University of Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme/specialisation: Master of Science - Industrial Economics	**Spring semester, 2020** **Open**
Author: Jawvanan Jehan Jørgen Storsveen	**Signature:** Jawvanan Jehan Jørgen Storsveen
Faculty supervisor: Reidar Brumer Bratvold, University of Stavanger	
Title of master's thesis: Debiasing Production Forecasts Through Reference Class Forecasting	
Credits: 60	
Keywords: Reference Class Forecasting Production Forecasting Debiasing NCS Uncertainty Calibration	Number of pages: ……73…… + supplemental material/other: ….15…. Stavanger, …30/06/2020……. date/year

Abstract

This thesis investigates past performance of production forecasts provided by operators on the NCS at the time of project sanction. Utilising a dataset comprising annual forecasted and actual production from 1995 to 2017, we demonstrate that operators on the NCS exhibit considerable optimism and overconfidence biases in their production forecasts. To debias these production forecasts, we develop and implement a reference class forecasting (RCF) methodology with the goal of producing well-calibrated forecasts. The debiased forecasts that are generated from this process are evaluated through a series of tests, providing strong evidence for bias reduction and enhanced forecasting performance. Prior to applying RCF adjustments, only 33% of all observations of actual production in the first six years fall within the 80% confidence interval defined by the forecasts. Applying RCF significantly reduces the overconfidence bias as the adjusted 80% confidence interval now captures 77% of the actual production levels. Moreover, RCF increases the fraction of fields whose actual production exceed the P50 estimate from 37% to 47%, implying reduced optimism.

Acknowledgement

This thesis concludes our journey towards graduation in the Industrial Economics Master's programme at the University of Stavanger. Working with this project has been both challenging and interesting and would not have been possible without support from others.

First and foremost, we would like to give a special thanks to our institute supervisor professor Reidar B. Bratvold for the opportunity to take part in this interesting study. Writing this thesis would be an infeasible task without his constructive feedback and excellent guidance. We would also like to thank the Norwegian Petroleum Directory for providing the data that enabled this study. The process of gaining knowledge about debiasing production forecasts on the Norwegian continental shelf, have been both a privilege and an awakening experience.

Finally, we would like to express our gratitude to family, friends, and each other for the support throughout the entire writing process, and for making this thesis rewarding.

Contents

List of Figures

List of Tables

Abbreviations

CDF	*Cumulative Density Function*	**PDF**	*Probability Density Function*
CF	*Cash Flow*	**PDO**	*Plan for Development and Operations*
CLT	*Central Limit Theorem*	**PRMS**	*Petroleum Resource Managment System*
F4Y	*First Four Years*	**PV**	*Present Value*
F6Y	*First Six Years*	**RC**	*Reference Class*
FID	*Final Investment Decision*	**RCF**	*Reference Class Forecasting*
FnY	*First Number of Years*	**RMSE**	*Root Mean Squared Error*
GRG	*Generalised Reduced Gradient*	**RSE**	*Root Squared Error*
ISF	*Inverse Survival Function*	**SD**	*Standard Deviation*
LB	*Lower Boundary*	**SEC**	*Securities and Exchange Commision*
ML	*Metalog*	**SF**	*Survival Function*
NCS	*Norwegian Continental Shelf*	**SPT**	*Symmetric Percentile Triplet*
NORSOK	*Norsk Sokkels Konkurransedisposisjon*	**TVM**	*Time Value of Money*
NPD	*Norwegian Petroleum Directorate*	**UB**	*Upper Boundary*
NPV	*Net Present Value*	**VBA**	*Visual Basic for Application*

1 Introduction

In the oil and gas industry, investment decisions require production forecasts. Together with estimates of cost and completion time, these production forecasts are used to formulate value estimates and, in turn, form the basis for deciding if and how fields should be developed. As biased or poorly informed production forecasts may lead to suboptimal decisions and poor capital efficiency, significant resources are devoted to forecasting future production in the oil and gas industry.

Despite its importance, optimistic and overconfident production forecasts is the norm rather than the exception for projects on the Norwegian Continental Shelf (NCS). In fact, an evaluation of forecast performance for development projects on the NCS performed by Bratvold et al. (2020) show that only around 30% of actual production outcomes in the first four years after production start fall within the expected 80% range. Moreover, they found that 84% of actual production outcomes for the same period was lower than the P50 production estimate, implying that production shortfalls are dominating on the NCS.

The key contribution of this thesis is to extend the work of Bratvold et al. in several ways. First, the time period of interest will be expanded to cover the first six years of production. Moreover, instead of solely evaluating the aggregated production within this time period, attention will be directed to each individual year with the goal of answering the following question:

Can reference class forecasting, when applied to each year individually, successfully reduce bias related to optimism and overconfidence?

After presenting findings from a literature study on how production forecasts are generated and on possible causes of underperformance in Section 2, we aim to answer this question by following the procedure presented in Figure 1.1. First, verification of original production forecasts for fields on the Norwegian continental shelf is conducted based on historical production forecasts and reported actual production for 56 fields. This entails data scrubbing and distribution fitting, which is discussed in Section 3 and Section 4, respectively. Next, with intentions of debiasing the original forecasts, a methodology for RCF is developed and implemented in Section 5. Finally, forecast calibration is evaluated relative to perfect calibration and further supported by in-sample and out-of-sample tests.

