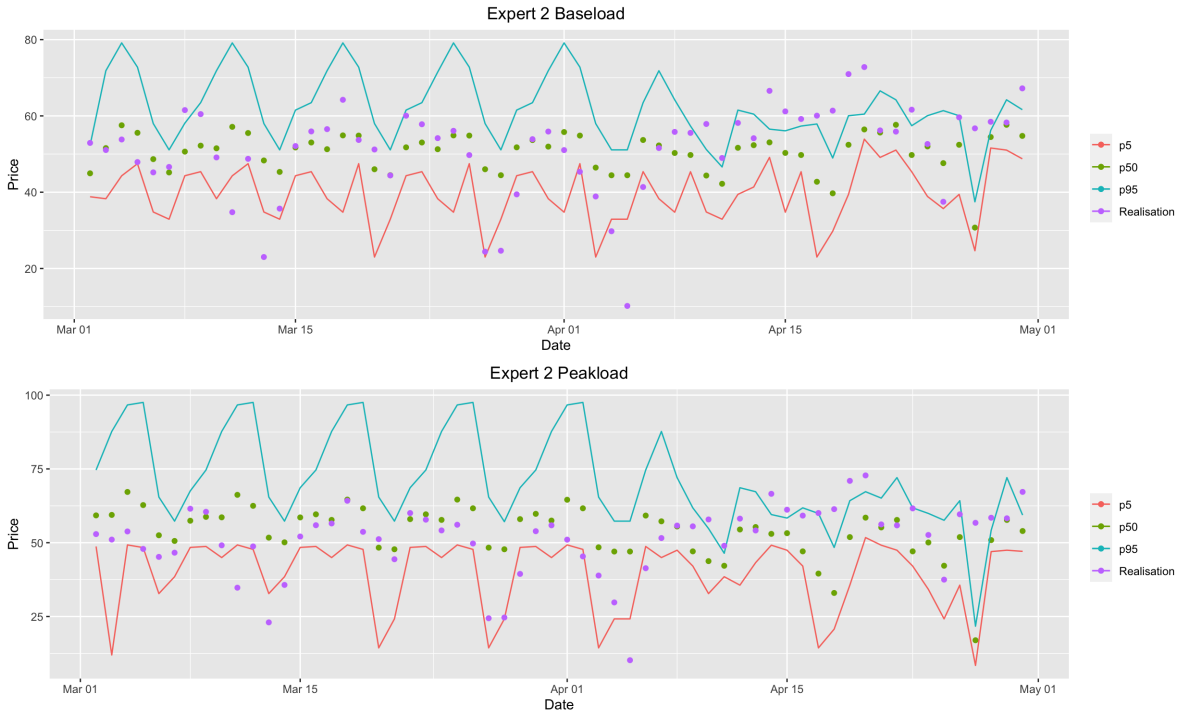

Figure C.2: Assessments of expert 2. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

Figure C.3: Assessments of expert 3. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

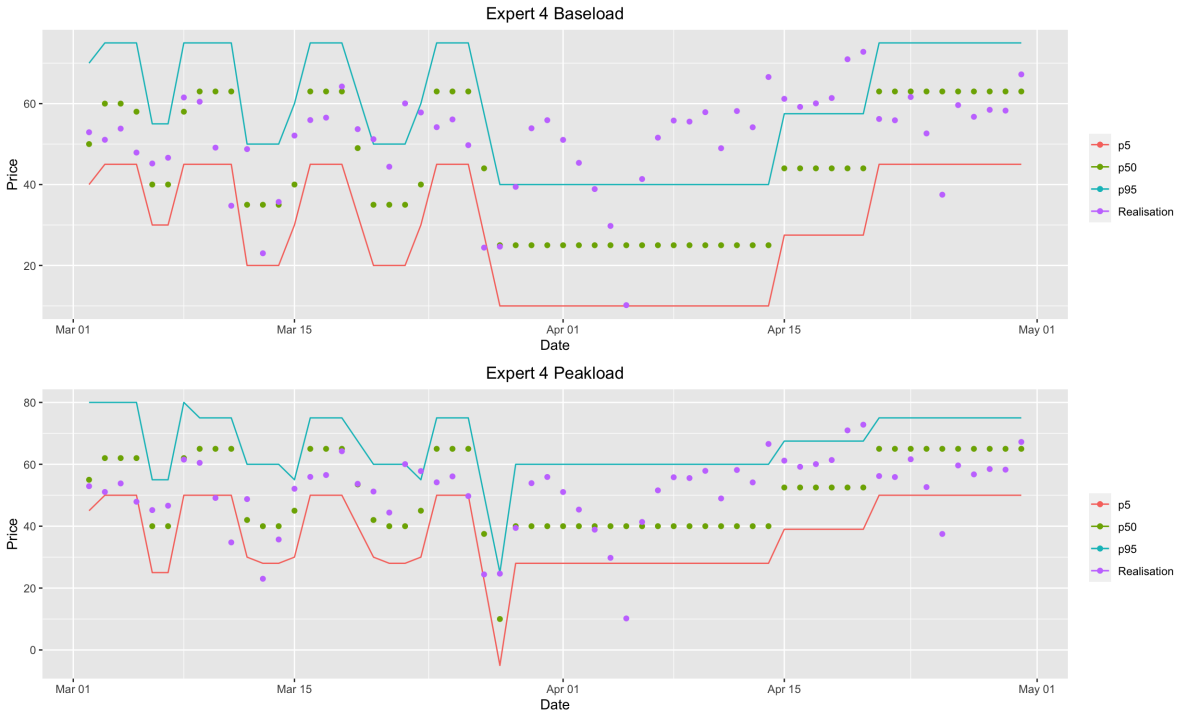

Figure C.4: Assessments of expert 4. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

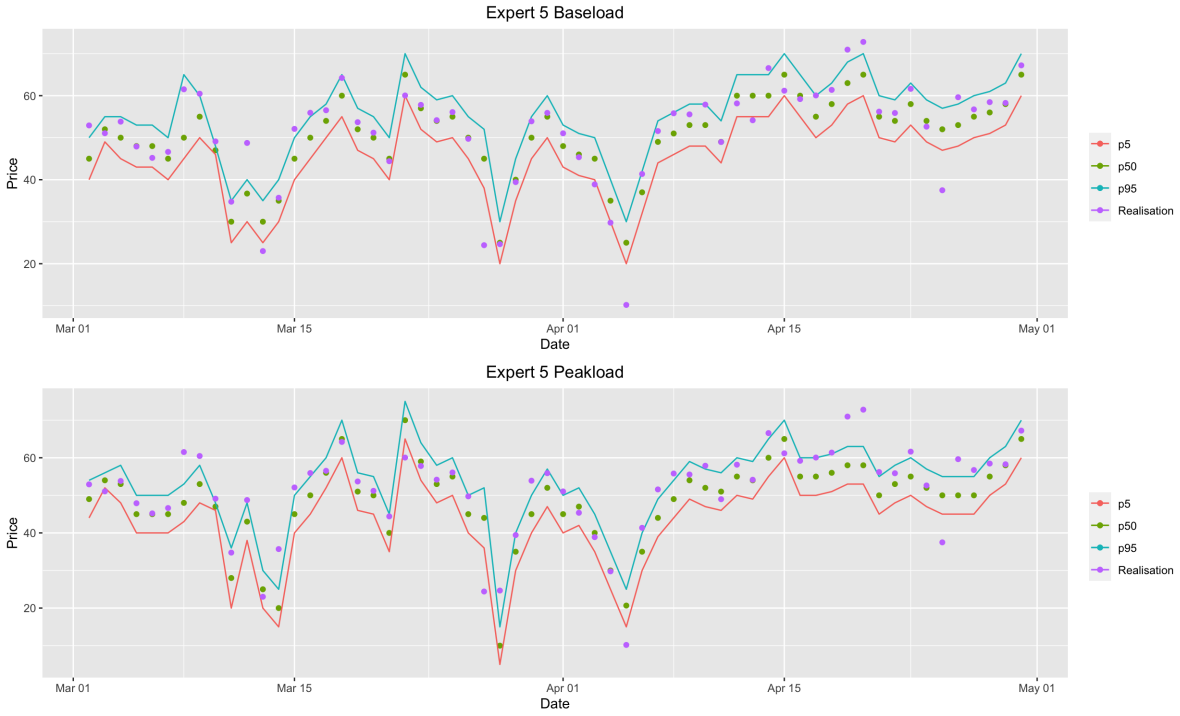

Figure C.5: Assessments of expert 5. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

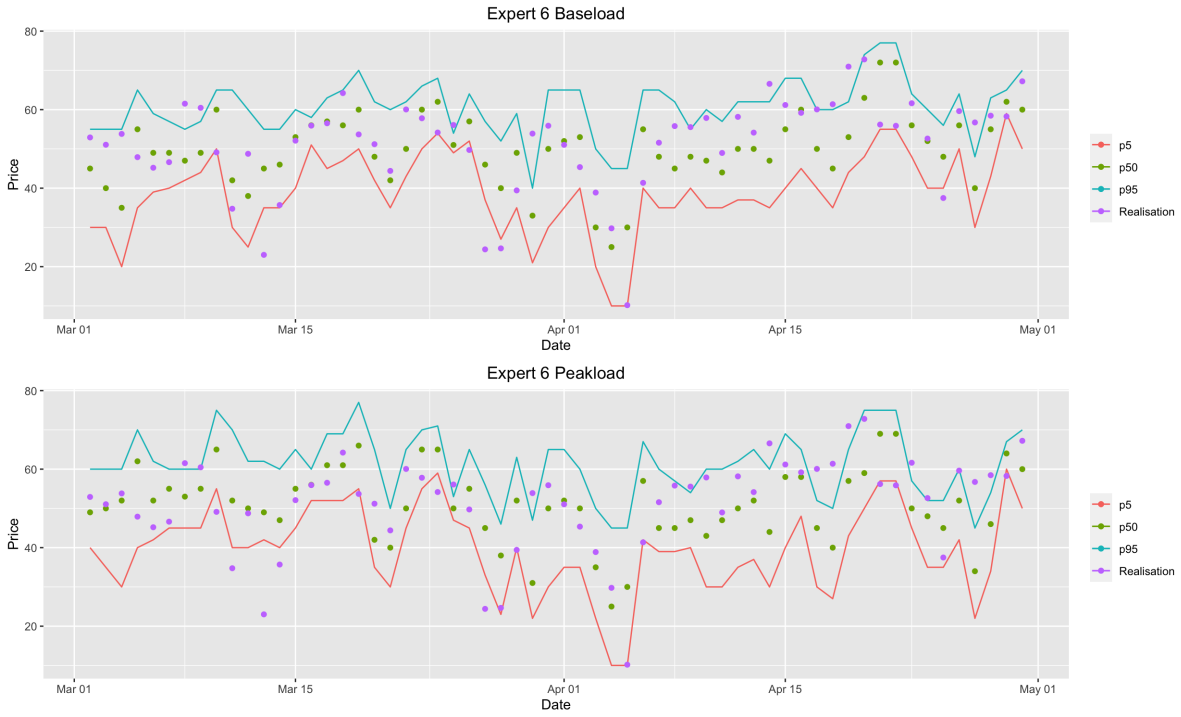

Figure C.6: Assessments of expert 6. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

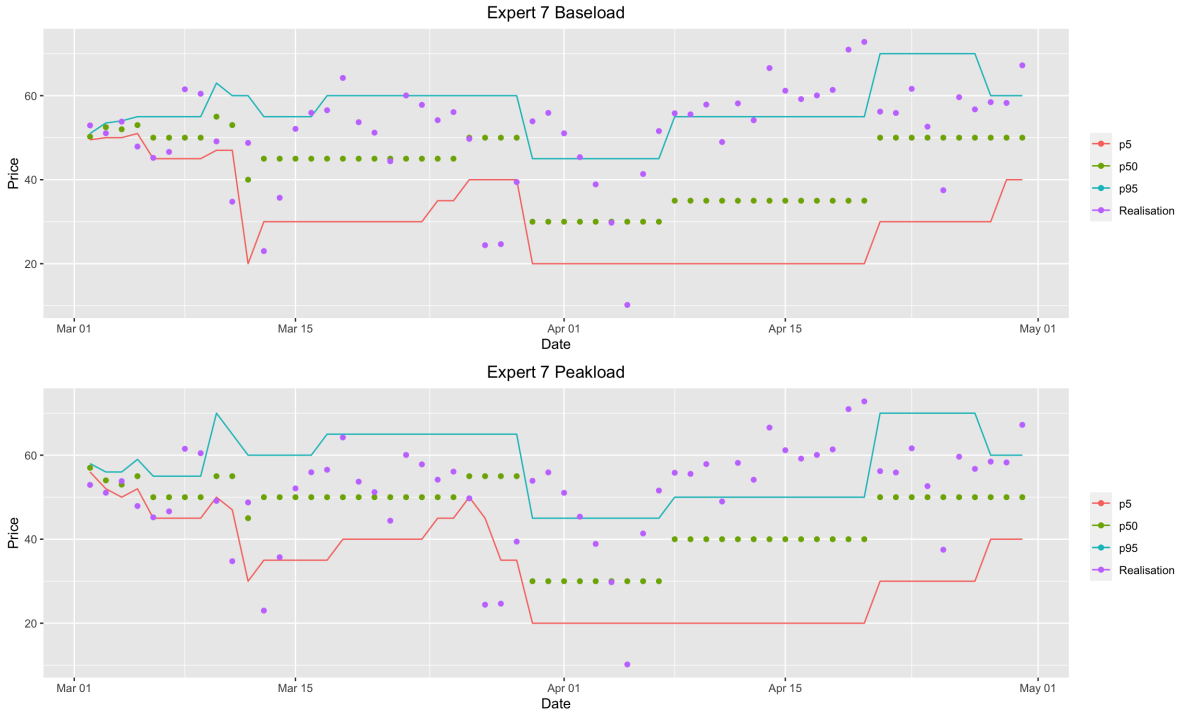

Figure C.7: Assessments of expert 7. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

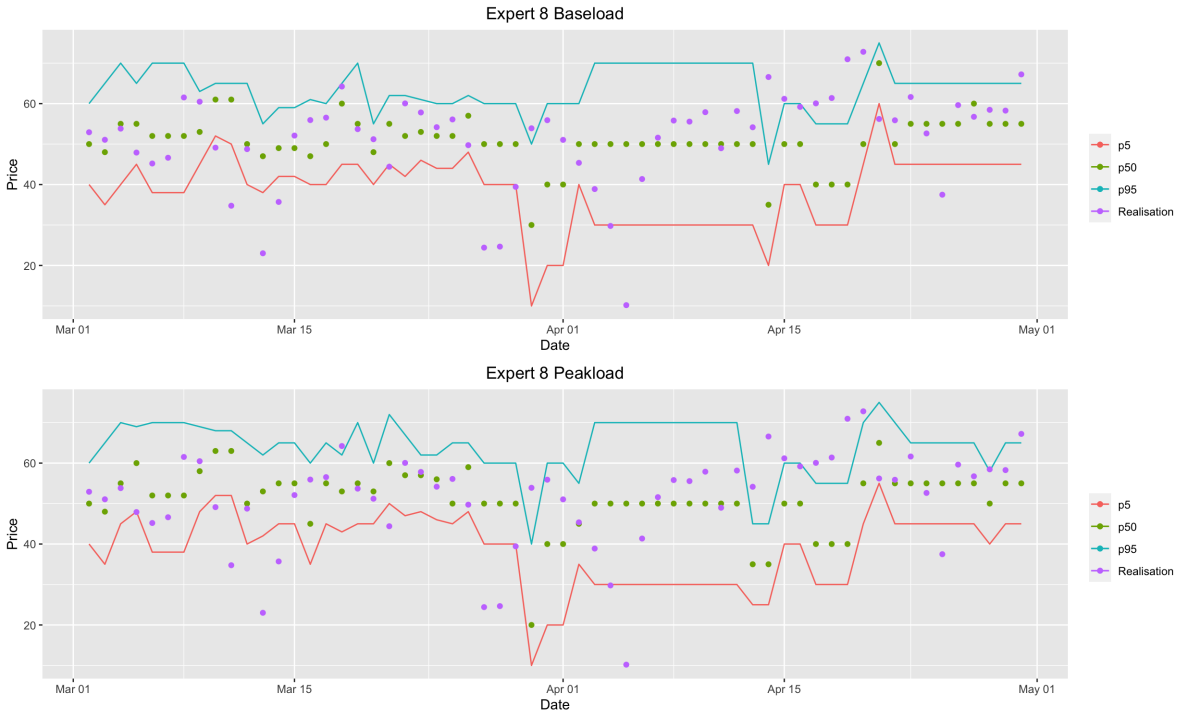

Figure C.8: Assessments of expert 8. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asesements and the peakload assessments.

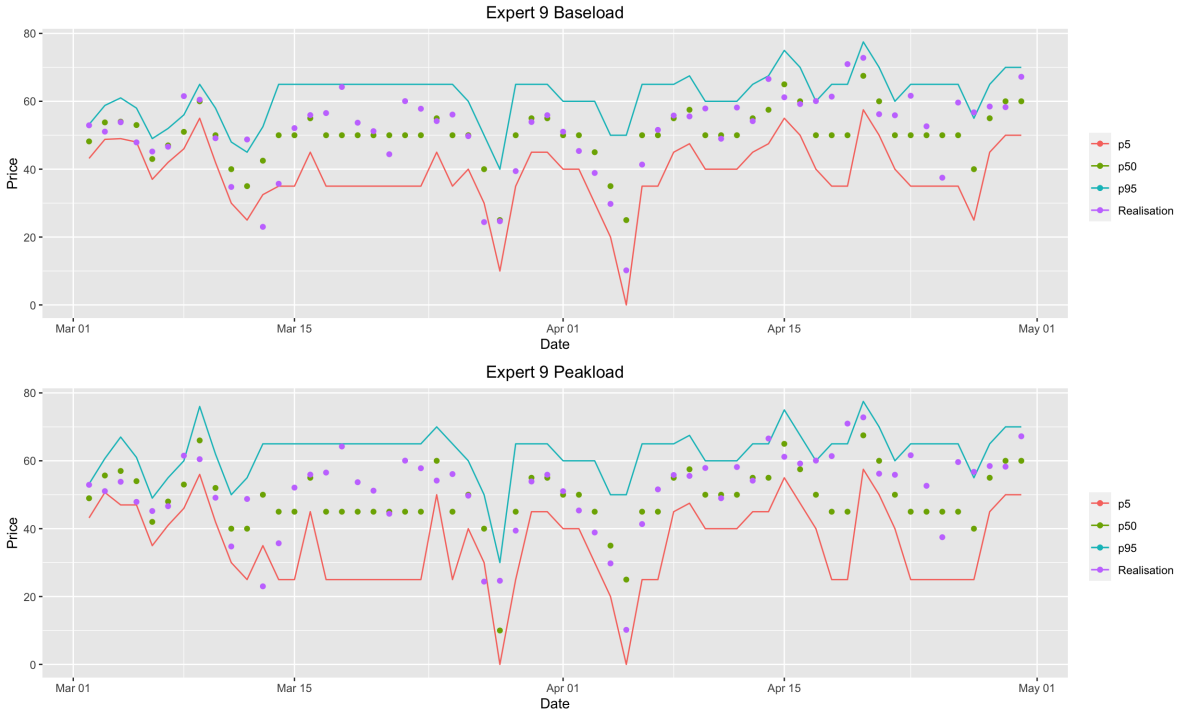

Figure C.9: Assessments of expert 9. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asesements and the peakload assessments.