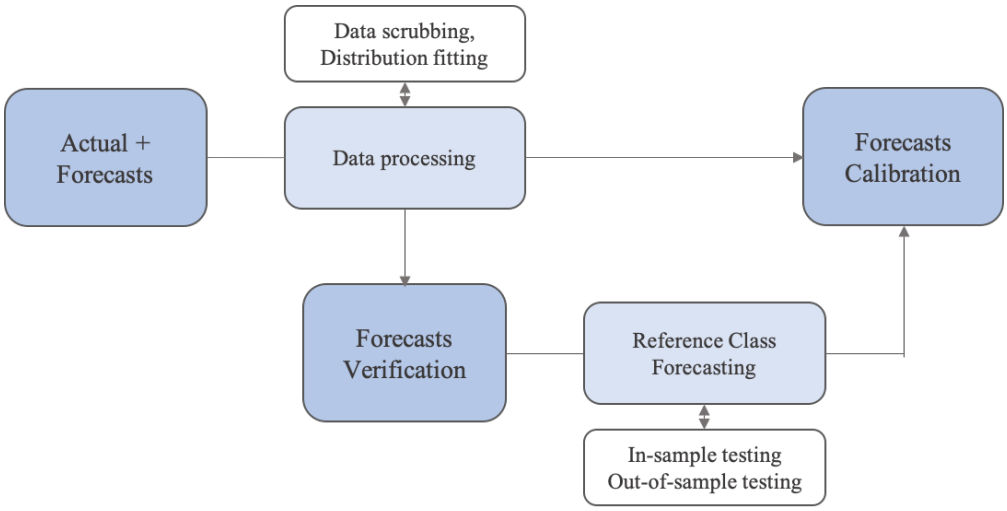


Figure 1.1: *Overview of thesis procedure*

2 Production forecasts in the oil and gas industry

When an investment decision for a project is made, incurring costs are weighed against the project's expected cash flows. This amounts to expected revenue and profitability, which is conventionally assessed by calculating the Internal Rate of Return (IRR) or the project's Net Present Value (NPV). For petroleum development projects, the profitability depends strongly on forecasted production of oil and gas (Meddaugh et al., 2017). Together with estimated costs and completion time, these forecasts represent the core of estimates for future cash flows and are, therefore, central for decision making processes in the oil and gas industry. Production excess or shortfall leads to suboptimal decisions and poor capital efficiency, adversely affecting both companies and shareholders. As a result, generating production forecasts that account for uncertainty related to actual production attainment becomes crucial for providing a well-informed decision-making basis for the final investment decision (FID). This section briefly describes how these forecasts traditionally are generated, evaluate the general performance of today's forecasts, and discuss possible factors that contribute to production shortfalls. For future reference, this will be referred to as underperformance.

2.1 Estimating future oil production

Due to the importance of well-informed production forecasts, companies in the oil and gas industry devote enormous amounts of resources to develop and improve forecasting methods (Nandurdikar et al., 2011). Estimates of future oil production from a particular reservoir are heavily reliant on data acquired from sources like seismic surveys, well logs, drilling, and core samples (PetroWiki, 2020). Knowledge generated by analysing data from these sources is used as input to advanced computer models for reservoir simulation, generally categorised as either static or dynamic models (Yeten et al., 2015). The former generally consists of a stratigraphic framework described by reservoir parameters like porosity and permeability distributions, fluid saturations, rock properties, and fluid contacts. Dynamic models are more advanced and typically comprise upscaled versions of static models. These models include additional input factors such as reservoir pressure, volume and temperature characteristics, and flow rates of the reservoir fluid, thereby acting to coarsen the resolution of the static model. For reliable production forecasts, both static and dynamic models that are representative of the specific reservoir are required.

2.1.1 Describing uncertainty

Approaches for describing uncertainty related to production forecasts generated by reservoir models can broadly be categorized as either deterministic or probabilistic (PetroWiki, 2016). Deterministic models are models where the output is fully determined by the explanatory variables and the initial conditions of these parameters (Rey, 2015). Probabilistic (or stochastic) models, on the other hand, incorporate ranges of values with corresponding probability distributions for each variable (Renard et al., 2013) and, in turn, yields a probability distribution for the model output. Based on the amount of available data and the strength of knowledge judgments, which points to an analyst's ability to produce a reasonable prediction of future production, one may resort to several different approaches for handling subsurface uncertainty. Bentley and Smith (2008) present three contrasting approaches; Rationalist approaches (1), Multiple stochastic approaches (2), and Multiple deterministic approaches (3). The rationalist approach is heavily shifted towards determinism, which is outlined through the presentation of a unique output − a single best guess − that may be accompanied by low and high estimates to account for uncertainty. The multiple stochastic approach probabilistically generates a large number of possible outcomes by assigning probability distributions to each input parameter. Each distribution is constructed from gathered reservoir data and, together, produce a cumulative probability curve for the model output, typically based on a Monte Carlo simulation approach. From this distribution, percentiles like P90, P50 and P10 production estimates may be retrieved (PetroWiki, 2016). For the final approach, multiple deterministic, a smaller number of models that each reflect an explicitly defined physical representation of the reservoir are created. Low, medium and high cases may, then, be retrieved by assigning probabilities to the various outputs.