Figure C.10: Assessments of expert 10. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

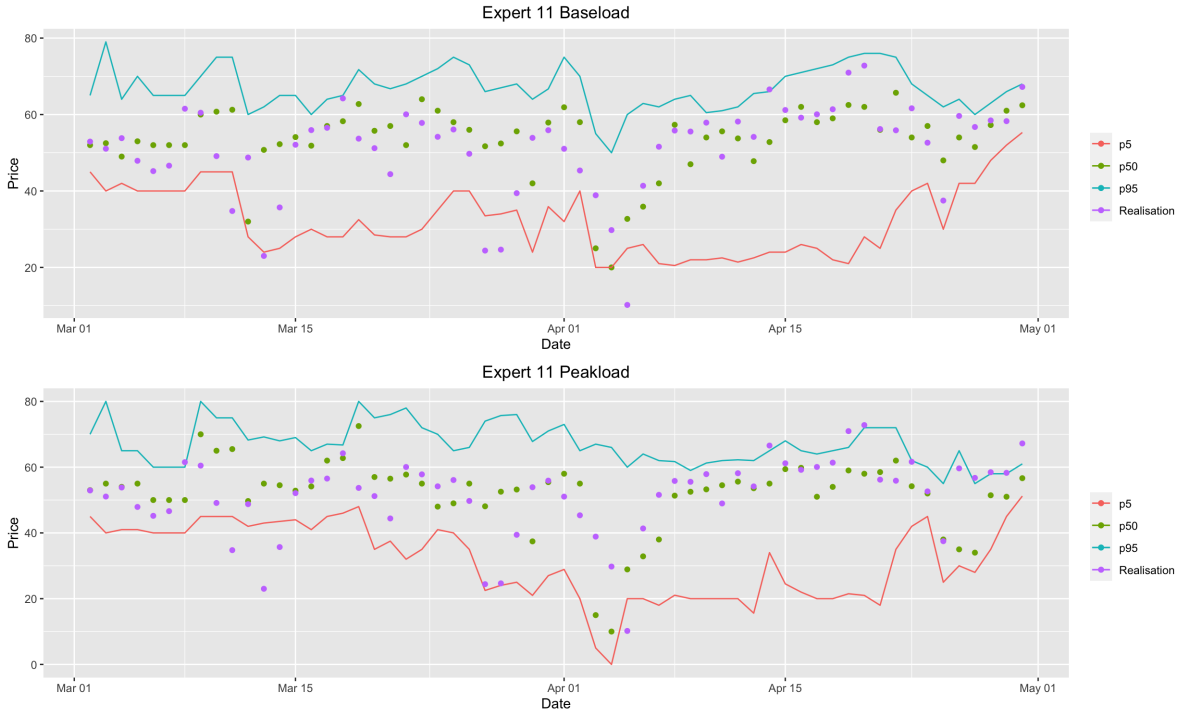

Figure C.11: Assessments of expert 11. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

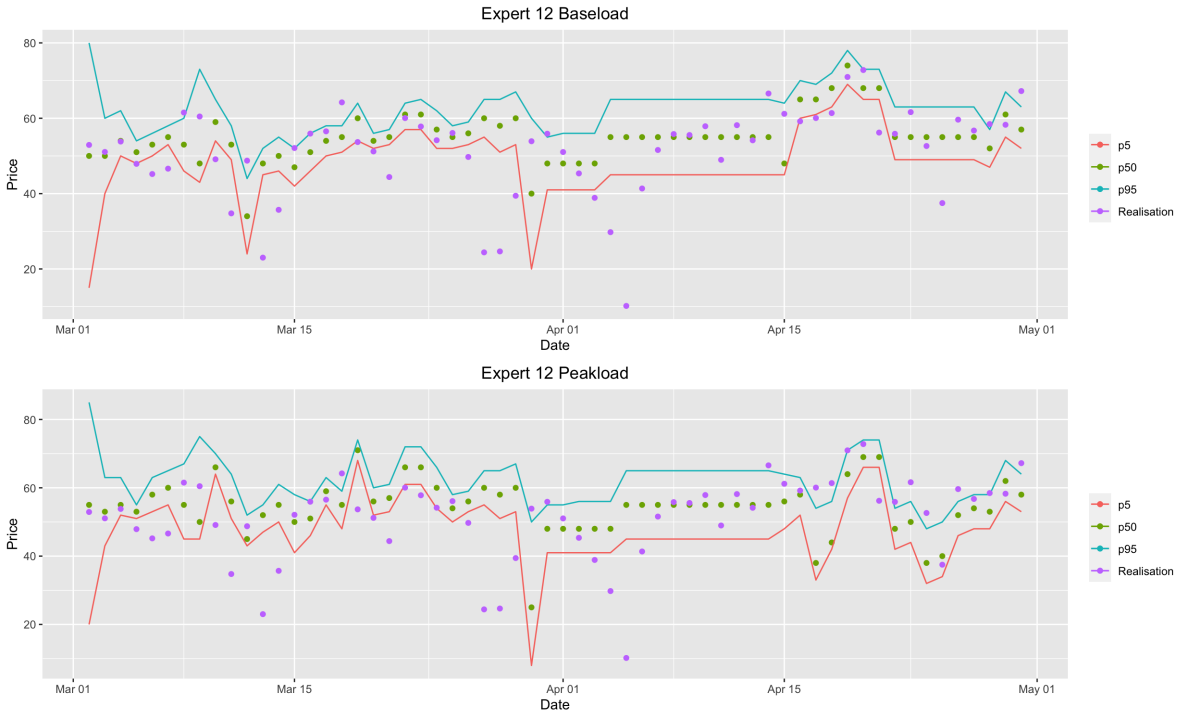

Figure C.12: Assessments of expert 12. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

Figure C.13: Assessments of expert 13. The red line represents the 5th percentile assessments, the blue line the 95th percentile assessments and the green dots the 50th percentile assessments. The purple points represent the realisations. We distinguish between the baseload asessements and the peakload assessments.

D

Expert performance plots first elicitation

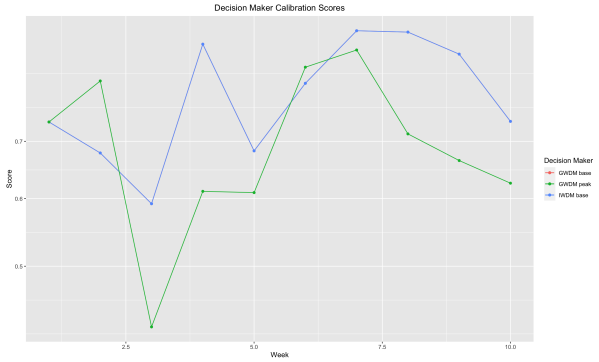

Figure D.1: Calibration scores of the decision makers throughout the weeks. We consider the full weeks here. We aggregate the weeks.

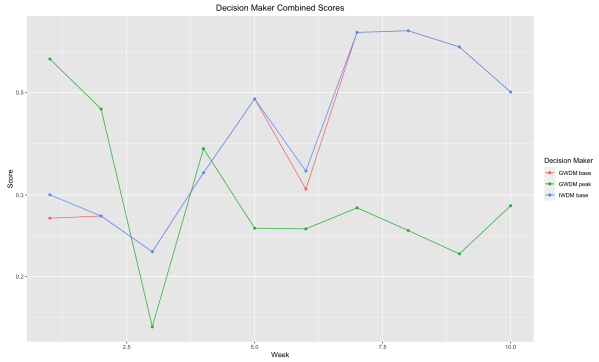

Figure D.2: Combined scores of the decision makers throughout the weeks. We consider the full weeks here. We aggregate the weeks.

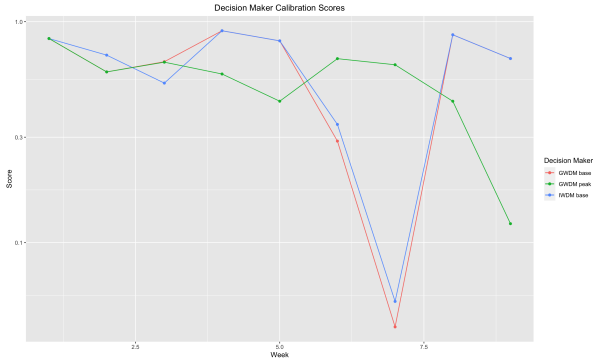

Figure D.3: Calibration scores of the decision makers throughout the weeks. We consider work-weeks here. We aggregate the weeks.

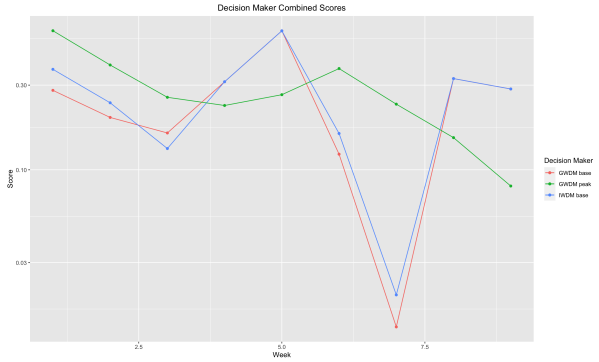

Figure D.4: Combined scores of the decision makers throughout the weeks. We consider work-weeks here. We aggregate the weeks.

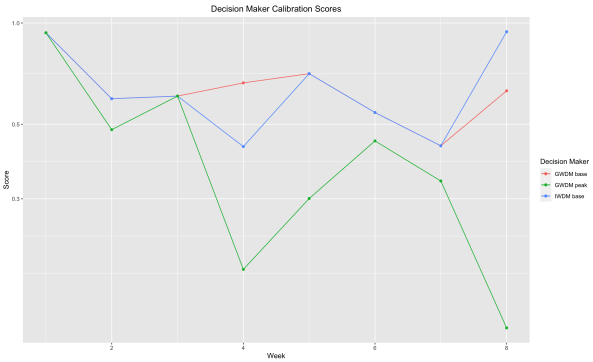

Figure D.5: Calibration scores of the decision makers throughout the weeks. We consider weekends here. We aggregate the weeks.

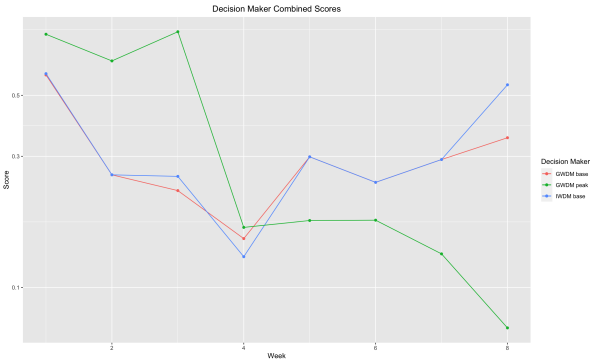

Figure D.6: Combined scores of the decision makers throughout the weeks. We consider weekends here. We aggregate the weeks.

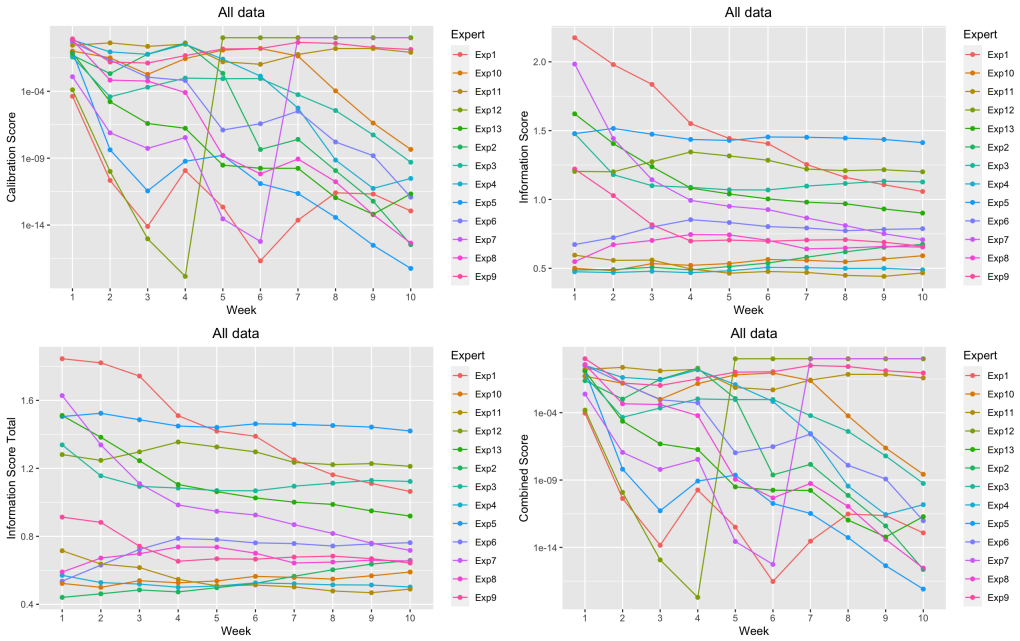

Figure D.7: Expert performance on a weekly basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. Weeks are aggregated, i.e. we keep adding 1 week to the data.

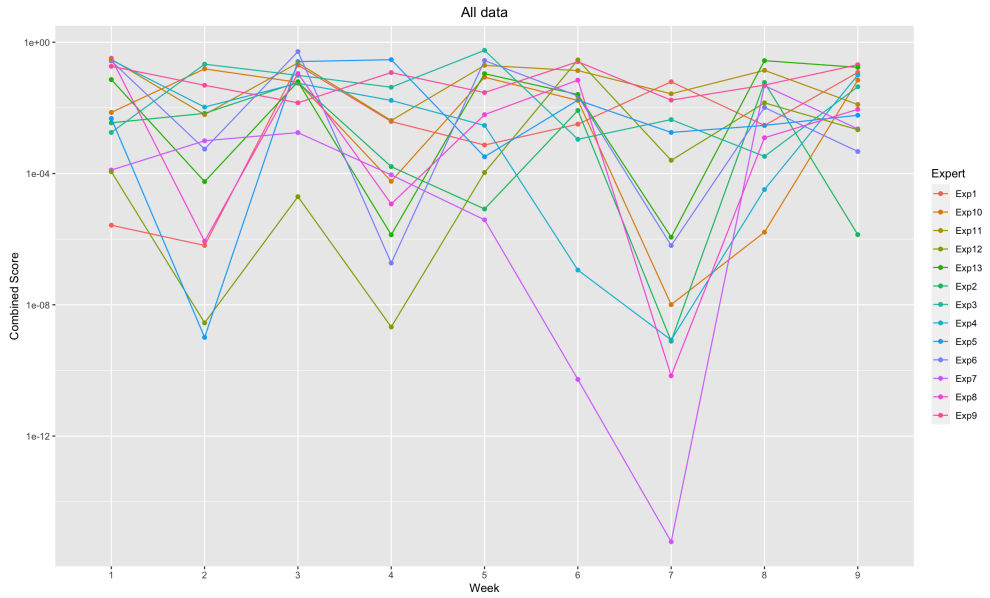

Figure D.8: Expert performance on a weekly basis. Presented are the combined scores. Weeks are not aggregated.

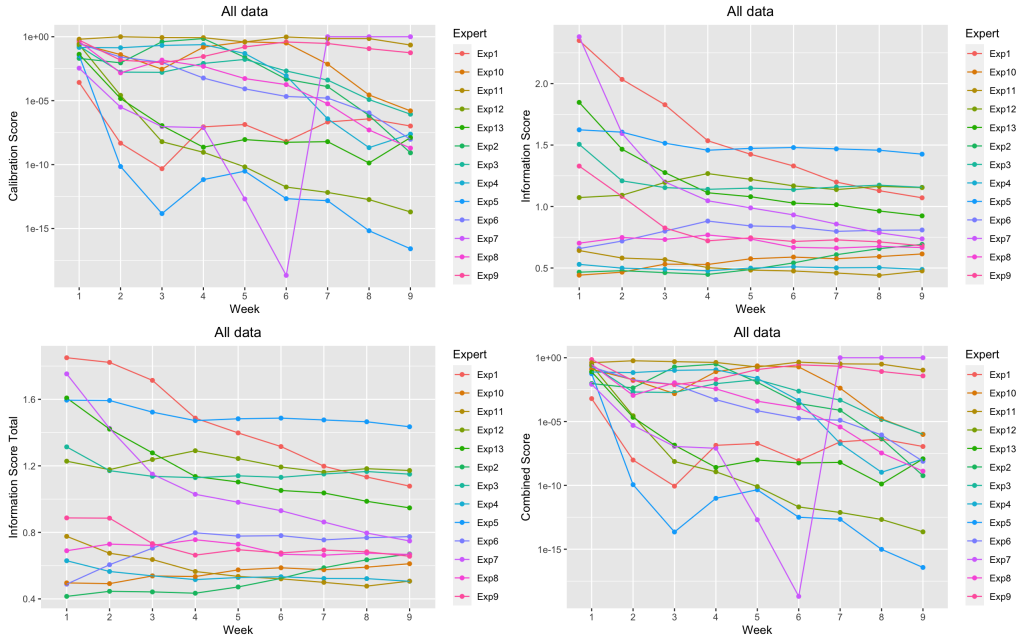

Figure D.9: Expert performance on a work-week basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. Weeks are aggregated, i.e. we keep adding 1 week to the data.