Despite its importance, uncertainty reflections of production forecasts has received a varying degree of attention in the past. Dating back to the 1980's, production forecasting for major development projects on the NCS was performed following the rationalist approach, i.e. by only expressing forecasted production by a single value. Yearly production forecasts for the anticipated production life of 10 to 30 years was generated by this methodology. Since then, a gradual shift towards probabilistic forecasting methods in the oil and gas industry has occurred, both in study and application. In the 22 year time period from 1995 to 2017, the number of published papers on the topic is found to grow more than 600%, as seen in Figure 2.1 (Bratvold et al., 2020).

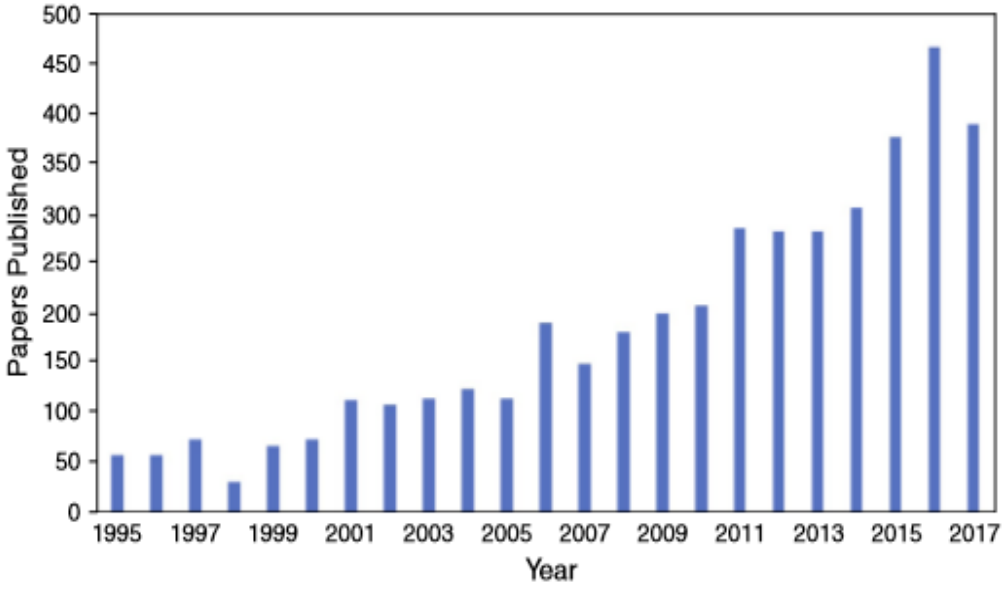


Figure 2.1: *Overview over the growth of published papers on probabilistic forecasting over a 22 year time period (Bratvold et al., 2020)*

In terms of expressing the uncertainty ranges in relation to production forecasts on the NCS, clear guidelines are provided by the Norwegian Petroleum Directorate (NPD). These guidelines are detailed in Table 2.1, emphasising the use of a multiple stochastic approach. This is also in alignment with guidelines provided by the Petroleum Resource Management System (PRMS) and the Securities and Exchange Commission (SEC), both of which describe the reserves and resources by low, medium and high estimates in terms of P90/P50/P10 ranges.

Following the guidelines provided by the NPD, this thesis expresses the low, base and high production estimates by P90, mean and P10 values, respectively. Thus, the following definitions of probabilistic forecasts apply:

- P90: There should be at least a 90% probability that the quantities actually recovered will equal or exceed the low estimate.
- P50: There should be at least a 50% probability that the quantities actually recovered will equal or exceed the best estimate.
- P10: There should be at least a 10% probability that the quantities actually recovered will equal or exceed the high estimate.

Table 2.1: *Overview of the Uncertainty category classifications and explanations provided by the NPD (Norwegian Petroleum Directorate, 2019) (modified)*

Uncertainty Category	Definition	Explanation
Low Estimate	Low estimate of petroleum volumes that are expected to be recovered from a project.	The low estimate must be lower than the base estimate. The probability of being able to recover the indicated estimate or more must be shown (e.g. P90). Compared with the base estimate, the low estimate should express potential negative changes with regard to mapping of the reservoir, reservoir/fluid parameters and/or recovery rate.
Base Estimate	Best estimate of petroleum volumes that are expected to be recovered from a project.	The base estimate must reflect the current understanding of the scope, properties and recovery rate of the reservoir. The base estimate will be calculated using a deterministic or stochastic method. If the base estimate was calculated using a stochastic method, the base estimate shall be stated as the expected value.
High Estimate	High estimate of petroleum volumes that are expected to be recovered from a project.	The high estimate must be higher than the base estimate. The probability of being able to recover the indicated estimate or more must be shown (e.g. P10). Compared with the base estimate, the high estimate should express potential positive changes with regard to mapping of the reservoir, reservoir/fluid parameters and/or recovery rate.

Figure 2.2 graphically illustrates the above definitions, showing the three percentile curves for production estimates and the actual oil production profile for a typical field.

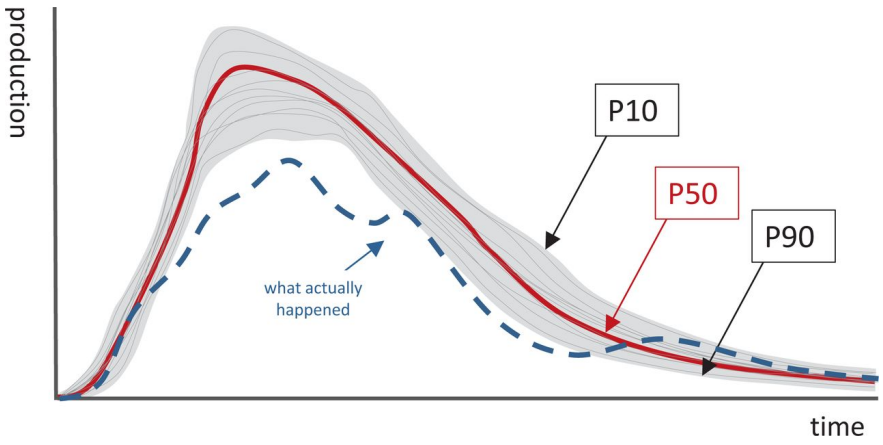


Figure 2.2: *Overview of P90, P50 and P10 cases from probabilistic forecasting from multiple stochastic models (Bentley, 2016)*

2.2 Current production forecast performance

In an ideal scenario, the forecasts fit the actual production profile without any deviations, thereby creating a well-informed basis for field development. Due to the presence of uncertainty, however, the majority of fields on the NCS fail to deliver on forecasted production. This is shown by Bratvold et al. (2020), who recently performed a study on the general forecast performance of fields on the NCS. Comparing historical actual production for the first four years after production start for 32 fields to their respective original production estimates, they found that only 31% of the fields had actual production that fell within the 80% confidence interval defined by the P90 and P10 estimates. Figure 2.3, showing actual production and the mean forecasted production for the same 55 fields investigated by Bratvold et al., further illustrates that production shortfalls in the first years of production has been the rule rather than the exception for fields on the NCS. For the first six years, actual production is seen to fall significantly short of the mean estimate, which is meant to reflect the expected value of future production volumes. Six years after production start, however, one experiences a shift between the actual and estimated production data. At this point, the former exceeds the latter, typically due to reinvestment and implementation of improved recovery methods for production.

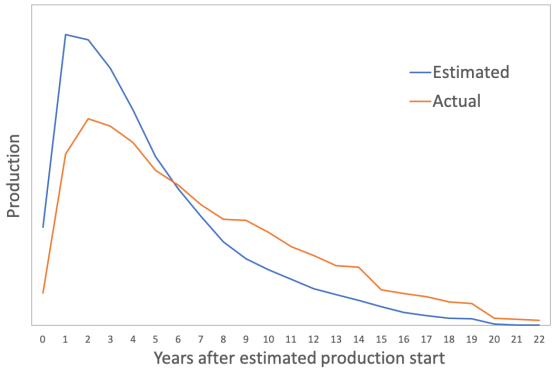


Figure 2.3: *Forecasted vs. Actual production*

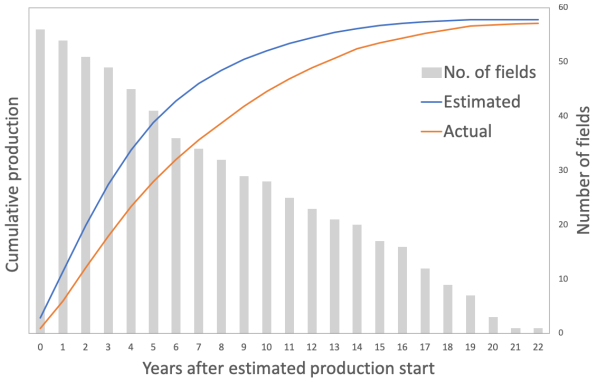


Figure 2.4: *Cumulative Production*

Figure 2.4 shows the cumulative actual and forecasted production for 22 years of production. Although the yearly actual production eventually exceeds forecasted production, and the cumulative actual production after 22 years of production is equal to that of the estimates, emphasis should be put on the NPV of these production volumes. Cash flows in initial years of a project carry more economical weight than those occurring at later stages of the project. This follows from the core principle of finance, stating that a sum of money in the future is worth less than the identical sum today – often referred to as the time value of money (TVM) (Chen, James, 2020). The clear tendency of overestimating production in the early

stages of the field's lifetime may, thus, result in value erosion that is not in line with the expectations of shareholders involved in the project.

2.3 Causes of underperformance

It is at least intuitively obvious that some of the deviation between forecasted and actual production source back to the high degree of uncertainty and complexity related to hydrocarbon production. Reservoirs on the NCS are typically heterogeneous and can reach depths of 4000 meters (Norwegian Petroleum Directorate, 2017), introducing uncertainty to the reservoir properties. Encapsulating the range of variation in all contributing variables and developing well-informed prediction models for 10 to 30 years of future production, thus, requires complex modelling of uncertainty. However, despite significantly increased understanding of uncertainty modelling over the past two decades, production forecasts are just as inaccurate today as they were 20 years ago (Bratvold et al., 2020). This may point to biased production estimates, owing to the various psychological factors presented in this section.