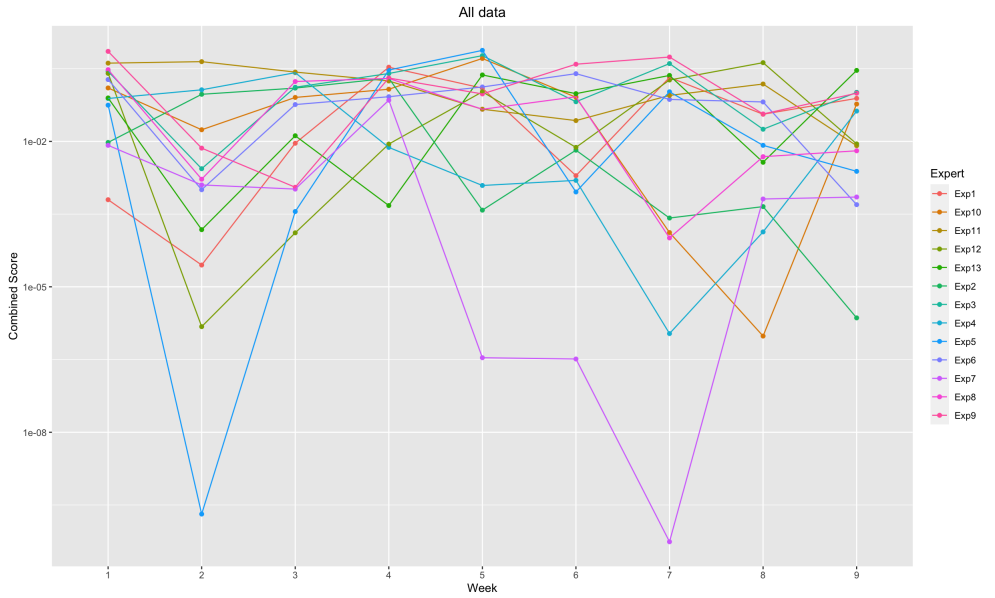

Figure D.10: Expert performance on a work-week basis. Presented are the combined scores. Weeks are not aggregated.

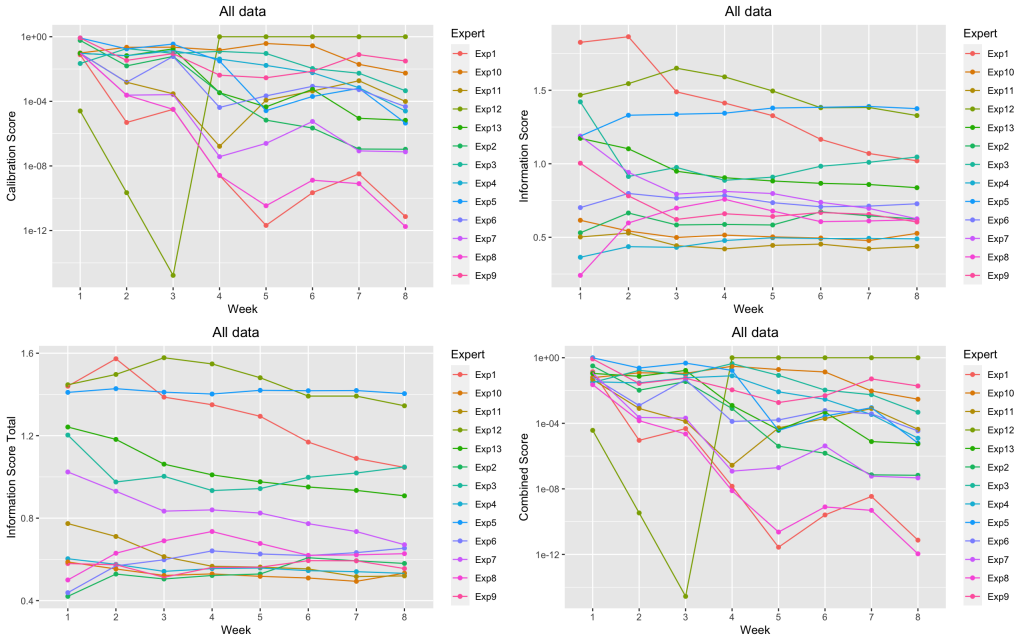

Figure D.11: Expert performance on a weekend basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. Weeks are aggregated, i.e. we keep adding 1 week to the data.

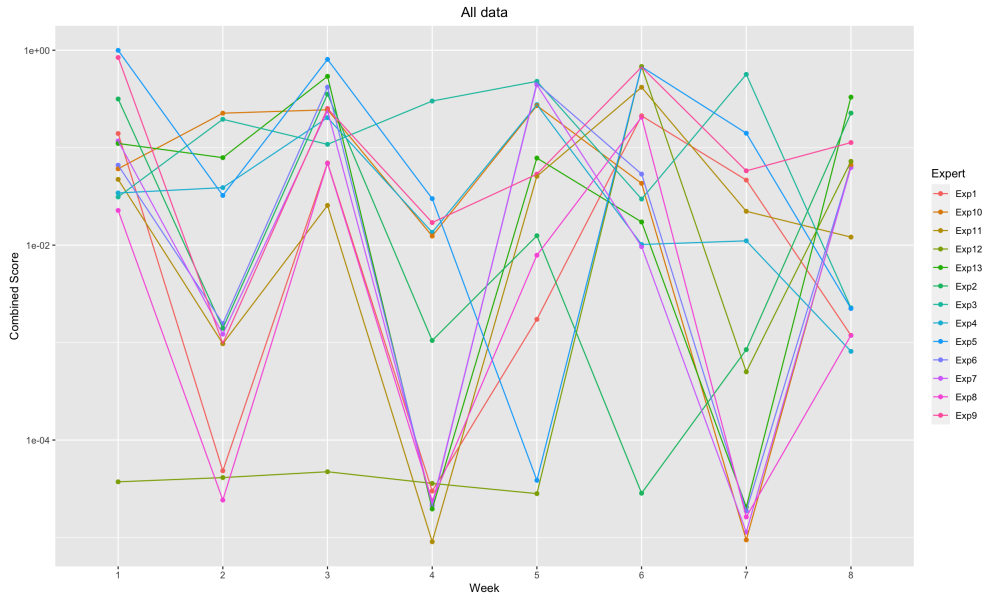

Figure D.12: Expert performance on a weekend basis. Presented are the combined scores. Weeks are not aggregated.

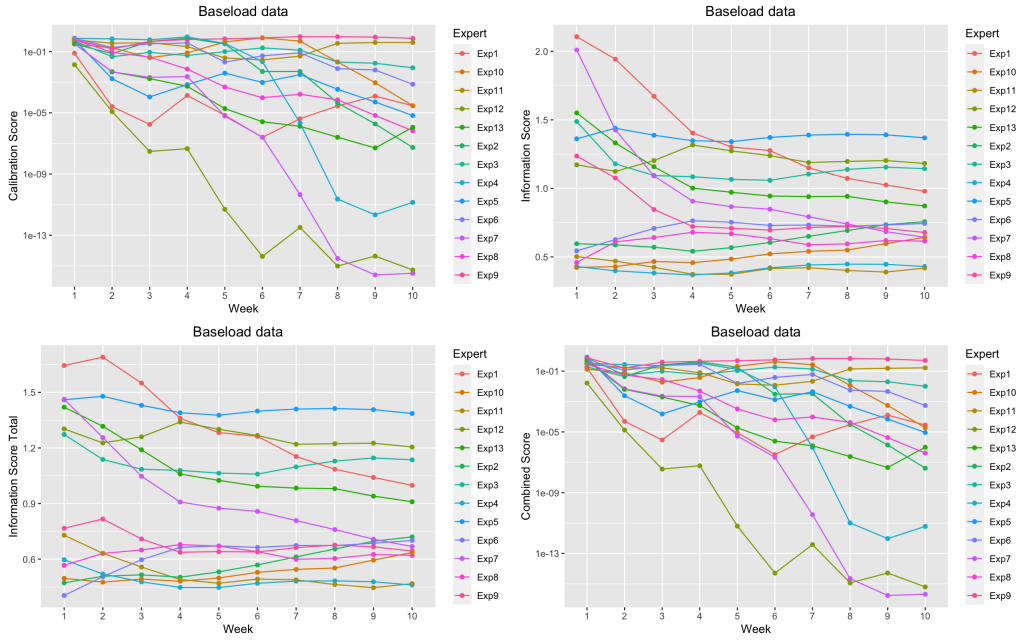

Figure D.13: Expert performance on a week basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. The scores are calculated for the baseload prices. Weeks are aggregated, i.e. we keep adding 1 week to the data.

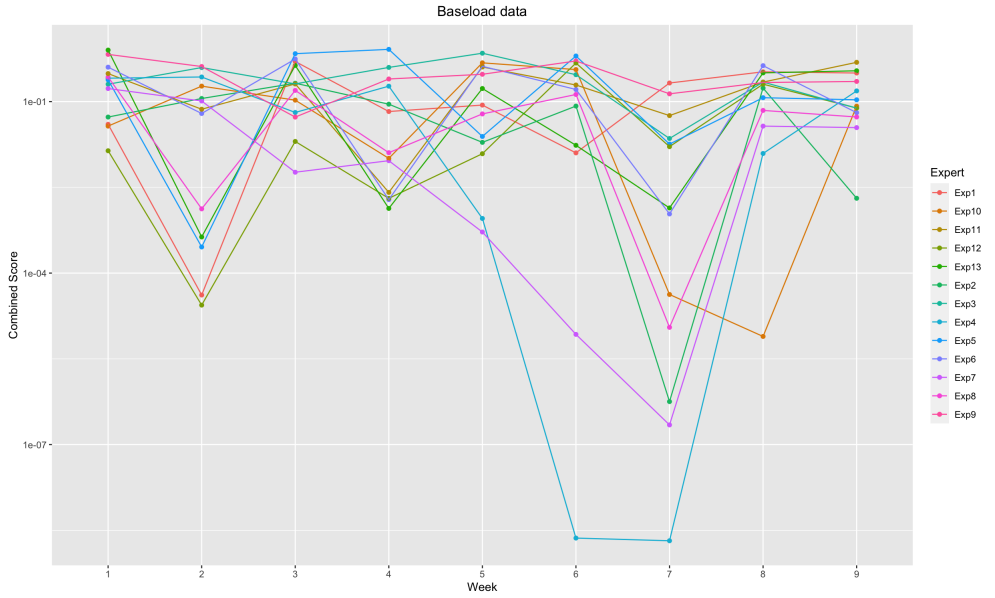

Figure D.14: Expert performance on a week basis. Presented are the combined scores. The scores are calculated for the baseload prices. Weeks are not aggregated.

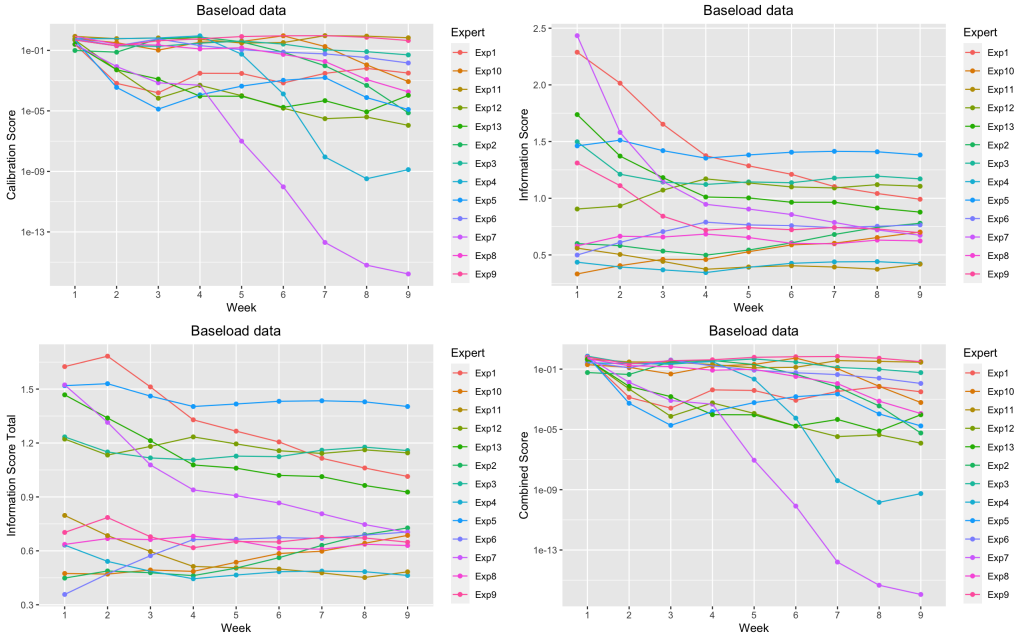

Figure D.15: Expert performance on a work-week basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. The scores are calculated for the baseload prices. Weeks are aggregated, i.e. we keep adding 1 week to the data.

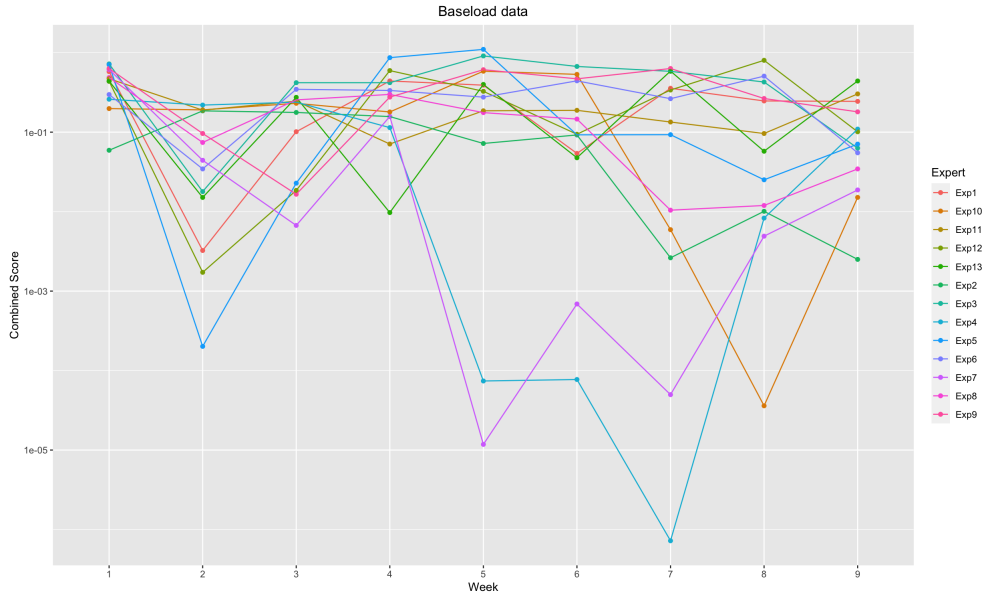

Figure D.16: Expert performance on a work-week basis. Presented are the combined scores. The scores are calculated for the baseload prices. Weeks are not aggregated.

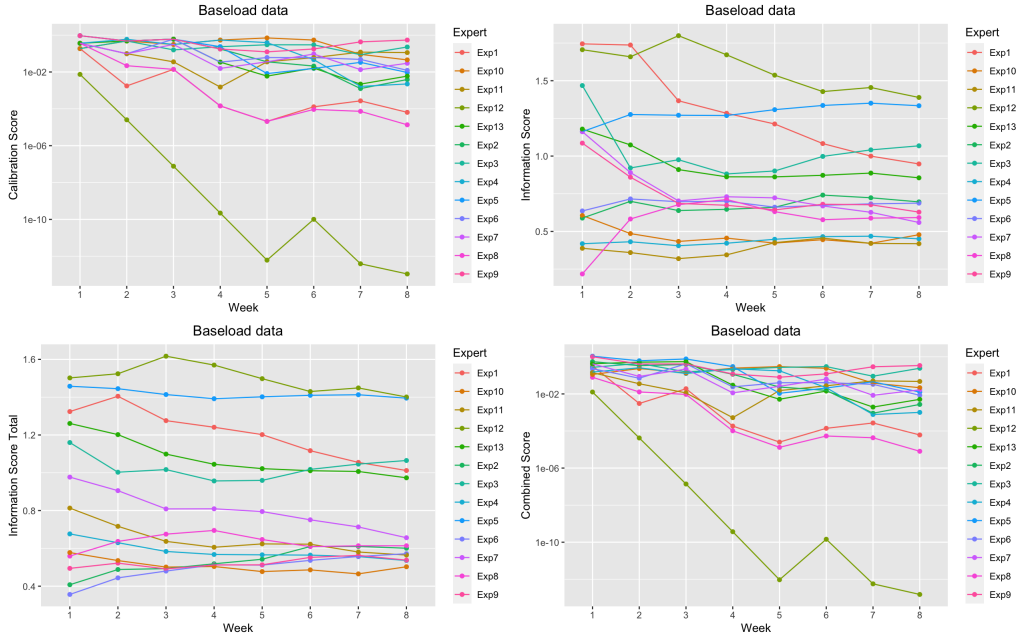

Figure D.17: Expert performance on a weekend basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. The scores are calculated for the baseload prices. Weeks are aggregated, i.e. we keep adding 1 week to the data.

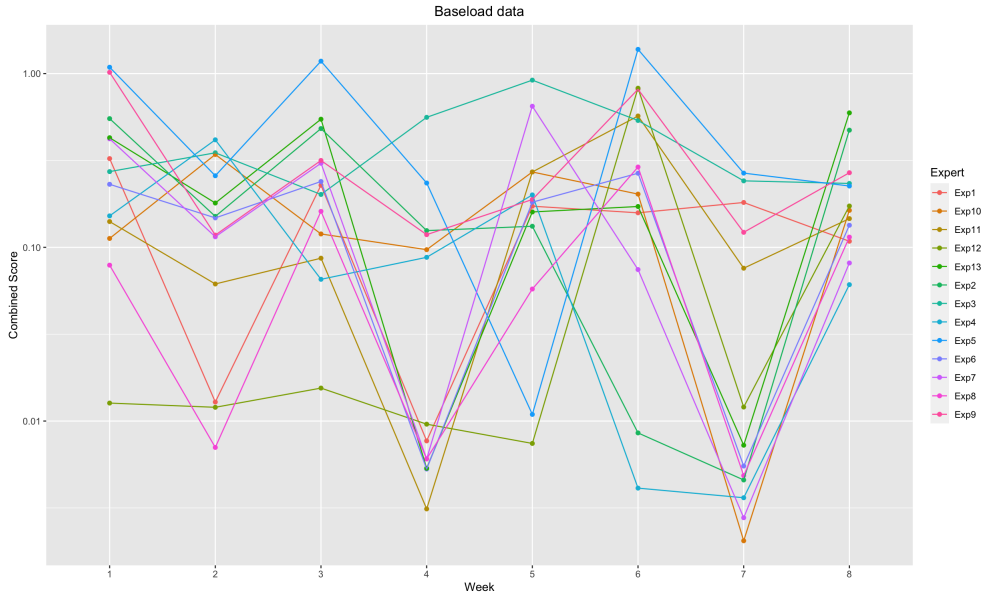

Figure D.18: Expert performance on a weekend basis. Presented are the combined scores. The scores are calculated for the baseload prices. Weeks are not aggregated.