Aside from the forecasts themselves, the Oil and Gas Authority (2017) presents five key contributing reasons for gaps between actual and estimated production in the oil and gas industry. These are; project- and organisation management, front-end loading, execution, and behaviour. Common for all these areas are incurring psychological and hierarchical factors that tend to drag a project beyond its predetermined target on production. Historically, these factors carry an undervalued perception. Rather than acknowledging the presence of biases in production estimates, production shortfalls are often explained by bad luck (Flyvbjerg et al., 2009). Goliat (Kongsnes, 2015), Martin Linge (Stangeland, 2015), Glitne (Norwegian Petroleum Directorate, 2011) and Yme (Skodje and Steneberg, 2011) are all fields that have expressed bad weather as explanation for cost overruns. Other prevalent explanations provided by Norwegian leaders when expressing the reasoning for failing to deliver on time and, in turn, expected production, are lack of quality from suppliers and change of complexity in the reservoir. While not denying the validity of such salient explanations, reported production data for fields on the NCS imply that these excuses may overshadow the presence of psychological biases.

Optimism and overconfidence due to lack of regard to distributable information is argued to be a common judgment trait of the human mind (Kahneman and Tversky, 1977; Kahneman, 1979). The concept of "planning fallacy" was introduced in the same papers, and can be understood as the tendency to believe that your own project will proceed as

planned, despite previous instances of similar projects with comparable scope and magnitude failing to perform according to expectations. A further expansion of the concept, making it applicable for projects in the petroleum industry, includes the underestimation of time and risk, which introduces a potential for production shortfalls and cost overruns. Decision-makers and top management tend to pursue projects that are unlikely to deliver on the trinity of estimated time, cost and returns (Flyvbjerg, 2007b). This tendency is further discussed by Flyvbjerg et al. (2009), deducing two main sources of biased forecast profiles; deception and delusion.

2.3.1 Deception

Deception is the term that is referred to whenever there is an advantage to be gained by a strategic misinterpretation of the project at hand, and relates to motivational bias. A project that falls under this category generally has an augmented perceived potential, typically caused by a principal-agent (P-A) problem where the primary incentives of the parties involved are although in alignment, not necessarily to the same degree (Flyvbjerg et al., 2009). Projects that are big in magnitude and consists of multiple tiers, such as offshore petroleum projects, are susceptible to P-A problems between every two levels of the supply chain.

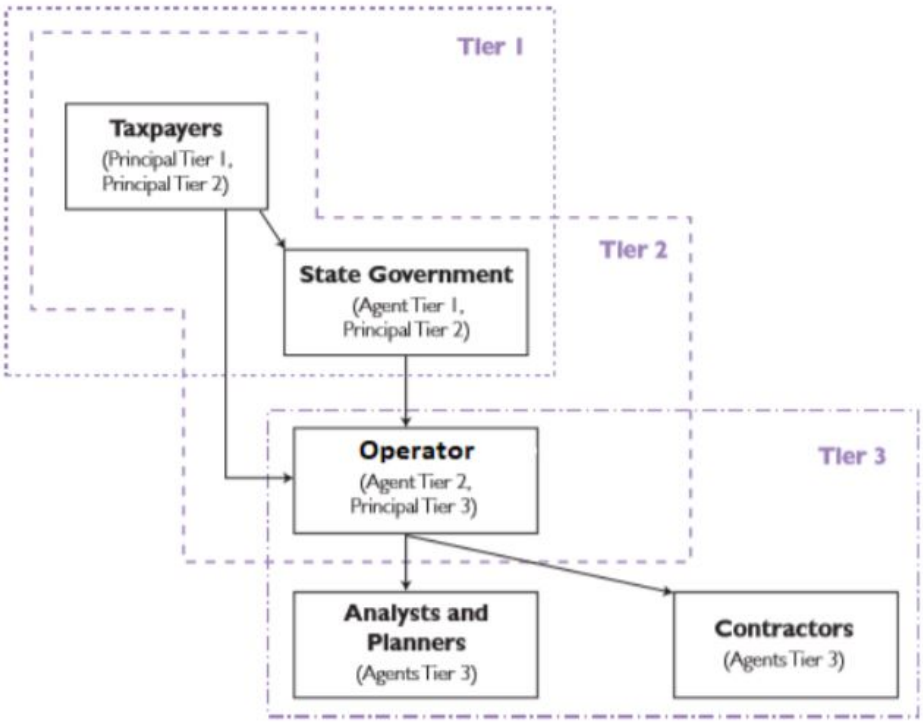


Figure 2.5: *Illustration of P-A tiers for a megaproject (Flyvbjerg et al., 2009)*

Figure 2.5, presents a typical principal agent system. As one proceeds through the entire system, there are clear benefits to gain through a strategic misinterpretation of the incentives that relates a specific principal to another specific agent of a megaproject. For instance, the first tier relating the taxpayers to the state government, the former may expect the latter to maximise the benefits and gains through an economic scope when retaining approval for the PDO. However, the state government may have other interests, such as only approving projects within predetermined limits of climate pollution or aquaculture. The same conflict may also be present between the operator and the government in the second tier. While the operators are responsible for preparing the PDO, utilising resource information provided by the government, one can suspect that the numbers may favour the incentives of the operators, while barely satisfying the needs of the government. Obviously, these conflicting interests affect the results that one perceived to gain before initiation. As the system comprise of more tiers, deviations from initial expectations tend to grow correspondingly larger. However, the tiers are necessary for the megaproject to initiate, survive and ultimately deliver.