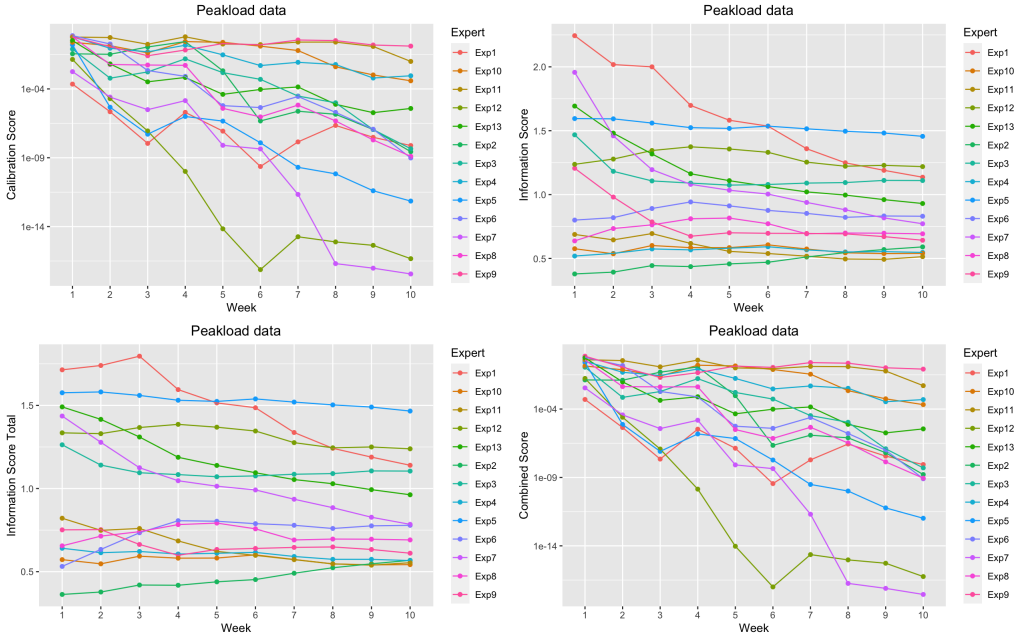

Figure D.19: Expert performance on a week basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. The scores are calculated for the peakload prices. Weeks are aggregated, i.e. we keep adding 1 week to the data.

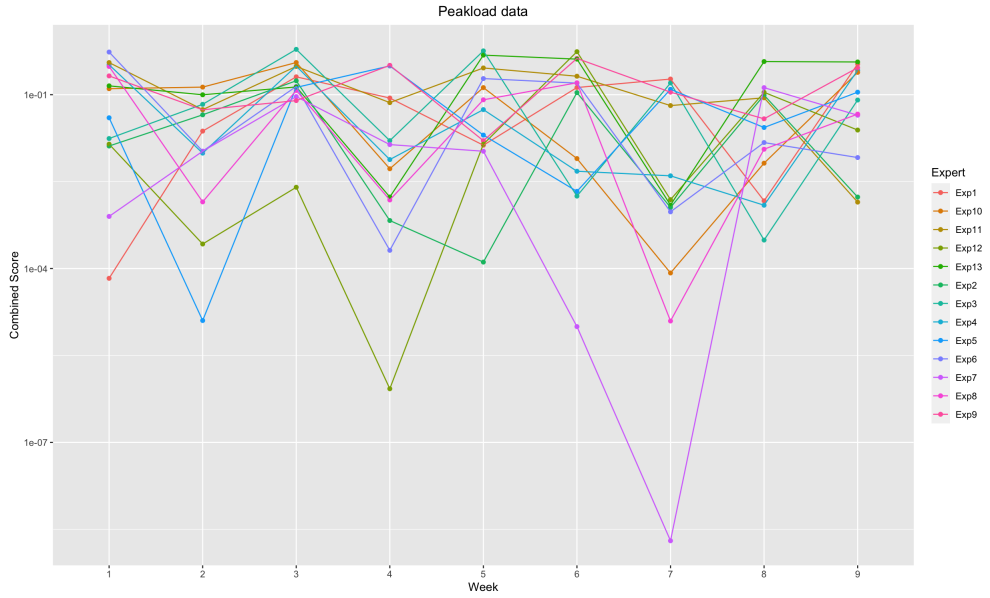

Figure D.20: Expert performance on a week basis. Presented are the combined scores. The scores are calculated for the peakload prices. Weeks are not aggregated.

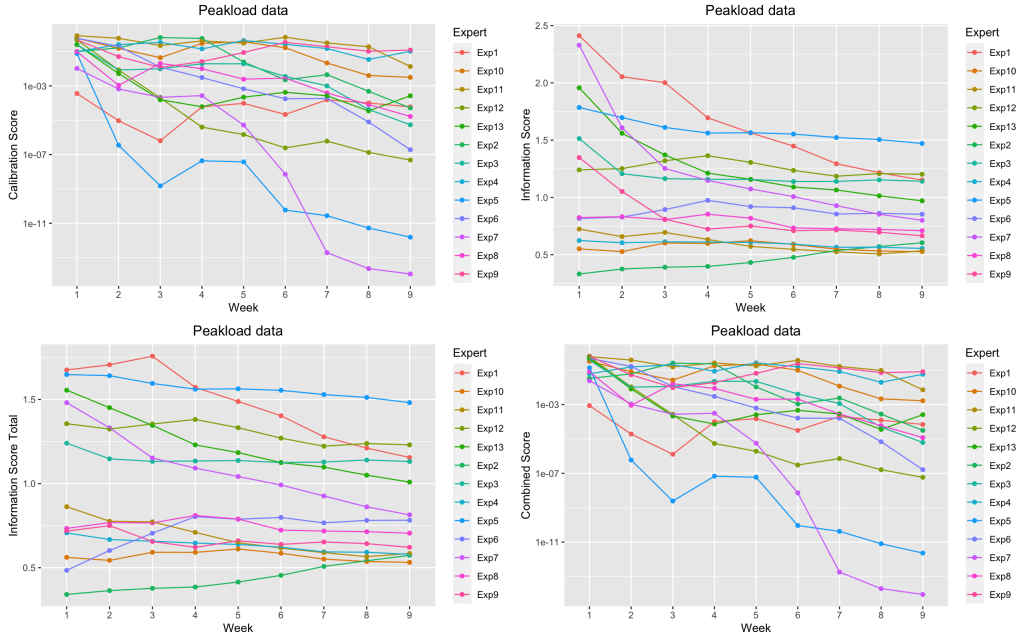

Figure D.21: Expert performance on a work-week basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. The scores are calculated for the peakload prices. Weeks are aggregated, i.e. we keep adding 1 week to the data.

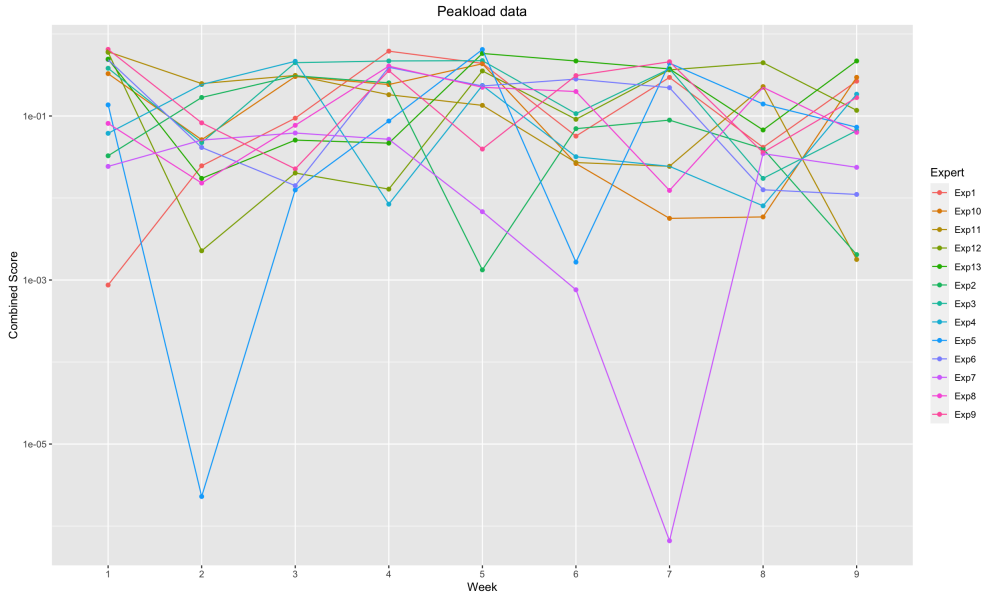

Figure D.22: Expert performance on a work-week basis. Presented are the combined scores. The scores are calculated for the peakload prices. Weeks are not aggregated.

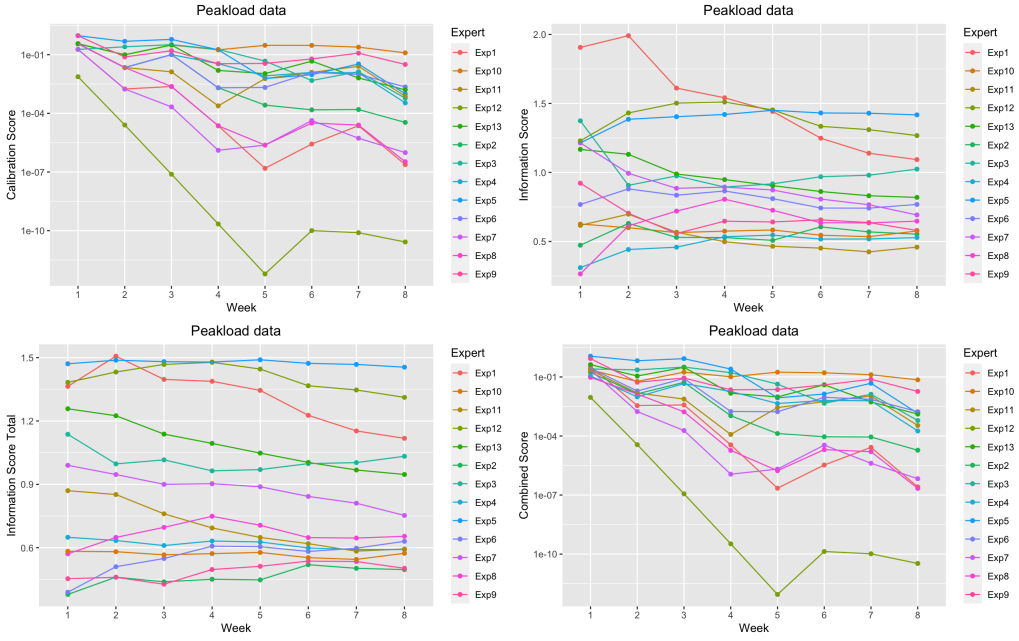

Figure D.23: Expert performance on a weekend basis. Presented are the calibration scores, information score all questions, information score seed questions and combined scores. The scores are calculated for the peakload prices. Weeks are aggregated, i.e. we keep adding 1 week to the data.

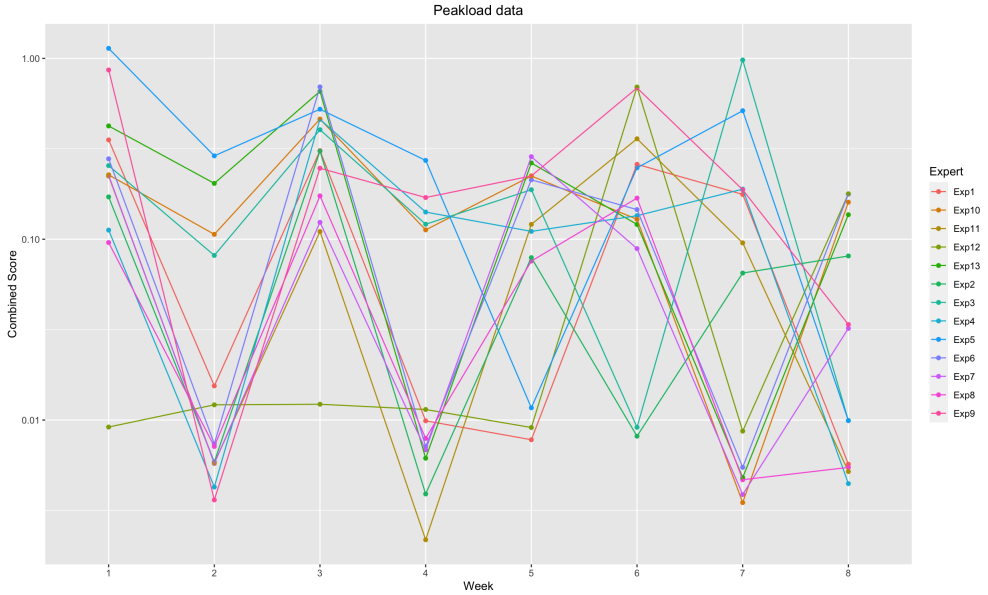

Figure D.24: Expert performance on a weekend basis. Presented are the combined scores. The scores are calculated for the peakload prices. Weeks are not aggregated.

E

Expert performance tables first elicitation

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$1.14 \cdot 10^{-13}$	1.064	1.058	$1.20 \cdot 10^{-13}$	$9.56 \cdot 10^{-13}$	$5.18 \cdot 10^{-13}$
Expert 2	$3.46 \cdot 10^{-16}$	0.658	0.674	$2.33 \cdot 10^{-16}$	$1.85 \cdot 10^{-15}$	$1.00 \cdot 10^{-15}$
Expert 3	$4.96 \cdot 10^{-10}$	1.123	1.127	$5.59 \cdot 10^{-10}$	$4.44 \cdot 10^{-09}$	$2.41 \cdot 10^{-09}$
Expert 4	$3.05 \cdot 10^{-11}$	0.502	0.489	$1.49 \cdot 10^{-11}$	$1.19 \cdot 10^{-10}$	$6.42 \cdot 10^{-11}$
Expert 5	$5.75 \cdot 10^{-18}$	1.420	1.413	$8.12 \cdot 10^{-18}$	$6.45 \cdot 10^{-17}$	$3.49 \cdot 10^{-17}$
Expert 6	$1.22 \cdot 10^{-12}$	0.763	0.788	$9.62 \cdot 10^{-13}$	$7.64 \cdot 10^{-12}$	$4.14 \cdot 10^{-12}$
Expert 7	0	0.717	0.707	0	0	0
Expert 8	$4.48 \cdot 10^{-16}$	0.655	0.654	$2.93 \cdot 10^{-16}$	$2.33 \cdot 10^{-15}$	$1.26 \cdot 10^{-15}$
Expert 9	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$	$6.99 \cdot 10^{-01}$	$3.79 \cdot 10^{-01}$
Expert 10	$4.51 \cdot 10^{-09}$	0.590	0.591	$2.67 \cdot 10^{-09}$	$2.12 \cdot 10^{-08}$	$1.15 \cdot 10^{-08}$
Expert 11	$8.14 \cdot 10^{-02}$	0.490	0.466	$3.80 \cdot 10^{-02}$	$3.01 \cdot 10^{-01}$	$1.62 \cdot 10^{-01}$
Expert 12	0	1.212	1.201	0	0	0
Expert 13	$2.13 \cdot 10^{-12}$	0.919	0.901	$1.92 \cdot 10^{-12}$	$1.52 \cdot 10^{-11}$	$8.25 \cdot 10^{-12}$
GWDM	$3.11 \cdot 10^{-01}$	0.340	0.342	$1.06 \cdot 10^{-01}$		$4.58 \cdot 10^{-01}$

Table E.1: Expert performance based on Global Weights and the Global Weight Decision Maker.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$1.14 \cdot 10^{-13}$	1.064	1.058	$1.20 \cdot 10^{-13}$	0	0
Expert 2	$3.46 \cdot 10^{-16}$	0.658	0.674	$2.33 \cdot 10^{-16}$	0	0
Expert 3	$4.96 \cdot 10^{-10}$	1.123	1.127	$5.59 \cdot 10^{-10}$	0	0
Expert 4	$3.05 \cdot 10^{-11}$	0.502	0.489	$1.49 \cdot 10^{-11}$	0	0
Expert 5	$5.75 \cdot 10^{-18}$	1.420	1.413	$8.12 \cdot 10^{-18}$	0	0
Expert 6	$1.22 \cdot 10^{-12}$	0.763	0.788	$9.62 \cdot 10^{-13}$	0	0
Expert 7	0	0.717	0.707	0	0	0
Expert 8	$4.48 \cdot 10^{-16}$	0.655	0.654	$2.93 \cdot 10^{-16}$	0	0
Expert 9	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$	$6.99 \cdot 10^{-01}$	$3.79 \cdot 10^{-01}$
Expert 10	$4.51 \cdot 10^{-09}$	0.590	0.591	$2.67 \cdot 10^{-09}$	0	0
Expert 11	$8.14 \cdot 10^{-02}$	0.490	0.466	$3.80 \cdot 10^{-02}$	$3.01 \cdot 10^{-01}$	$1.63 \cdot 10^{-01}$
Expert 12	0	1.212	1.201	0	0	0
Expert 13	$2.13 \cdot 10^{-12}$	0.919	0.901	$1.92 \cdot 10^{-12}$	0	0
GWDM	$3.11 \cdot 10^{-01}$	0.340	0.342	$1.06 \cdot 10^{-01}$		$4.58 \cdot 10^{-01}$

Table E.2: Expert performance based on Global Weights and the Global Weight Decision Maker with $\alpha = 0.08138$.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$1.14 \cdot 10^{-13}$	1.064	1.058	$1.20 \cdot 10^{-13}$	0	0
Expert 2	$3.46 \cdot 10^{-16}$	0.658	0.674	$2.33 \cdot 10^{-16}$	0	0
Expert 3	$4.96 \cdot 10^{-10}$	1.123	1.127	$5.59 \cdot 10^{-10}$	0	0
Expert 4	$3.05 \cdot 10^{-11}$	0.502	0.489	$1.49 \cdot 10^{-11}$	0	0
Expert 5	$5.75 \cdot 10^{-18}$	1.420	1.413	$8.12 \cdot 10^{-18}$	0	0
Expert 6	$1.22 \cdot 10^{-12}$	0.763	0.788	$9.62 \cdot 10^{-13}$	0	0
Expert 7	0	0.717	0.707	0	0	0
Expert 8	$4.48 \cdot 10^{-16}$	0.655	0.654	$2.93 \cdot 10^{-16}$	0	0
Expert 9	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$	$6.99 \cdot 10^{-01}$	$3.79 \cdot 10^{-01}$
Expert 10	$4.51 \cdot 10^{-09}$	0.590	0.591	$2.67 \cdot 10^{-09}$	0	0
Expert 11	$8.14 \cdot 10^{-02}$	0.490	0.466	$3.80 \cdot 10^{-02}$	0	0
Expert 12	0	1.212	1.201	0	0	0
Expert 13	$2.13 \cdot 10^{-12}$	0.919	0.901	$1.92 \cdot 10^{-12}$	0	0
GWDM	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$		$5.00 \cdot 10^{-01}$