2.3.2 Delusion

Delusion describes underlying psychological effects that eludes a task performer to underestimate the upcoming workload and relates to cognitive bias. Flyvbjerg et al. (2009) emphasise that managers often make delusional and highly optimistic decisions, rather than basing decisions on a rational weighting of gains, losses and probabilities. Put in other terms, delusion is an involuntary mistake that forecasters are prone to whenever estimates are made. In hindsight, one finds that many mistakes source back to executives taking an inside view on the decision at hand. Rather than grasping the entire picture of the project with a long-term plan in mind, the focus is directed towards the specifics in a short-sighted scope. As argued by Flyvbjerg (2007b) and Kahneman (1979), this leads to a constant state of planning fallacy where the final output on cost, time and production are far off the initial expectations. Assessments of distributions for variables such as average porosity, net-to-gross, and formation volume factors are exposed to subjectivity and, thus, susceptible to cognitive bias. The problem gets elevated for decisions related to large oil and gas projects, because they are made on the basis of many subjective estimates, all of which are likely to be affected by cognitive bias. Whenever there is a trace of a delusional approach to a petroleum development project, it can be grouped into one or several of four delusional bias categories; information availability, anchoring, overconfidence, and trust heuristics.

Information availability

Information is the foundation of which any decision is made. It is therefore crucial to be conscious about the source and, more importantly, the validity of the information at hand − no information might be better than disinformation. Bratvold et al. (2010) describe the human perception of reality as distorted due to the excess information available. Further, they argue that there is a tendency to drift more towards most recent and vivid information. From a decision-maker perspective in any industry, past proceedings may also have a significant contributing psychological factor for the project at hand. If a project manager has recently been involved in a successful project, he might find it easier to pass on that feeling to upcoming projects. However, this induces a possibility for overconfidence and complacency, in which case the project will be less likely to deliver at planned pace. On the other hand, if a manager is taking up a new project after recently being involved in a failed one, he might lack the confidence to run crucial operation procedures. In turn, this might make him unable or reluctant to pass on crucial information to the right receivers at the right time. Also, when creating a production forecast, technical information is key. The model can only be as good as the information it is built on. Both reservoir data and information from comparable projects are important to consider when generating a well-informed forecasting model.

Anchoring

Anchoring is another consequence of the inside view thinking that leads to optimistic forecasts (Flyvbjerg et al., 2009), and can be understood as the tendency of putting too much trust into base estimates for production forecasts in spite of wide uncertainty ranges (Bentley and Smith, 2008). Once anchored, the willingness to explore uncertainty ranges are sure to diminish, resulting in a prediction model that is overly influenced by the anchor points without enough care for the ranges. Anchoring is therefore a well-understood cognitive behaviour where the resulting estimates are more likely to be over- rather than underconfident (Welsh et al., 2010). Figure 2.6 illustrates a typical case of anchoring in reservoir modelling, showing that although low and high cases are provided, these may also be anchored on the base estimate.

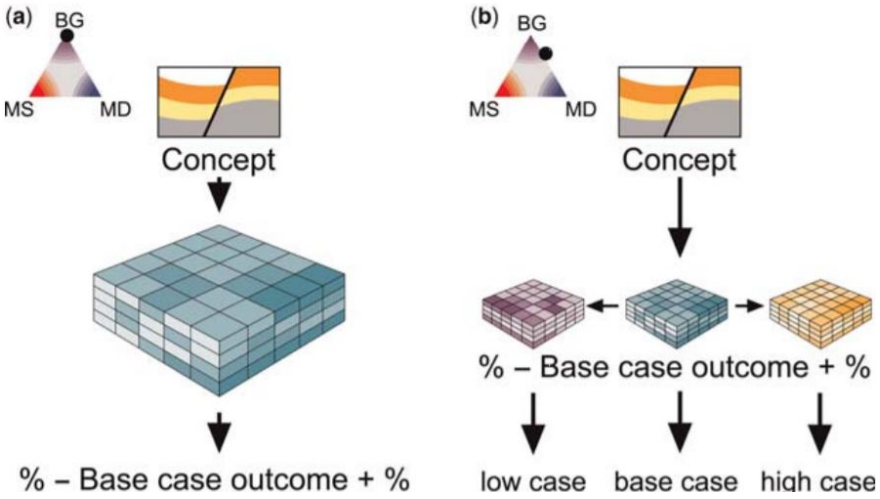


Figure 2.6: *Diagramatic representation of anchoring: (a) the extreme end-member case is the single best guess; (b) even with the addition of a +/- spread, the approach is still anchored on the initial best guess (Bentley and Smith, 2008)*

Overconfidence

Overconfidence is perhaps cognitive bias at its most well-known form. The nature of this bias is to cause an individual to overestimate the strength of knowledge that one possesses. As a result, the bounds of the possibility range for any event or parameter are narrowed. Welsh et al. (2007) investigated the economic impact of overconfidence on large development decisions by assuming a triangular distribution model for the minimum, most likely and maximum values of reservoir parameters like porosity, water saturation, net-to-gross, area, thickness. Their results, which are presented in Figure 2.7, illustrate the clear impact of overestimation on the project NPV.