Table E.3: Expert performance based on Global Weights and the Global Weight Decision Maker with $\alpha = 0.1332$.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$1.14 \cdot 10^{-13}$	1.064	1.058	$1.20 \cdot 10^{-13}$	$7.69 \cdot 10^{-02}$	$8.68 \cdot 10^{-13}$
Expert 2	$3.46 \cdot 10^{-16}$	0.658	0.674	$2.33 \cdot 10^{-16}$	$7.69 \cdot 10^{-02}$	$1.68 \cdot 10^{-15}$
Expert 3	$4.96 \cdot 10^{-10}$	1.123	1.127	$5.59 \cdot 10^{-10}$	$7.69 \cdot 10^{-02}$	$4.03 \cdot 10^{-09}$
Expert 4	$3.05 \cdot 10^{-11}$	0.502	0.489	$1.49 \cdot 10^{-11}$	$7.69 \cdot 10^{-02}$	$1.08 \cdot 10^{-10}$
Expert 5	$5.75 \cdot 10^{-18}$	1.420	1.413	$8.12 \cdot 10^{-18}$	$7.69 \cdot 10^{-02}$	$5.86 \cdot 10^{-17}$
Expert 6	$1.22 \cdot 10^{-12}$	0.763	0.788	$9.62 \cdot 10^{-13}$	$7.69 \cdot 10^{-02}$	$6.94 \cdot 10^{-12}$
Expert 7	0	0.717	0.707	0	$7.69 \cdot 10^{-02}$	0
Expert 8	$4.48 \cdot 10^{-16}$	0.655	0.654	$2.93 \cdot 10^{-16}$	$7.69 \cdot 10^{-02}$	$2.12 \cdot 10^{-15}$
Expert 9	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$	$7.69 \cdot 10^{-02}$	$6.35 \cdot 10^{-01}$
Expert 10	$4.51 \cdot 10^{-09}$	0.590	0.591	$2.67 \cdot 10^{-09}$	$7.69 \cdot 10^{-02}$	$1.92 \cdot 10^{-08}$
Expert 11	$8.14 \cdot 10^{-02}$	0.490	0.466	$3.80 \cdot 10^{-02}$	$7.69 \cdot 10^{-02}$	$2.74 \cdot 10^{-01}$
Expert 12	0	1.212	1.201	0	0	0
Expert 13	$2.13 \cdot 10^{-12}$	0.919	0.901	$1.92 \cdot 10^{-12}$	$7.69 \cdot 10^{-02}$	0
EWDM	$6.17 \cdot 10^{-02}$	0.207	0.206	$1.27 \cdot 10^{-02}$		$9.17 \cdot 10^{-02}$

Table E.4: Expert performance based on Equal Weights and the Equal Weight Decision Maker.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$1.14 \cdot 10^{-13}$	1.064	1.058	$1.20 \cdot 10^{-13}$		$3.56 \cdot 10^{-13}$
Expert 2	$3.46 \cdot 10^{-16}$	0.658	0.674	$2.33 \cdot 10^{-16}$		$6.89 \cdot 10^{-16}$
Expert 3	$4.96 \cdot 10^{-10}$	1.123	1.127	$5.59 \cdot 10^{-10}$		$1.66 \cdot 10^{-09}$
Expert 4	$3.05 \cdot 10^{-11}$	0.502	0.489	$1.49 \cdot 10^{-11}$		$4.42 \cdot 10^{-11}$
Expert 5	$5.75 \cdot 10^{-18}$	1.420	1.413	$8.12 \cdot 10^{-18}$		$2.40 \cdot 10^{-17}$
Expert 6	$1.22 \cdot 10^{-12}$	0.763	0.788	$9.62 \cdot 10^{-13}$		$2.85 \cdot 10^{-12}$
Expert 7	0	0.717	0.707	0		0
Expert 8	$4.48 \cdot 10^{-16}$	0.655	0.654	$2.93 \cdot 10^{-16}$		$8.68 \cdot 10^{-16}$
Expert 9	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$		$2.61 \cdot 10^{-01}$
Expert 10	$4.51 \cdot 10^{-09}$	0.590	0.591	$2.67 \cdot 10^{-09}$		$7.89 \cdot 10^{-09}$
Expert 11	$8.14 \cdot 10^{-02}$	0.490	0.466	$3.80 \cdot 10^{-02}$		$1.12 \cdot 10^{-01}$
Expert 12	0	1.212	1.201	0		0
Expert 13	$2.13 \cdot 10^{-12}$	0.919	0.901	$1.92 \cdot 10^{-12}$		$5.68 \cdot 10^{-12}$
IWDM	$5.64 \cdot 10^{-01}$	0.376	0.376	$2.12 \cdot 10^{-01}$		$6.27 \cdot 10^{-01}$

Table E.5: Expert performance based on Item Weights and the Item Weight Decision Maker.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$1.14 \cdot 10^{-13}$	1.064	1.058	$1.20 \cdot 10^{-13}$		0
Expert 2	$3.46 \cdot 10^{-16}$	0.658	0.674	$2.33 \cdot 10^{-16}$		0
Expert 3	$4.96 \cdot 10^{-10}$	1.123	1.127	$5.59 \cdot 10^{-10}$		0
Expert 4	$3.05 \cdot 10^{-11}$	0.502	0.489	$1.49 \cdot 10^{-11}$		0
Expert 5	$5.75 \cdot 10^{-18}$	1.420	1.413	$8.12 \cdot 10^{-18}$		0
Expert 6	$1.22 \cdot 10^{-12}$	0.763	0.788	$9.62 \cdot 10^{-13}$		0
Expert 7	0	0.717	0.707	0		0
Expert 8	$4.48 \cdot 10^{-16}$	0.655	0.654	$2.93 \cdot 10^{-16}$		0
Expert 9	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$		$2.61 \cdot 10^{-01}$
Expert 10	$4.51 \cdot 10^{-09}$	0.590	0.591	$2.67 \cdot 10^{-09}$		0
Expert 11	$8.14 \cdot 10^{-02}$	0.490	0.466	$3.80 \cdot 10^{-02}$		$1.12 \cdot 10^{-01}$
Expert 12	0	1.212	1.201	0		0
Expert 13	$2.13 \cdot 10^{-12}$	0.919	0.901	$1.92 \cdot 10^{-12}$		0
IWDM	$5.64 \cdot 10^{-01}$	0.376	0.376	$2.12 \cdot 10^{-01}$		$6.27 \cdot 10^{-01}$

Table E.6: Expert performance based on Item Weights and the Item Weight Decision Maker with $\alpha = 0.08138$.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$1.14 \cdot 10^{-13}$	1.064	1.058	$1.20 \cdot 10^{-13}$		0
Expert 2	$3.46 \cdot 10^{-16}$	0.658	0.674	$2.33 \cdot 10^{-16}$		0
Expert 3	$4.96 \cdot 10^{-10}$	1.123	1.127	$5.59 \cdot 10^{-10}$		0
Expert 4	$3.05 \cdot 10^{-11}$	0.502	0.489	$1.49 \cdot 10^{-11}$		0
Expert 5	$5.75 \cdot 10^{-18}$	1.420	1.413	$8.12 \cdot 10^{-18}$		0
Expert 6	$1.22 \cdot 10^{-12}$	0.763	0.788	$9.62 \cdot 10^{-13}$		0
Expert 7	0	0.717	0.707	0		0
Expert 8	$4.48 \cdot 10^{-16}$	0.655	0.654	$2.93 \cdot 10^{-16}$		0
Expert 9	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$		$5.00 \cdot 10^{-01}$
Expert 10	$4.51 \cdot 10^{-09}$	0.590	0.591	$2.67 \cdot 10^{-09}$		0
Expert 11	$8.14 \cdot 10^{-02}$	0.490	0.466	$3.80 \cdot 10^{-02}$		0
Expert 12	0	1.212	1.201	0		0
Expert 13	$2.13 \cdot 10^{-12}$	0.919	0.901	$1.92 \cdot 10^{-12}$		0
IWDM	$1.33 \cdot 10^{-01}$	0.643	0.660	$8.80 \cdot 10^{-02}$		$5.00 \cdot 10^{-01}$

Table E.7: Expert performance based on Item Weights and the Item Weight Decision Maker with $\alpha = 0.1332$.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$2.90 \cdot 10^{-05}$	0.998	0.980	$2.84 \cdot 10^{-05}$	$4.20 \cdot 10^{-05}$	$4.12 \cdot 10^{-05}$
Expert 2	$5.40 \cdot 10^{-08}$	0.720	0.757	$4.09 \cdot 10^{-08}$	$6.04 \cdot 10^{-08}$	$5.93 \cdot 10^{-08}$
Expert 3	$8.77 \cdot 10^{-03}$	1.136	1.144	$1.00 \cdot 10^{-02}$	$1.48 \cdot 10^{-02}$	$1.46 \cdot 10^{-02}$
Expert 4	$1.40 \cdot 10^{-11}$	0.460	0.430	$6.04 \cdot 10^{-12}$	$8.92 \cdot 10^{-12}$	$8.75 \cdot 10^{-12}$
Expert 5	$6.65 \cdot 10^{-06}$	1.386	1.369	$9.10 \cdot 10^{-06}$	$1.35 \cdot 10^{-05}$	$1.32 \cdot 10^{-05}$
Expert 6	$7.30 \cdot 10^{-04}$	0.701	0.745	$5.44 \cdot 10^{-04}$	$8.00 \cdot 10^{-04}$	$7.90 \cdot 10^{-04}$
Expert 7	$3.24 \cdot 10^{-16}$	0.668	0.643	$2.08 \cdot 10^{-16}$	$3.08 \cdot 10^{-16}$	$3.02 \cdot 10^{-16}$
Expert 8	$6.56 \cdot 10^{-07}$	0.620	0.615	$4.03 \cdot 10^{-07}$	$5.96 \cdot 10^{-07}$	$5.85 \cdot 10^{-07}$
Expert 9	$7.39 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$	$7.41 \cdot 10^{-01}$	$7.26 \cdot 10^{-01}$
Expert 10	$2.90 \cdot 10^{-05}$	0.636	0.642	$1.86 \cdot 10^{-05}$	$2.76 \cdot 10^{-05}$	$2.70 \cdot 10^{-05}$
Expert 11	$3.94 \cdot 10^{-01}$	0.468	0.419	$1.65 \cdot 10^{-01}$	$2.44 \cdot 10^{-01}$	$2.39 \cdot 10^{-01}$
Expert 12	$5.39 \cdot 10^{-16}$	1.205	1.182	$6.37 \cdot 10^{-16}$	$9.41 \cdot 10^{-16}$	$9.23 \cdot 10^{-16}$
Expert 13	$1.13 \cdot 10^{-06}$	0.910	0.872	$9.82 \cdot 10^{-07}$	$1.45 \cdot 10^{-06}$	$1.42 \cdot 10^{-06}$
GWDM	$3.84 \cdot 10^{-02}$	0.346	0.350	$1.35 \cdot 10^{-02}$		$1.95 \cdot 10^{-02}$

Table E.8: Expert performance based on Global Weights and the Global Weight Decision Maker. Scores are based on baseload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$2.90 \cdot 10^{-05}$	0.998	0.980	$2.84 \cdot 10^{-05}$	0	0
Expert 2	$5.40 \cdot 10^{-08}$	0.720	0.757	$4.09 \cdot 10^{-08}$	0	0
Expert 3	$8.77 \cdot 10^{-03}$	1.136	1.144	$1.01 \cdot 10^{-02}$	0	0
Expert 4	$1.40 \cdot 10^{-11}$	0.460	0.430	$6.04 \cdot 10^{-12}$	0	0
Expert 5	$6.65 \cdot 10^{-06}$	1.386	1.369	$9.10 \cdot 10^{-06}$	0	0
Expert 6	$7.30 \cdot 10^{-04}$	0.701	0.745	$5.44 \cdot 10^{-04}$	0	0
Expert 7	$3.24 \cdot 10^{-16}$	0.668	0.643	$2.08 \cdot 10^{-16}$	0	0
Expert 8	$6.56 \cdot 10^{-07}$	0.620	0.615	$4.03 \cdot 10^{-07}$	0	0
Expert 9	$7.39 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$	1	$5.00 \cdot 10^{-01}$
Expert 10	$2.90 \cdot 10^{-05}$	0.636	0.642	$1.86 \cdot 10^{-05}$	0	0
Expert 11	$3.94 \cdot 10^{-01}$	0.468	0.419	$1.65 \cdot 10^{-01}$	0	0
Expert 12	$5.39 \cdot 10^{-16}$	1.205	1.182	$6.37 \cdot 10^{-16}$	0	0
Expert 13	$1.13 \cdot 10^{-06}$	0.910	0.872	$9.82 \cdot 10^{-07}$	0	0
GWDM	$7.39 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$

Table E.9: Expert performance based on Global Weights and the optimized Global Weight Decision Maker. Scores are based on baseload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$2.90 \cdot 10^{-05}$	0.998	0.980	$2.84 \cdot 10^{-05}$	0	0
Expert 2	$5.40 \cdot 10^{-08}$	0.720	0.757	$4.09 \cdot 10^{-08}$	0	0
Expert 3	$8.77 \cdot 10^{-03}$	1.136	1.144	$1.01 \cdot 10^{-02}$	0	0
Expert 4	$1.40 \cdot 10^{-11}$	0.460	0.430	$6.04 \cdot 10^{-12}$	0	0
Expert 5	$6.65 \cdot 10^{-06}$	1.386	1.369	$9.10 \cdot 10^{-06}$	0	0
Expert 6	$7.30 \cdot 10^{-04}$	0.701	0.745	$5.44 \cdot 10^{-04}$	0	0
Expert 7	$3.24 \cdot 10^{-16}$	0.668	0.643	$2.08 \cdot 10^{-16}$	0	0
Expert 8	$6.56 \cdot 10^{-07}$	0.620	0.615	$4.03 \cdot 10^{-07}$	0	0
Expert 9	$7.39 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$	$7.53 \cdot 10^{-01}$	$7.38 \cdot 10^{-01}$
Expert 10	$2.90 \cdot 10^{-05}$	0.636	0.642	$1.86 \cdot 10^{-05}$	0	0
Expert 11	$3.94 \cdot 10^{-01}$	0.468	0.419	$1.65 \cdot 10^{-01}$	$2.48 \cdot 10^{-01}$	$2.43 \cdot 10^{-01}$
Expert 12	$5.39 \cdot 10^{-16}$	1.205	1.182	$6.37 \cdot 10^{-16}$	0	0
Expert 13	$1.13 \cdot 10^{-06}$	0.910	0.872	$9.82 \cdot 10^{-07}$	0	0
GWDM	$3.84 \cdot 10^{-02}$	0.347	0.351	$1.35 \cdot 10^{-02}$		$1.99 \cdot 10^{-02}$

Table E.10: Expert performance based on Global Weights and the Global Weight Decision Maker with $\alpha = 0.3$. Scores are based on baseload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$2.90 \cdot 10^{-05}$	0.998	0.980	$2.84 \cdot 10^{-05}$	0.07692	$4.09 \cdot 10^{-05}$
Expert 2	$5.40 \cdot 10^{-08}$	0.720	0.757	$4.09 \cdot 10^{-08}$	0.07692	$5.89 \cdot 10^{-08}$
Expert 3	$8.77 \cdot 10^{-03}$	1.136	1.144	$1.01 \cdot 10^{-02}$	0.07692	$1.45 \cdot 10^{-02}$
Expert 4	$1.40 \cdot 10^{-11}$	0.460	0.430	$6.04 \cdot 10^{-12}$	0.07692	$8.69 \cdot 10^{-12}$
Expert 5	$6.65 \cdot 10^{-06}$	1.386	1.369	$9.10 \cdot 10^{-06}$	0.07692	$1.31 \cdot 10^{-05}$
Expert 6	$7.30 \cdot 10^{-04}$	0.701	0.745	$5.44 \cdot 10^{-04}$	0.07692	$7.80 \cdot 10^{-04}$
Expert 7	$3.24 \cdot 10^{-16}$	0.668	0.643	$2.08 \cdot 10^{-16}$	0.07692	$3.00 \cdot 10^{-16}$
Expert 8	$6.56 \cdot 10^{-07}$	0.620	0.615	$4.03 \cdot 10^{-07}$	0.07692	$5.81 \cdot 10^{-07}$
Expert 9	$7.39 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$	0.07692	$7.21 \cdot 10^{-01}$
Expert 10	$2.90 \cdot 10^{-05}$	0.636	0.642	$1.86 \cdot 10^{-05}$	0.07692	$2.68 \cdot 10^{-05}$
Expert 11	$3.94 \cdot 10^{-01}$	0.468	0.419	$1.65 \cdot 10^{-01}$	0.07692	$2.37 \cdot 10^{-01}$
Expert 12	$5.39 \cdot 10^{-16}$	1.205	1.182	$6.37 \cdot 10^{-16}$	0.07692	$9.17 \cdot 10^{-16}$
Expert 13	$1.13 \cdot 10^{-06}$	0.910	0.872	$9.82 \cdot 10^{-07}$	0.07692	$1.41 \cdot 10^{-06}$
EWDM	$9.39 \cdot 10^{-02}$	0.196	0.193	$1.81 \cdot 10^{-02}$		$2.61 \cdot 10^{-02}$

Table E.11: Expert performance based on Equal Weights and the Equal Weight Decision Maker. Scores are based on baseload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$2.90 \cdot 10^{-05}$	0.998	0.980	$2.84 \cdot 10^{-05}$		$4.06 \cdot 10^{-05}$
Expert 2	$5.40 \cdot 10^{-08}$	0.720	0.757	$4.09 \cdot 10^{-08}$		$5.83 \cdot 10^{-08}$
Expert 3	$8.77 \cdot 10^{-03}$	1.136	1.144	$1.00 \cdot 10^{-02}$		$1.43 \cdot 10^{-02}$
Expert 4	$1.40 \cdot 10^{-11}$	0.460	0.430	$6.04 \cdot 10^{-12}$		$8.61 \cdot 10^{-12}$
Expert 5	$6.65 \cdot 10^{-06}$	1.386	1.369	$9.10 \cdot 10^{-06}$		$1.30 \cdot 10^{-05}$
Expert 6	$7.30 \cdot 10^{-04}$	0.701	0.745	$5.44 \cdot 10^{-04}$		$7.80 \cdot 10^{-04}$
Expert 7	$3.24 \cdot 10^{-16}$	0.668	0.643	$2.08 \cdot 10^{-16}$		$2.97 \cdot 10^{-16}$
Expert 8	$6.56 \cdot 10^{-07}$	0.620	0.615	$4.03 \cdot 10^{-07}$		$5.76 \cdot 10^{-07}$
Expert 9	$7.39 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$		$7.15 \cdot 10^{-01}$
Expert 10	$2.90 \cdot 10^{-05}$	0.636	0.642	$1.86 \cdot 10^{-05}$		$2.66 \cdot 10^{-05}$
Expert 11	$3.94 \cdot 10^{-01}$	0.468	0.419	$1.65 \cdot 10^{-01}$		$2.35 \cdot 10^{-01}$
Expert 12	$5.39 \cdot 10^{-16}$	1.205	1.182	$6.37 \cdot 10^{-16}$		$9.09 \cdot 10^{-16}$
Expert 13	$1.13 \cdot 10^{-06}$	0.910	0.872	$9.82 \cdot 10^{-07}$		$1.40 \cdot 10^{-06}$
IWDM	$6.34 \cdot 10^{-02}$	0.383	0.383	$2.34 \cdot 10^{-02}$		$3.47 \cdot 10^{-02}$