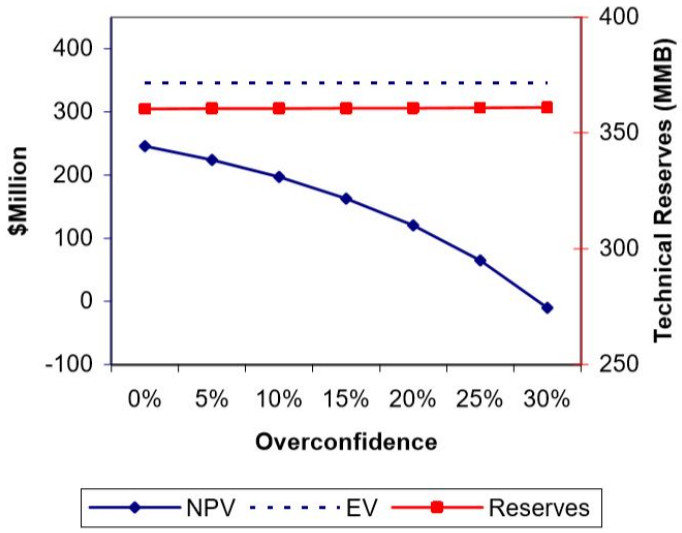


Figure 2.7: *Effect of Overconfidence on NPV (Welsh et al., 2007)*

It is observed that the expected value of the project remains constant at $346 million as the expected value for the input distributions remained constant between each condition. However, the results of the simulation model never reach the expected value of return. The maximum value that the project can attain is an NPV of $246 million. This can be explained by the non-linearity arising from the complexity of the model itself. Further, the results show a clear profile of an accelerating decline in the NPV as the rate of overconfidence steadily increases. At 5% overconfidence rate, the NPV drops to $224 million. Further extrapolation to 30% overconfidence, the project retrieves a negative NPV of -$10 million. This implies that a company that are 30% overconfident about their parametric input values compared to the true underlying uncertainty values would predict an NPV of $246 million, whereas the actual NPV would be -$10 million, resulting in an error of $256 million due to overconfidence. From these results, the impact of overconfidence bias to the potential financial losses are evident and are imperative to be accounted for and reduced to a minimal in any prediction model.

Trust heuristics

The last, and probably most overlooked, delusional bias affecting estimates is trust heuristics, which can be understood as the tendency of managers to rely on the judgment of the most trusted team member(s) when making a decision. By doing so, one may overlook valuable expertise from other team members that might have provided important input to the objective at hand. This contributes to estimation errors in the oil and gas industry, simply by not utilising all expertise knowledge that is available.

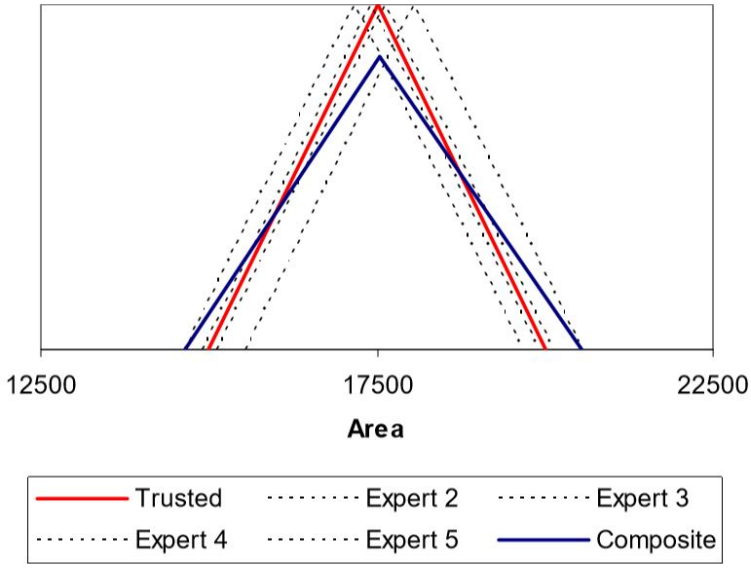


Figure 2.8: *Illustration of the knowledge of 5 different experts and the composite knowledge following a triangular distribution (Welsh et al., 2007)*

Figure 2.8 illustrates the beliefs of five experts on the parametric value of area. As illustrated, the beliefs are triangular distribution PDFs of equal shape, differing only in their ability to reflect the differences between each expert opinion. The red distribution represents that of the most trusted team member, which is noticeably narrower than the composite blue triangular PDF distribution. The difference in range between the red and blue distributions may not withhold crucial information about the true underlying parametric value of the area. This particular, and perhaps crucial, information is not processed if only the expertise provided by the most trusted team member is regarded.

Welsh et al. (2007) present a model that displays the effect of trust heuristics on overconfidence. This model aims to show how multiple experts with varying information input and varying degrees of agreement affects overconfidence. From the results presented in Figure 2.9, it can be observed that including even a single additional expert in the decision making process contributes to reducing the overconfidence by around 5 to 10% on average, depending on whether there is a high or low degree of agreement amongst the experts. Proceeding to add, say, 4 additional experts to the group of decision makers induces an average reduction in overconfidence by 8 to 17%. The potential economic impacts of trust heuristics can be retrieved by comparing Figure 2.9 with Figure 2.7. Assuming the same scenario as was modelled in the discussion of overconfidence, a 5% change in overconfidence equates to an error in calculating the project's NPV of between \$22 and \$75 million, depending on how overconfident the trusted expert was to start with. While this research assumes a similar distribution for the individual knowledge of the experts based on individual assessments rather than consensus, the results are important to not overlook. It is apparent that the oil and gas industry have yet to become better at exploiting the knowledge base of the experts at hand. Improvements on this area may yield significantly better results in the economic and performance portfolios of the companies.