Table E.12: Expert performance based on Item Weights and the Item Weight Decision Maker. Scores are based on baseload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$2.90 \cdot 10^{-05}$	0.998	0.980	$2.84 \cdot 10^{-05}$		0
Expert 2	$5.40 \cdot 10^{-08}$	0.720	0.757	$4.09 \cdot 10^{-08}$		0
Expert 3	$8.77 \cdot 10^{-03}$	1.136	1.144	$1.00 \cdot 10^{-02}$		0
Expert 4	$1.40 \cdot 10^{-11}$	0.460	0.430	$6.04 \cdot 10^{-12}$		0
Expert 5	$6.65 \cdot 10^{-06}$	1.386	1.369	$9.10 \cdot 10^{-06}$		0
Expert 6	$7.30 \cdot 10^{-04}$	0.701	0.745	$5.44 \cdot 10^{-04}$		0
Expert 7	$3.24 \cdot 10^{-16}$	0.668	0.643	$2.08 \cdot 10^{-16}$		0
Expert 8	$6.56 \cdot 10^{-07}$	0.620	0.615	$4.03 \cdot 10^{-07}$		0
Expert 9	$7.39 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$
Expert 10	$2.90 \cdot 10^{-05}$	0.636	0.642	$1.86 \cdot 10^{-05}$		0
Expert 11	$3.94 \cdot 10^{-01}$	0.468	0.419	$1.65 \cdot 10^{-01}$		0
Expert 12	$5.39 \cdot 10^{-16}$	1.205	1.182	$6.37 \cdot 10^{-16}$		0
Expert 13	$1.13 \cdot 10^{-06}$	0.910	0.872	$9.82 \cdot 10^{-07}$		0
IWDM	$7.39 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$

Table E.13: Expert performance based on Item Weights and the optimized Item Weight Decision Maker. Scores are based on baseload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$2.90 \cdot 10^{-05}$	0.998	0.980	$2.84 \cdot 10^{-05}$		0
Expert 2	$5.40 \cdot 10^{-08}$	0.720	0.757	$4.09 \cdot 10^{-08}$		0
Expert 3	$8.77 \cdot 10^{-03}$	1.136	1.144	$1.00 \cdot 10^{-02}$		0
Expert 4	$1.40 \cdot 10^{-11}$	0.460	0.430	$6.04 \cdot 10^{-12}$		0
Expert 5	$6.65 \cdot 10^{-06}$	1.386	1.369	$9.10 \cdot 10^{-06}$		0
Expert 6	$7.30 \cdot 10^{-04}$	0.701	0.745	$5.44 \cdot 10^{-04}$		0
Expert 7	$3.24 \cdot 10^{-16}$	0.668	0.643	$2.08 \cdot 10^{-16}$		0
Expert 8	$6.56 \cdot 10^{-07}$	0.620	0.615	$4.03 \cdot 10^{-07}$		0
Expert 9	$7.40 \cdot 10^{-01}$	0.644	0.678	$5.01 \cdot 10^{-01}$		$7.26 \cdot 10^{-01}$
Expert 10	$2.90 \cdot 10^{-05}$	0.636	0.642	$1.86 \cdot 10^{-05}$		0
Expert 11	$3.94 \cdot 10^{-01}$	0.468	0.419	$1.65 \cdot 10^{-01}$		$2.39 \cdot 10^{-01}$
Expert 12	$5.39 \cdot 10^{-16}$	1.205	1.182	$6.37 \cdot 10^{-16}$		0
Expert 13	$1.13 \cdot 10^{-06}$	0.910	0.872	$9.82 \cdot 10^{-07}$		0
IWDM	$6.34 \cdot 10^{-02}$	0.382	0.383	$2.43 \cdot 10^{-02}$		$3.52 \cdot 10^{-02}$

Table E.14: Expert performance based on Item Weights and the Item Weight Decision Maker with $\alpha = 0.3$. Scores are based on baseload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$7.74 \cdot 10^{-09}$	1.140	1.136	$8.78 \cdot 10^{-09}$	$9.89 \cdot 10^{-08}$	$2.42 \cdot 10^{-08}$
Expert 2	$2.76 \cdot 10^{-09}$	0.568	0.591	$1.63 \cdot 10^{-09}$	$1.83 \cdot 10^{-08}$	$4.48 \cdot 10^{-09}$
Expert 3	$4.51 \cdot 10^{-09}$	1.105	1.110	$5.01 \cdot 10^{-09}$	$5.64 \cdot 10^{-08}$	$1.38 \cdot 10^{-08}$
Expert 4	$9.01 \cdot 10^{-04}$	0.567	0.547	$4.93 \cdot 10^{-04}$	$5.55 \cdot 10^{-03}$	$1.36 \cdot 10^{-03}$
Expert 5	$7.14 \cdot 10^{-13}$	1.466	1.456	$1.04 \cdot 10^{-12}$	$1.17 \cdot 10^{-11}$	$2.87 \cdot 10^{-12}$
Expert 6	$9.70 \cdot 10^{-10}$	0.779	0.830	$8.05 \cdot 10^{-10}$	$9.07 \cdot 10^{-09}$	$2.22 \cdot 10^{-09}$
Expert 7	$3.63 \cdot 10^{-18}$	0.785	0.772	$2.80 \cdot 10^{-18}$	$3.16 \cdot 10^{-17}$	$7.72 \cdot 10^{-18}$
Expert 8	$1.27 \cdot 10^{-09}$	0.691	0.693	$8.80 \cdot 10^{-10}$	$9.91 \cdot 10^{-09}$	$2.43 \cdot 10^{-09}$
Expert 9	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-02}$	$9.34 \cdot 10^{-01}$	$2.29 \cdot 10^{-01}$
Expert 10	$3.90 \cdot 10^{-04}$	0.542	0.540	$2.11 \cdot 10^{-04}$	$2.37 \cdot 10^{-03}$	$5.80 \cdot 10^{-03}$
Expert 11	$9.97 \cdot 10^{-03}$	0.554	0.514	$5.12 \cdot 10^{-03}$	$5.77 \cdot 10^{-02}$	$1.41 \cdot 10^{-02}$
Expert 12	$4.67 \cdot 10^{-17}$	1.239	1.219	$5.69 \cdot 10^{-17}$	$6.41 \cdot 10^{-16}$	$1.57 \cdot 10^{-16}$
Expert 13	$3.81 \cdot 10^{-06}$	0.963	0.930	$3.55 \cdot 10^{-06}$	$3.99 \cdot 10^{-05}$	$9.77 \cdot 10^{-06}$
GWDM	$6.25 \cdot 10^{-01}$	0.425	0.438	$2.74 \cdot 10^{-01}$		$7.55 \cdot 10^{-01}$

Table E.15: Expert performance based on Global Weights and the Global Weight Decision Maker. Scores are based on peakload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$7.74 \cdot 10^{-09}$	1.140	1.136	$8.78 \cdot 10^{-09}$	0	0
Expert 2	$2.76 \cdot 10^{-09}$	0.568	0.591	$1.63 \cdot 10^{-09}$	0	0
Expert 3	$4.51 \cdot 10^{-09}$	1.105	1.110	$5.01 \cdot 10^{-09}$	0	0
Expert 4	$9.01 \cdot 10^{-04}$	0.567	0.547	$4.93 \cdot 10^{-04}$	0	0
Expert 5	$7.14 \cdot 10^{-13}$	1.466	1.456	$1.04 \cdot 10^{-12}$	0	0
Expert 6	$9.70 \cdot 10^{-10}$	0.779	0.830	$8.05 \cdot 10^{-10}$	0	0
Expert 7	$3.63 \cdot 10^{-18}$	0.785	0.772	$2.80 \cdot 10^{-18}$	0	0
Expert 8	$1.27 \cdot 10^{-09}$	0.691	0.693	$8.80 \cdot 10^{-10}$	0	0
Expert 9	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-01}$	$9.42 \cdot 10^{-01}$	$2.23 \cdot 10^{-01}$
Expert 10	$3.90 \cdot 10^{-04}$	0.542	0.540	$2.11 \cdot 10^{-04}$	0	0
Expert 11	$9.97 \cdot 10^{-03}$	0.554	0.514	$5.12 \cdot 10^{-03}$	$5.81 \cdot 10^{-02}$	$1.36 \cdot 10^{-02}$
Expert 12	$4.67 \cdot 10^{-17}$	1.239	1.219	$5.69 \cdot 10^{-17}$	0	0
Expert 13	$3.81 \cdot 10^{-06}$	0.963	0.930	$3.55 \cdot 10^{-06}$	0	0
GWDM	$6.25 \cdot 10^{-01}$	0.440	0.454	$2.85 \cdot 10^{-01}$		$7.64 \cdot 10^{-01}$

Table E.16: Expert performance based on Global Weights and the optimized Global Weight Decision Maker. Scores are based on peakload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$7.74 \cdot 10^{-09}$	1.140	1.136	$8.78 \cdot 10^{-09}$	0	0
Expert 2	$2.76 \cdot 10^{-09}$	0.568	0.591	$1.63 \cdot 10^{-09}$	0	0
Expert 3	$4.51 \cdot 10^{-09}$	1.105	1.110	$5.01 \cdot 10^{-09}$	0	0
Expert 4	$9.01 \cdot 10^{-04}$	0.567	0.547	$4.93 \cdot 10^{-04}$	0	0
Expert 5	$7.14 \cdot 10^{-13}$	1.466	1.456	$1.04 \cdot 10^{-12}$	0	0
Expert 6	$9.70 \cdot 10^{-10}$	0.779	0.830	$8.05 \cdot 10^{-10}$	0	0
Expert 7	$3.63 \cdot 10^{-18}$	0.785	0.772	$2.80 \cdot 10^{-18}$	0	0
Expert 8	$1.27 \cdot 10^{-09}$	0.691	0.693	$8.80 \cdot 10^{-10}$	0	0
Expert 9	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-01}$	1	$5.00 \cdot 10^{-01}$
Expert 10	$3.90 \cdot 10^{-04}$	0.542	0.540	$2.11 \cdot 10^{-04}$	0	0
Expert 11	$9.97 \cdot 10^{-03}$	0.554	0.514	$5.12 \cdot 10^{-03}$	0	0
Expert 12	$4.67 \cdot 10^{-17}$	1.239	1.219	$5.69 \cdot 10^{-17}$	0	0
Expert 13	$3.81 \cdot 10^{-06}$	0.963	0.930	$3.55 \cdot 10^{-06}$	0	0
GWDM	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$

Table E.17: Expert performance based on Global Weights and the Global Weight Decision Maker with $\alpha = 0.1292$. Scores are based on peakload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$7.74 \cdot 10^{-09}$	1.140	1.136	$8.78 \cdot 10^{-09}$	0.07692	$5.39 \cdot 10^{-08}$
Expert 2	$2.76 \cdot 10^{-09}$	0.568	0.591	$1.63 \cdot 10^{-09}$	0.07692	$1.00 \cdot 10^{-08}$
Expert 3	$4.51 \cdot 10^{-09}$	1.105	1.110	$5.01 \cdot 10^{-09}$	0.07692	$3.07 \cdot 10^{-08}$
Expert 4	$9.01 \cdot 10^{-04}$	0.567	0.547	$4.93 \cdot 10^{-04}$	0.07692	$3.03 \cdot 10^{-03}$
Expert 5	$7.14 \cdot 10^{-13}$	1.466	1.456	$1.04 \cdot 10^{-12}$	0.07692	$6.39 \cdot 10^{-12}$
Expert 6	$9.70 \cdot 10^{-10}$	0.779	0.830	$8.05 \cdot 10^{-10}$	0.07692	$4.95 \cdot 10^{-09}$
Expert 7	$3.63 \cdot 10^{-18}$	0.785	0.772	$2.80 \cdot 10^{-18}$	0.07692	$1.72 \cdot 10^{-17}$
Expert 8	$1.27 \cdot 10^{-09}$	0.691	0.693	$8.80 \cdot 10^{-10}$	0.07692	$5.41 \cdot 10^{-09}$
Expert 9	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-01}$	0.07692	$5.10 \cdot 10^{-01}$
Expert 10	$3.90 \cdot 10^{-04}$	0.542	0.540	$2.11 \cdot 10^{-04}$	0.07692	$1.29 \cdot 10^{-03}$
Expert 11	$9.97 \cdot 10^{-03}$	0.554	0.514	$5.12 \cdot 10^{-03}$	0.07692	$3.15 \cdot 10^{-02}$
Expert 12	$4.67 \cdot 10^{-17}$	1.239	1.219	$5.69 \cdot 10^{-17}$	0.07692	$3.50 \cdot 10^{-16}$
Expert 13	$3.81 \cdot 10^{-06}$	0.963	0.930	$3.55 \cdot 10^{-06}$	0.07692	$2.18 \cdot 10^{-05}$
EWDM	$3.38 \cdot 10^{-01}$	0.220	0.219	$7.40 \cdot 10^{-02}$		$4.55 \cdot 10^{-01}$

Table E.18: Expert performance based on Equal Weights and the Equal Weight Decision Maker. Scores are based on peakload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$7.74 \cdot 10^{-09}$	1.140	1.136	$8.78 \cdot 10^{-09}$		$2.49 \cdot 10^{-08}$
Expert 2	$2.76 \cdot 10^{-09}$	0.568	0.591	$1.63 \cdot 10^{-09}$		$4.61 \cdot 10^{-09}$
Expert 3	$4.51 \cdot 10^{-09}$	1.105	1.110	$5.01 \cdot 10^{-09}$		$1.42 \cdot 10^{-08}$
Expert 4	$9.01 \cdot 10^{-04}$	0.567	0.547	$4.93 \cdot 10^{-04}$		$1.40 \cdot 10^{-03}$
Expert 5	$7.14 \cdot 10^{-13}$	1.466	1.456	$1.04 \cdot 10^{-12}$		$2.95 \cdot 10^{-12}$
Expert 6	$9.70 \cdot 10^{-10}$	0.779	0.830	$8.05 \cdot 10^{-10}$		$2.28 \cdot 10^{-09}$
Expert 7	$3.63 \cdot 10^{-18}$	0.785	0.772	$2.80 \cdot 10^{-18}$		$7.94 \cdot 10^{-18}$
Expert 8	$1.27 \cdot 10^{-09}$	0.691	0.693	$8.80 \cdot 10^{-10}$		$2.49 \cdot 10^{-09}$
Expert 9	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-01}$		$2.35 \cdot 10^{-01}$
Expert 10	$3.90 \cdot 10^{-04}$	0.542	0.540	$2.11 \cdot 10^{-04}$		$5.97 \cdot 10^{-04}$
Expert 11	$9.97 \cdot 10^{-03}$	0.554	0.514	$5.12 \cdot 10^{-03}$		$1.45 \cdot 10^{-02}$
Expert 12	$4.67 \cdot 10^{-17}$	1.239	1.219	$5.69 \cdot 10^{-17}$		$1.61 \cdot 10^{-16}$
Expert 13	$3.81 \cdot 10^{-06}$	0.963	0.930	$3.55 \cdot 10^{-06}$		$1.00 \cdot 10^{-05}$
IWDM	$5.57 \cdot 10^{-01}$	0.459	0.475	$2.64 \cdot 10^{-01}$		$7.48 \cdot 10^{-01}$

Table E.19: Expert performance based on Item Weights and the Item Weight Decision Maker. Scores are based on peakload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$7.74 \cdot 10^{-09}$	1.140	1.136	$8.78 \cdot 10^{-09}$		0
Expert 2	$2.76 \cdot 10^{-09}$	0.568	0.591	$1.63 \cdot 10^{-09}$		0
Expert 3	$4.51 \cdot 10^{-09}$	1.105	1.110	$5.01 \cdot 10^{-09}$		0
Expert 4	$9.01 \cdot 10^{-04}$	0.567	0.547	$4.93 \cdot 10^{-04}$		0
Expert 5	$7.14 \cdot 10^{-13}$	1.466	1.456	$1.04 \cdot 10^{-12}$		0
Expert 6	$9.70 \cdot 10^{-10}$	0.779	0.830	$8.05 \cdot 10^{-10}$		0
Expert 7	$3.63 \cdot 10^{-18}$	0.785	0.772	$2.80 \cdot 10^{-18}$		0
Expert 8	$1.27 \cdot 10^{-09}$	0.691	0.693	$8.80 \cdot 10^{-10}$		0
Expert 9	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-01}$		$2.29 \cdot 10^{-01}$
Expert 10	$3.90 \cdot 10^{-04}$	0.542	0.540	$2.11 \cdot 10^{-04}$		0
Expert 11	$9.97 \cdot 10^{-03}$	0.554	0.514	$5.12 \cdot 10^{-03}$		$1.41 \cdot 10^{-02}$
Expert 12	$4.67 \cdot 10^{-17}$	1.239	1.219	$5.69 \cdot 10^{-17}$		0
Expert 13	$3.81 \cdot 10^{-06}$	0.963	0.930	$3.55 \cdot 10^{-06}$		0
IWDM	$5.57 \cdot 10^{-01}$	0.475	0.492	$2.74 \cdot 10^{-01}$		$7.57 \cdot 10^{-01}$

Table E.20: Expert performance based on Item Weights and the optimized Item Weight Decision Maker. Scores are based on peakload price assessments.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$7.74 \cdot 10^{-09}$	1.140	1.136	$8.78 \cdot 10^{-09}$		0
Expert 2	$2.76 \cdot 10^{-09}$	0.568	0.591	$1.63 \cdot 10^{-09}$		0
Expert 3	$4.51 \cdot 10^{-09}$	1.105	1.110	$5.01 \cdot 10^{-09}$		0
Expert 4	$9.01 \cdot 10^{-04}$	0.567	0.547	$4.93 \cdot 10^{-04}$		0
Expert 5	$7.14 \cdot 10^{-13}$	1.466	1.456	$1.04 \cdot 10^{-12}$		0
Expert 6	$9.70 \cdot 10^{-10}$	0.779	0.830	$8.05 \cdot 10^{-10}$		0
Expert 7	$3.63 \cdot 10^{-18}$	0.785	0.772	$2.80 \cdot 10^{-18}$		0
Expert 8	$1.27 \cdot 10^{-09}$	0.691	0.693	$8.80 \cdot 10^{-10}$		0
Expert 9	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$
Expert 10	$3.90 \cdot 10^{-04}$	0.542	0.540	$2.12 \cdot 10^{-04}$		0
Expert 11	$9.97 \cdot 10^{-03}$	0.554	0.514	$5.12 \cdot 10^{-03}$		0
Expert 12	$4.67 \cdot 10^{-17}$	1.239	1.219	$5.69 \cdot 10^{-17}$		0
Expert 13	$3.81 \cdot 10^{-06}$	0.963	0.930	$3.55 \cdot 10^{-06}$		0
IWDM	$1.29 \cdot 10^{-01}$	0.611	0.643	$8.30 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$

Table E.21: Expert performance based on Item Weights and the Item Weight Decision Maker with $\alpha = 0.1292$. Scores are based on peakload price assessments.

F

Calibration Questions second elicitation

We measure expert performance and quantify their uncertainty using seed variables. We use calibration questions to objectively evaluate the uncertainty assessments of the experts. There were 21 calibration questions in total.

1. According to Centraal Bureau Statistiek (CBS), electricity import from Germany to the Netherlands decreased in 2020 compared to 2019. What was the percentage decrease in 2020 compared to 2019 concerning electricity import to the Netherlands from Germany according to CBS?

2. According to DNV GL's 'Energy Transition Outlook, 2020 A global and regional forecast to 2050', the share of transport, i.e. electrical vehicles, in the total electricity demand will grow significantly in the upcoming decades. In 2018 this share equaled 1.1%. The Dutch government aims at 100% of new passenger car sales to be electric by 2030. What is the share of transport in the total electricity demand estimated to be in 2050 in the Netherlands according to DNV GL?

3. According to the National Climate Agreement, the Netherlands has set a goal to reduce CO2-emmision significantly by 2030. Suppose the reductions are calculated in comparison to the emission values of 1990. What is the estimated decrease, in percentage, of CO2 emissions, according to the Government of the Netherlands?

4. According to the Planbureau voor de Leefomgeving's (PBL) Netherlands Climate and Energy Outlook 2020, the average emission prices in the Netherlands increased with 212,5% between the beginning of 2015 and the end of 2019. According to the same source, the emission prices will continue to rise in the upcoming decades. What is the estimated percentage increase in 2030 compared to 2015 according to their projection? Note that we are comparing the average emission year price of 2030 with the average emission year price of 2015.

5. Coal-fired power plants will slowly phase out during the upcoming decades. The Netherlands currently has 5 coal-fired power plants. How many power plants are estimated to be left in 2026 in the Netherlands, according to Montels 2019 prediction?

6. Electricity consumption in the Netherlands has grown rapidly in the last two decades. Between 1990 and 2008 energy consumption grew from 77.5 TWh to 118 TWh, according to the International Energy Agency (IEA). What was the percentage increase or decrease in consumption in 2019 when compared to 2008 according to the IEA? (Please denote a decrease with ' - ').

7. In 2019, the total offshore wind capacity of the Netherlands was equal to 0.9 GW (Government of the Netherlands). Based on the Dutch offshore wind tender schedule, what is the projected increase in percentages of offshore wind capacity (in GW) in the Netherlands in 2030 compared to the offshore wind capacity in 2019 according to the Government of the Netherlands?

8. The average year temperature in the Bilt, the Netherlands, has increased with 1.7 degrees Celsius between 1952 and 2017 (CBS). In 2017 the average year temperature measured in the Bilt equaled 10,93 degrees Celsius. What was the percentage increase or decrease in temperature in 2020 compared to 2017 according to CBS data?

9. The Dutch government set out to achieve an onshore wind capacity of 6 GW in 2020 (Rijksoverheid). What was the actual onshore wind capacity in the Netherlands by the end of 2020 according to the Government of the Netherlands?

10. What was the absolute difference in euro's between the average baseload price in May and the average baseload price in December?

11. The Netherlands has a large base of gas-fired power plants for district heating, industrial heating and horticultural heating. What was the capacity of these plants in 2019 in GW according to DNV GL?

12. The Netherlands has electricity interconnections with Germany, Belgium, Norway and Great-Britain (Centraal Bureau Statistiek, CBS). This structure enables electricity to flow between electrical grids. A connection to Denmark was completed in 2019. What is the capacity of this connection in MW?

13. What was the absolute difference in euro's between the average peakload price in May and the average peakload price in December?

14. What was the absolute difference in euro's between the average peakload price in May and the average baseload price in May?

15. What was the absolute difference in euro's between the average peakload price in December and the average baseload price in December?

16. What was the difference between the average 2019 price of Dutch Wind Guarantee of Origins and the price of European Wind Guarantee of Origins?

17. What was the percentage share of coal and oil in the Dutch generation mix in 2019 concerning electricity generation according to the International Energy Agency?

18. What was the percentage share of natural gas in the Dutch generation mix in 2019 concerning electricity generation according to the International Energy Agency?

19. What was the percentage share of wind in the Dutch generation mix in 2019 concerning electricity generation according to the International Energy Agency?

20. What was the percentage share of solar PV in the Dutch generation mix in 2019 concerning electricity generation according to the International Energy Agency?

21. With the growth of demand, electricity production increased as well. Between 2000 and 2018 the production of electricity* in the Netherlands grew from 90 TWh to 114 TWh (IEA). Since 2000, the Netherlands has been a net importer of electricity on a yearly basis. How much electricity, in TWh, did the Netherlands import on average, on a year basis, between 2000 and 2020?

G

Range graphs based second elicitation

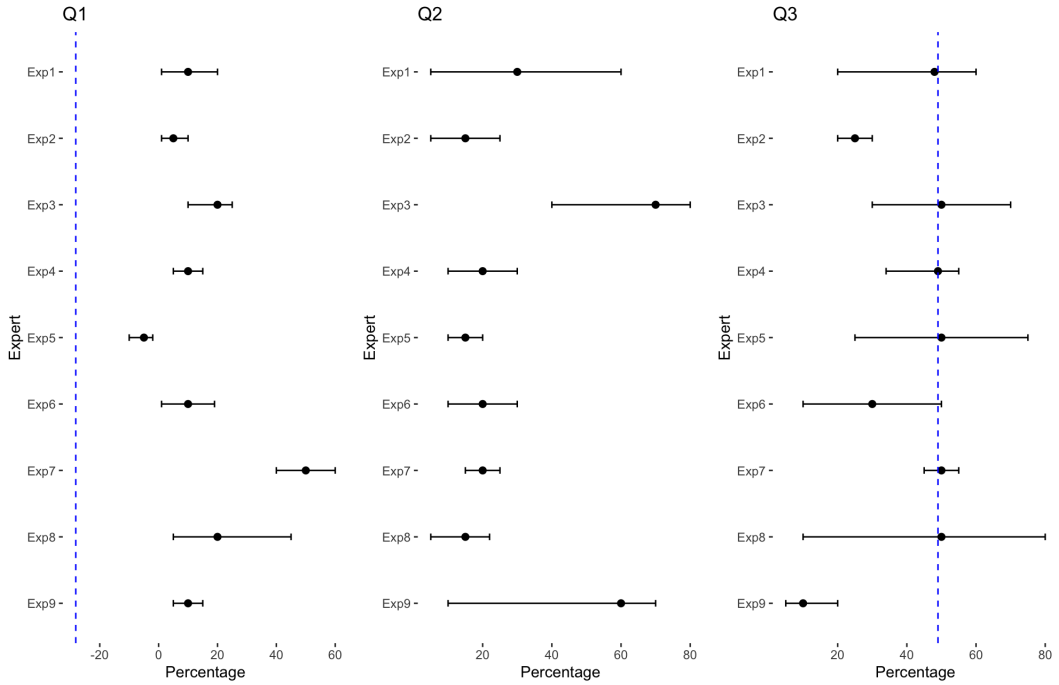

Figure G.1: Experts assessments per question. Seed variables 1, 2 and 3 are presented here. The blue dashed line represents the realization value. The realization to the second question is omitted due to confidentiality regarding the source.

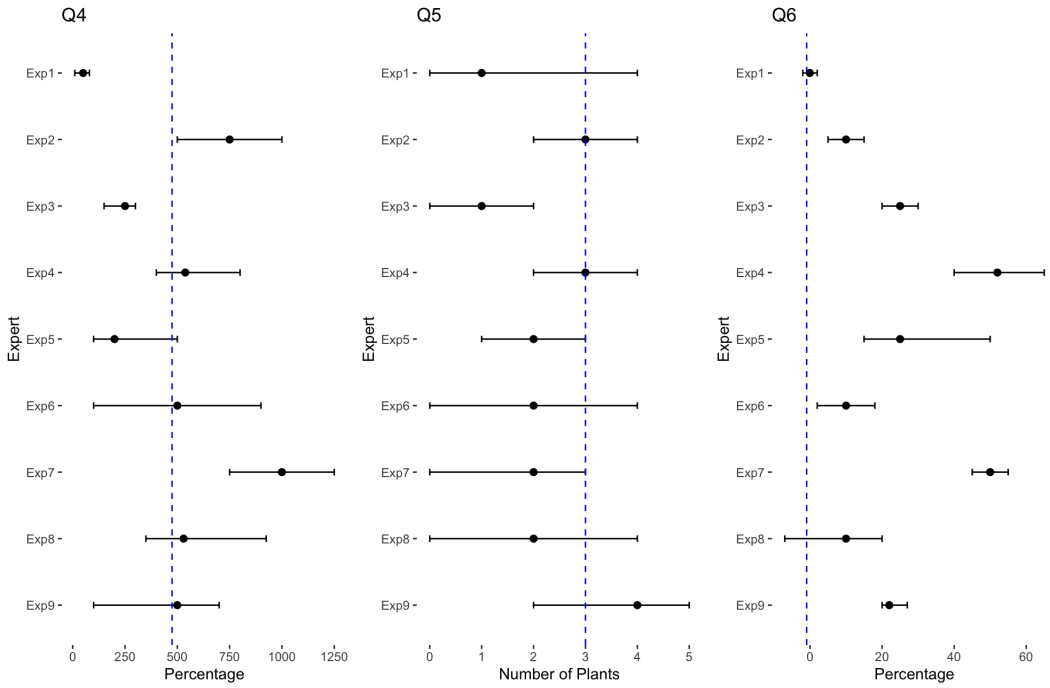

Figure G.2: Experts assessments per question. Seed variables 4, 5 and 6 are presented here. The blue dashed line represents the realization value.

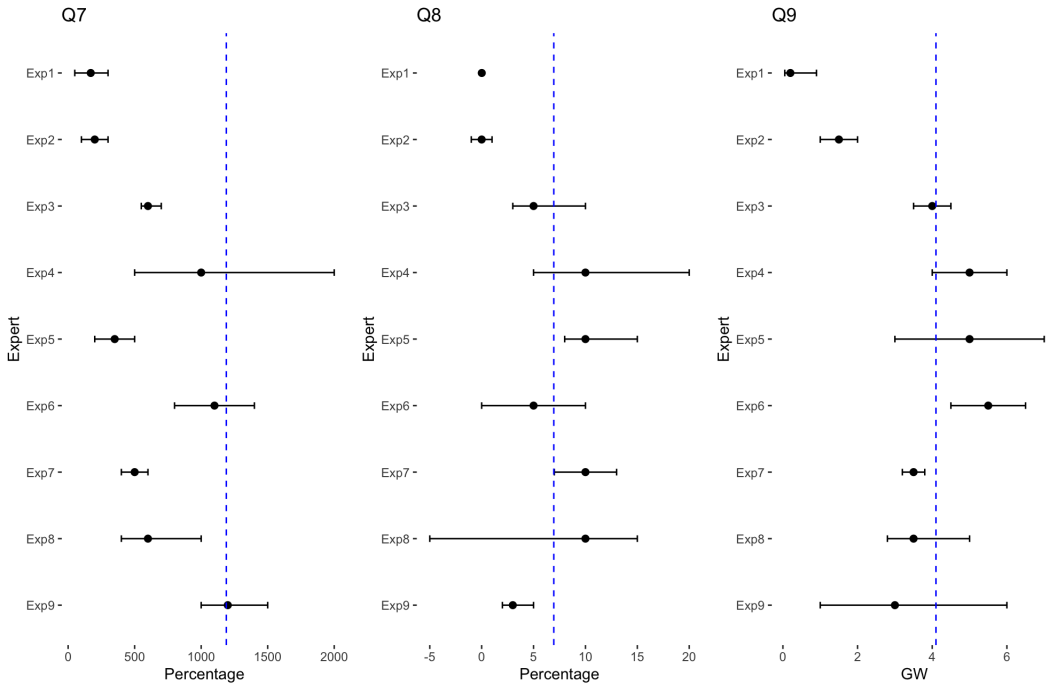

Figure G.3: Experts assessments per question. Seed variables 7, 8 and 9 are presented here. The blue dashed line represents the realization value.

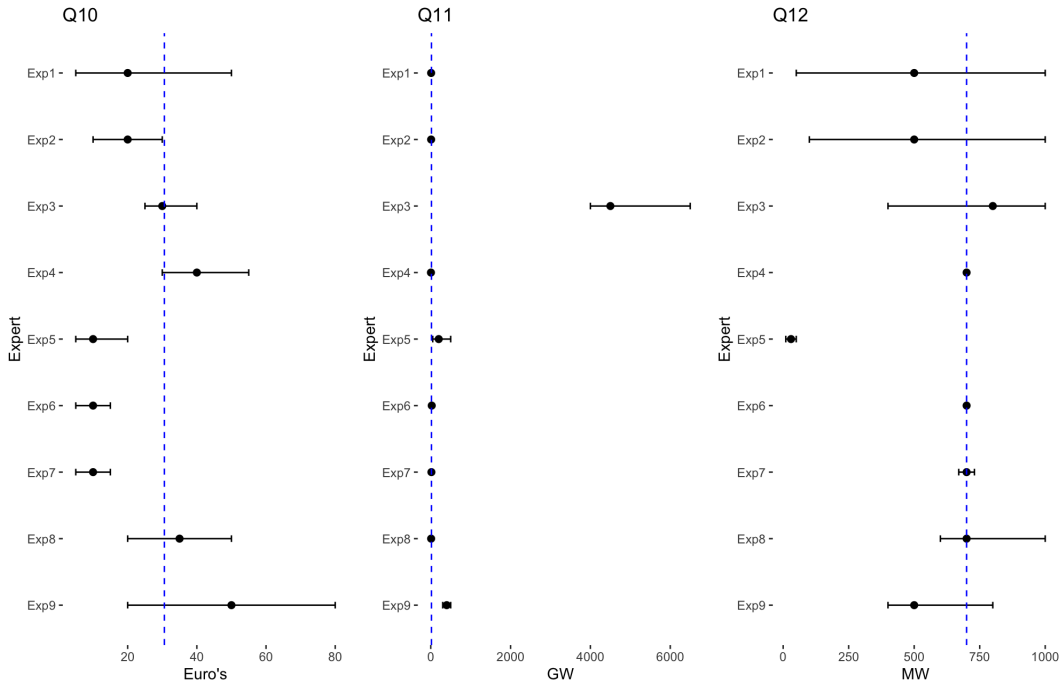

Figure G.4: Experts assessments per question. Seed variables 10, 11 and 12 are presented here. The blue dashed line represents the realization value.

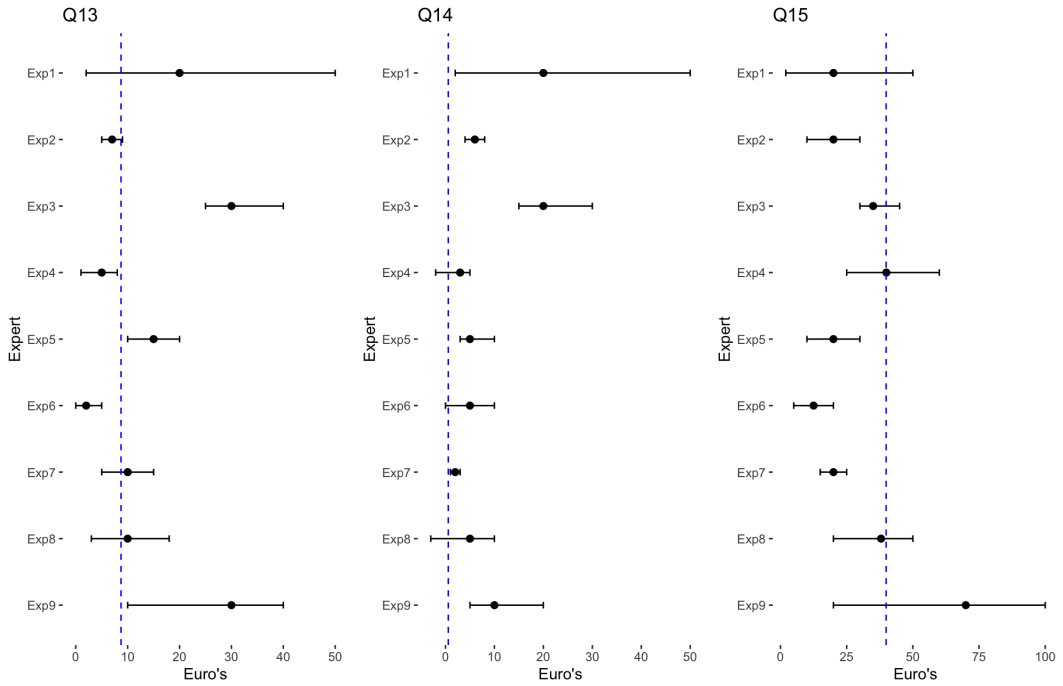

Figure G.5: Experts assessments per question. Seed variables 13, 14 and 15 are presented here. The blue dashed line represents the realization value.

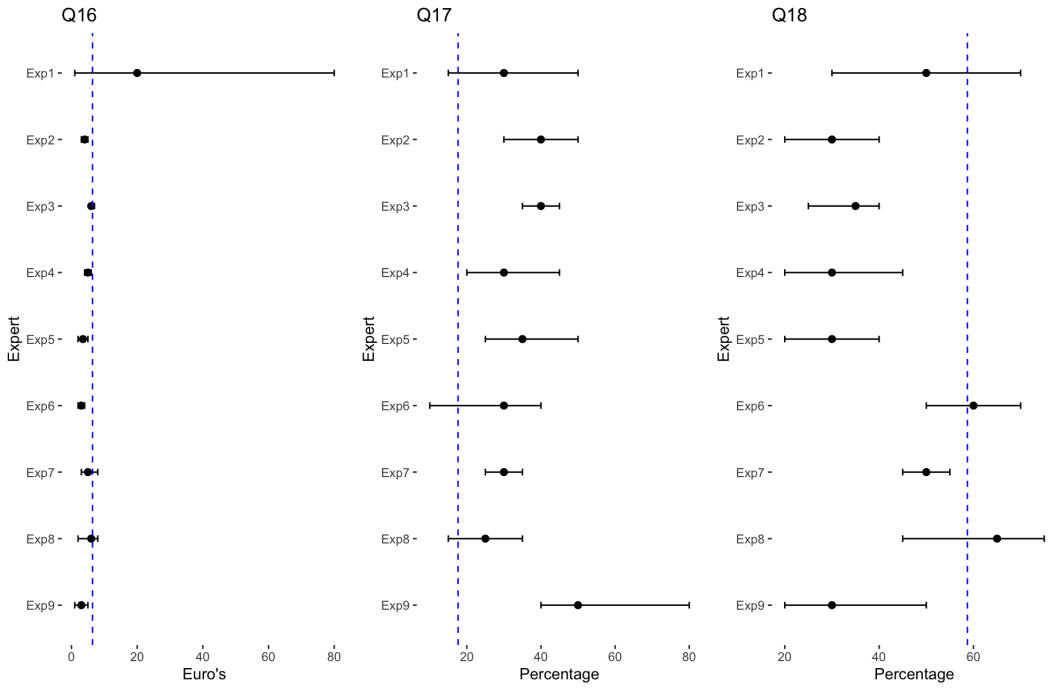

Figure G.6: Experts assessments per question. Seed variables 16, 17 and 18 are presented here. The blue dashed line represents the realization value.

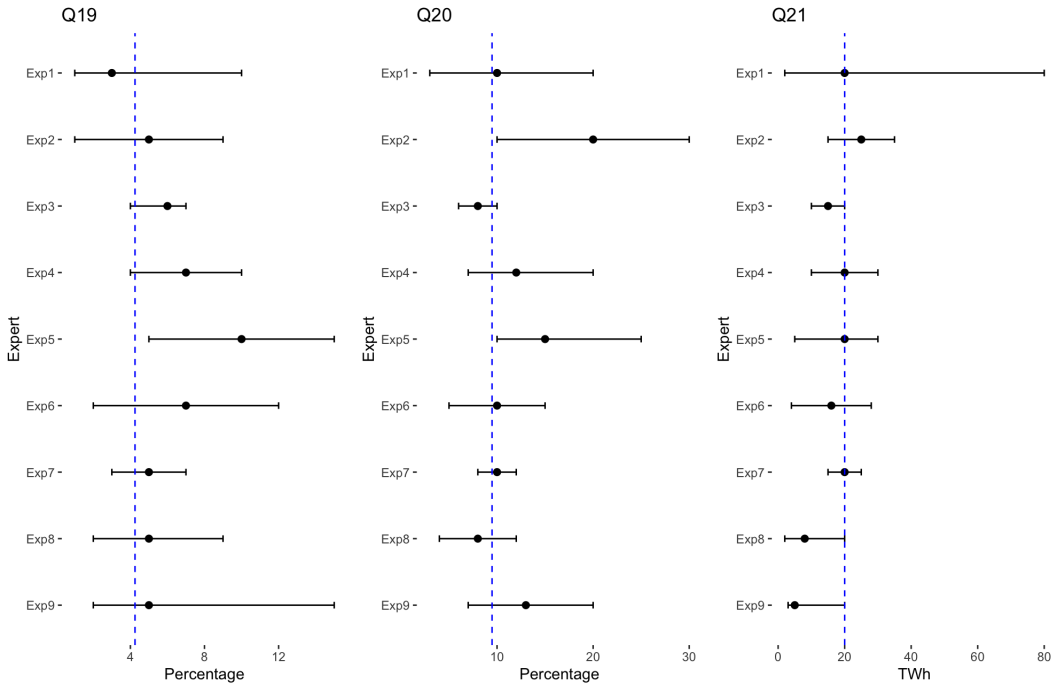

Figure G.7: Experts assessments per question. Seed variables 19, 20 and 21 are presented here. The blue dashed line represents the realization value.

H

Expert performance tables second elicitation

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$	$7.66 \cdot 10^{-05}$	$5.48 \cdot 10^{-05}$
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$	$8.33 \cdot 10^{-01}$	$5.96 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$	$2.72 \cdot 10^{-05}$	$1.95 \cdot 10^{-05}$
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$	$4.13 \cdot 10^{-02}$	$2.96 \cdot 10^{-02}$
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$	$9.23 \cdot 10^{-10}$	$6.61 \cdot 10^{-10}$
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$	$2.10 \cdot 10^{-04}$	$1.50 \cdot 10^{-04}$
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$	$1.42 \cdot 10^{-05}$	$1.02 \cdot 10^{-05}$
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$	$1.13 \cdot 10^{-9}$	$8.12 \cdot 10^{-10}$
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$	$1.26 \cdot 10^{-01}$	$9.00 \cdot 10^{-02}$
GWDM	$5.38 \cdot 10^{-01}$	0.555	0.628	$3.38 \cdot 10^{-01}$		$2.84 \cdot 10^{-01}$

Table H.1: Expert performance based on Global Weights and the Global Weight Decision Maker.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$	0	0
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$	1	$5.00 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$	0	0
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$	0	0
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$	0	0
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$	0	0
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$	0	0
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$	0	0
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$	0	0
GWDM	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.10 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$

Table H.2: Expert performance based on Global Weights and the optimized Global Weight Decision Maker.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$	0	0
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$	$8.69 \cdot 10^{-01}$	$5.01 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$	0	0
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$	0	0
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$	0	0
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$	0	0
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$	0	0
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$	0	0
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$	$1.31 \cdot 10^{-01}$	$7.56 \cdot 10^{-02}$
GWDM	$9.23 \cdot 10^{-01}$	0.578	0.652	$6.02 \cdot 10^{-01}$		$4.24 \cdot 10^{-01}$

Table H.3: Expert performance based on Global Weights and the Global Weight Decision Maker with $\alpha = 0.09331$.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$	$1.10 \cdot 10^{-01}$	$6.58 \cdot 10^{-05}$
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$	$1.10 \cdot 10^{-01}$	$7.16 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$	$1.10 \cdot 10^{-01}$	$2.34 \cdot 10^{-05}$
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$	$1.10 \cdot 10^{-01}$	$3.55 \cdot 10^{-02}$
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$	$1.10 \cdot 10^{-01}$	$7.94 \cdot 10^{-10}$
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$	$1.10 \cdot 10^{-01}$	$1.80 \cdot 10^{-04}$
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$	$1.10 \cdot 10^{-01}$	$1.22 \cdot 10^{-05}$
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$	$1.10 \cdot 10^{-01}$	$9.75 \cdot 10^{-10}$
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$	$1.10 \cdot 10^{-01}$	$1.08 \cdot 10^{-01}$
EWDM	$3.96 \cdot 10^{-01}$	0.312	0.351	$1.39 \cdot 10^{-01}$		$1.40 \cdot 10^{-01}$

Table H.4: Expert performance based on Equal Weights and the Equal Weight Decision Maker.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$		$4.11 \cdot 10^{-05}$
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$		$4.47 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$		$1.46 \cdot 10^{-05}$
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$		$2.21 \cdot 10^{-02}$
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$		$4.95 \cdot 10^{-10}$
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$		$1.12 \cdot 10^{-04}$
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$		$7.61 \cdot 10^{-06}$
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$		$6.08 \cdot 10^{-10}$
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$		$6.74 \cdot 10^{-02}$
IWDM	$9.23 \cdot 10^{-01}$	0.694	0.799	$7.37 \cdot 10^{-01}$		$4.64 \cdot 10^{-01}$

Table H.5: Expert performance based on Item Weights and the Item Weight Decision Maker.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$		0
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$		$4.47 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$		0
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$		$2.21 \cdot 10^{-02}$
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$		0
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$		0
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$		0
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$		0
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$		$6.74 \cdot 10^{-02}$
IWDM	$9.23 \cdot 10^{-01}$	0.694	0.800	$7.38 \cdot 10^{-01}$		$4.64 \cdot 10^{-01}$

Table H.6: Expert performance based on Item Weights and the optimized Item Weight Decision Maker.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$		0
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$		0
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$		0
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$		0
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$		0
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$		0
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$		0
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$		0
IWDM	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.10 \cdot 10^{-01}$		$5.00 \cdot 10^{-01}$

Table H.7: Expert performance based on Item Weights and the Item Weight Decision Maker with $\alpha = 0.6546$.

ID	Calibration Score	Information score all questions	Information score seed questions	Combined Score	Normalized Weight	Normalized Weight with DM
Expert 1	$6.32 \cdot 10^{-05}$	1.039	1.033	$6.53 \cdot 10^{-05}$		0
Expert 2	$6.55 \cdot 10^{-01}$	0.936	1.085	$7.11 \cdot 10^{-01}$		$5.06 \cdot 10^{-01}$
Expert 3	$1.25 \cdot 10^{-05}$	1.800	1.855	$2.32 \cdot 10^{-05}$		0
Expert 4	$2.08 \cdot 10^{-02}$	1.404	1.696	$3.52 \cdot 10^{-02}$		0
Expert 5	$6.20 \cdot 10^{-10}$	1.214	1.270	$7.87 \cdot 10^{-10}$		0
Expert 6	$1.16 \cdot 10^{-04}$	1.363	1.541	$1.79 \cdot 10^{-04}$		0
Expert 7	$8.77 \cdot 10^{-06}$	1.310	1.381	$1.21 \cdot 10^{-05}$		0
Expert 8	$6.20 \cdot 10^{-10}$	1.259	1.559	$9.67 \cdot 10^{-10}$		0
Expert 9	$9.33 \cdot 10^{-02}$	0.984	1.149	$1.07 \cdot 10^{-01}$		$7.64 \cdot 10^{-02}$
IWDM	$7.07 \cdot 10^{-01}$	0.707	0.813	$5.86 \cdot 10^{-01}$		$4.18 \cdot 10^{-01}$

Table H.8: Expert performance based on Item Weights and the Item Weight Decision Maker with $\alpha = 0.09331$.

I

Training as coded in Minerva for second elicitation

WELCOME!

Welcome to a structured expert judgment project focusing on electricity spot price predictions! You will participate along with other experts in an expert elicitation. During the study, you are asked to answer questions concerning the Dutch electricity market. That is, questions concerning energy conservation, energy demand developments, the development of the emissions prices, capacity cross-boarder, the energy mix and its development, flow based market coupling etc. The questions concern historical data, but also projections about the future given by various sources.

Your assessments, along with other experts' assessments will be objectively evaluated and will be included in a mathematical model to provide the best possible prediction. During this training, you will learn how to make the assessments, how are your assessments used in the model, as well as how to use the functionalities of this platform. Practical matters are also included and for any question you can of course reach us at a.d.s.bachasingh@student.tudelft.nl.

Structured expert judgment is an accepted tool in risk and decision analysis for supplementing data shortfalls, quantifying uncertainty and building rational consensus. It has been used in studies sponsored by the European Union, the US NOAA, EPA and CDC, Health Canada, Bill and Melinda Gates Foundation, Shell, among many others, to characterize uncertainty in a wide variety of relationships not amenable to repeated experimentation. To pick a few examples, these include the effects of medical procedures, risks from nuclear power plants, and risks of invasive species. Structured expert judgment can also be used in unforeseen and highly uncertain circumstances, such as crises.

QUESTIONS AND ASSESSMENTS

You will provide assessments for uncertain quantities. The objective is to provide best estimates, along with a characterization of the uncertainty around the estimates. The uncertainty is described by a lower and an upper bound. The lower bound is given by a 5% quantile, whereas the upper bound is given by a 95% quantile of your uncertain distribution. The best estimate is given by the 50% quantile, or the median of the distribution. Each of the question you will answer during this project will have this format. Let's take, for example, the following question:

Dutch electricity export to Germany and Belgium increased in 2020 when compared to 2019 according to Centraal Bureau Statistiek. What was the percentage increase in 2020 when compared to 2019, concerning electricity export from the Netherlands to Belgium according to Centraal Bureau Statistiek?

5%	50%	95%

Suppose that the answer to the question is unknown at the moment an expert provides uncertainty assessments. The true value (or the realization) is therefore unknown for the expert.

- The 50% quantile is that number for which the expert judges the chance ½ that the true value is above or below. We refer to this quantity as the best estimate.
- The 5% quantile tile is that number for which the expert judges the chance that the true value is BELOW the assessment to be 0.05 and the chance that the true value is ABOVE the assessment to be 0.95. We refer to this quantity as the lower bound.
- The 95% quantile is that number for which the chance that the true value is BELOW the assessment to be 0.95, and the chance that the true value is ABOVE the assessment to be 0.05. We refer to this quantity as the upper bound.

EXAMPLE

Suppose the expert provided the following assessments:

Dutch electricity export to Germany and Belgium increased in 2020 when compared to 2019 according to Centraal Bureau Statistiek. What was the percentage increase in 2020 when compared to 2019, concerning electricity export from the Netherlands to Belgium according to Centraal Bureau Statistiek?

5%	50%	95%
3.2%	15.4%	32.4%

According to expert's assessments,

- the true value is equally likely to be above or below 15.4%
- there is a 90% chance that the true value lies between 3.2% and 32.4%
- there is a 5% chance that the true value is smaller than 3.2%
- there is a 5% chance that the true value is higher than 32.4%

Note that the assessments always need to be in strictly ascending order: 5% quantile < 50% quantile < 95% quantile. If, by chance, your assessments will not meet this constraint, you will receive a notification.

VALIDATING ASSESSMENTS

Suppose the expert provided the following assessments:

Dutch electricity export to Germany and Belgium increased in 2020 when compared to 2019 according to Centraal Bureau Statistiek. What was the percentage increase in 2020 when compared to 2019, concerning electricity export from the Netherlands to Belgium according to Centraal Bureau Statistiek?

5%	50%	95%
3.2%	6.9%	10.8%

A good probability assessor is one whose assessments capture the true values with the long run correct relative frequencies (statistically accurate), with distributions that are as narrow as possible (informative). Informativeness is gauged by 'how far apart the percentiles are' relative to an appropriate background.

Measuring statistical accuracy requires the true values for a set of assessments. The true value for the above question is 11.69%. It falls above the 95% quantile. If the expert's assessments are statistically accurate, then in the long run, 5% of the answers should fall within this inter-percentile interval. Similarly, 90% of the answers should fall between the 5% quantile and the 95% quantile.

In gauging overall performance, statistical accuracy is more important than informativeness. Non-informative but statistically accurate assessments are useful, as they sensitize us to how large the uncertainties may be; highly informative but statistically very inaccurate assessments are not useful. Do not shy away from wide distributions if that reflects your real uncertainty.

If you consider you have little knowledge about an item, this fact by itself does NOT disqualify you as an uncertainty assessor. Knowing little means that your percentiles should be 'far apart'. If other experts are

more informative, without sacrificing statistical accuracy, then they will exert more influence on the decision maker. But if there are no statistically accurate experts with more informative assessments, then the uninformative assessments accurately depict the uncertainty. That in itself is VERY important information.

We will use the statistical accuracy and informativeness of your assessments to determine performance-based weights, which will be used to aggregate all the experts' assessments into final assessments.

TIME TO PRACTICE

What was the total offshore wind capacity of the Netherlands in 2019 in GW according to the Dutch Government?

5%	50%	95%

According to the Planbureau voor de Leefomgeving's (PBL) Netherlands Climate and Energy Outlook 2020, the emission prices in the Netherlands increased significantly between 2015 and 2019. What was the average percentage increase in 2019 compared to 2015?

5%	50%	95%

What was the percentage share of Hydro in the European* generation mix in 2018 concerning electricity generation according to the International Energy Agency? *Germany, France, UK, Italy, Turkey, Spain, Poland, Ukraine, Netherlands, Belgium, Sweden, Czech Republic, Finland, Romania, Austria, Norway, Belarus, Hungary, Switzerland, Greece, Israel, Portugal, Bulgaria, Slovak Republic, Denmark, Serbia, Ireland, Croatia, Lithuania, Bosnia and Herzegovina, Slovenia, Estonia, Iceland, Latvia, Moldova, Luxembourg, Kosovo, North Macedonia, Albania, Cyprus, Montenegro, Malta, Gibralta.

5%	50%	95%

What was the percentage share of wind and solar in the European* generation mix in 2018 concerning electricity generation according to the International Energy Agency? *Germany, France, UK, Italy, Turkey, Spain, Poland, Ukraine, Netherlands, Belgium, Sweden, Czech Republic, Finland, Romania, Austria, Norway, Belarus, Hungary, Switzerland, Greece, Israel, Portugal, Bulgaria, Slovak Republic, Denmark, Serbia, Ireland, Croatia, Lithuania, Bosnia and Herzegovina, Slovenia, Estonia, Iceland, Latvia, Moldova, Luxembourg, Kosovo, North Macedonia, Albania, Cyprus, Montenegro, Malta, Gibralta.

5%	50%	95%

WANT TO PRACTICE MORE?

How much has the import of electricity in the Netherlands grown in 2019 compared to 2014 according to the International Energy Agency?

5%	50%	95%

The following questions concern the 2019 APX electricity spot prices published on Epexspot in 2019. Here, the peak hours concern all hours between 08:00 and 20:00. Thus, peakload prices refer to the average of spot prices between 08:00 and 20:00. The baseload price refers to the average of spot prices between 00:00 and 23:59.

What was the absolute difference in euro's between the average baseload price in January and the average baseload price in June?

5%	50%	95%

What was the absolute difference in euro's between the average peakload price in January and the average peakload price in June?

5%	50%	95%

What was the absolute difference in euro's between the average peakload price in January and the average baseload price in January?

5%	50%	95%

What was the absolute difference in euro's between the average peakload price in June and the average baseload price in June?

5%	50%	95%

PRACTICAL MATTERS

You are now ready to start providing assessments. Please check the 'Settings' of the project for more details about the questions. Also, please check your email regularly. You will receive reminders to provide assessments, but also information about the project. Once more, if you have questions or encounter issues of any kind with the platform, contact us at a.d.s.bachasingh@student.tudelft.nl.

ENJOY BEING AN EXPERT FOR THIS PROJECT!