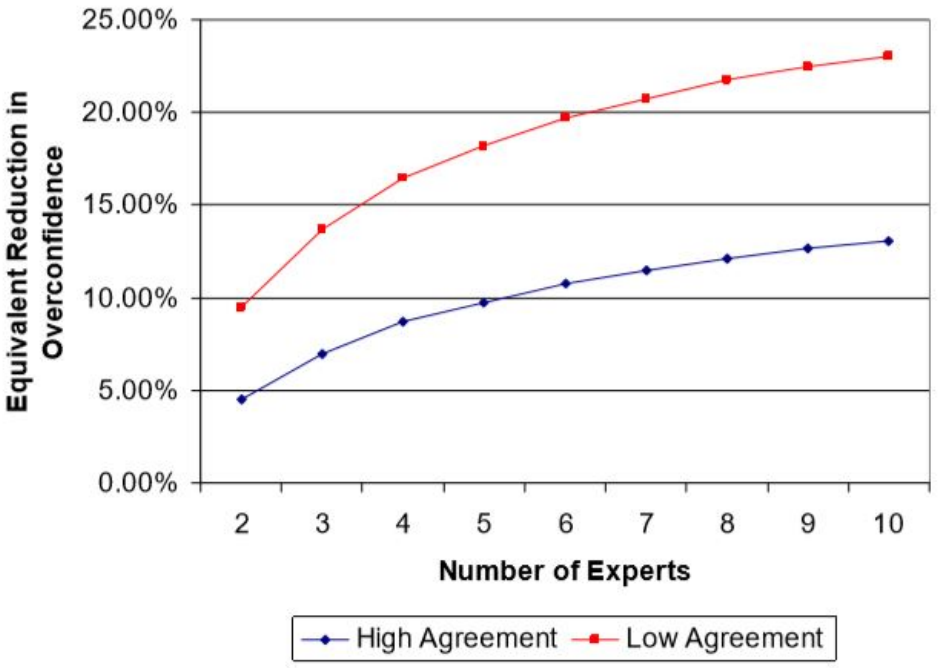


Figure 2.9: *Illustration of the impact of agreement among experts and overconfidence (Welsh et al., 2007)*

As presented in this section, the causes of underperformance are many and bear significant impact when comparing the estimates to the actual production quantity. The NCS production data utilised in this thesis may exhibit one or several of the presented elements that translates to biased forecasts. In this work, focus is directed towards providing relevant discussion on overconfidence and optimism debiasing.

3 Data and data scrubbing

Forecast performance for fields on the Norwegian continental shelf is, in this thesis, investigated based on the same dataset utilised by Bratvold et al. (2020), which was briefly introduced in Section 2.2. While Bratvold et al. consistently focus on aggregated production data for the first four years (F4Y) of production, our work extends this time period to cover the first six years. Moreover, rather than studying the aggregated production data for this time period, attention is directed to each individual year. This section aims to provide a detailed description of the dataset, the operations applied for eliminating the effect of schedule delays, and the process of data scrubbing utilised to filter out unreliable data.

3.1 Data

Evaluating performance of production forecasts on the NCS requires actual and estimated production data at field-level. While actual production data are acquired from the operators' annual reporting to the revised national budget (Norwegian Petroleum Directorate, 2020), estimated production data for fields on the NCS is not public information and, therefore, not easily attainable. However, before a field is approved for development, operators on the NCS are required to submit a report on the Plan for Development and Operations (PDO) to the NPD. Furthermore, it is a prerequisite that the production forecasts supporting the FID is included in this report. Estimated production data provided by operators at the time of FID is acquired through a non-disclosure agreement with the NPD. Consequently, no actual field name with production estimate will be presented. If a field name is used, it is to show public data. Furthermore, axis-values are removed in cases that inherent revelation possibilities of fields that are being discussed.

The dataset provided by the NPD comprises 85 oil and gas fields on the NCS, all approved for development from 1995 to 2017. Excluding fields with either poor or missing data, as well as forecasts for gas, natural gas liquids and condensate production, yields a final dataset with 56 fields. For each of these fields, year-by-year low, medium and high production estimates are provided for their projected lifetime. In total, this adds up to an extensive dataset consisting of 602 production years. Guidelines provided by the NPD suggests that medium estimates should reflect the expected value (mean), while low and high estimates preferably should represent the P90 and P10 values, respectively. Although the early PDO guidelines failed to rigidly specify corresponding probabilities for the low and high estimates (Norwegian Petroleum Directorate, 2000; Ministry of Petroleum and Energy, 2010), no additional information is given to contradict the current guidelines.

Thus, it is assumed that the provided estimates are consistent with the NPD guidelines presented in Section 2.1.1, i.e. that the low, medium and high production estimates reflect P90, mean and P10 values, respectively. Production forecasts with these characteristics are said to be well-calibrated – or unbiased – if; 1) 80% of the actual outcomes lie within the forecasted P90/P10 range, and 2) 50% of the actual observations lie above the mean estimate while the other 50% lie below it (assuming approximately normally distributed data) (Bratvold et al., 2020).

3.1.1 Time shifting the data

Bratvold et al. (2020) found that 17 percent of the fields started production earlier than scheduled, while 69 percent experienced schedule delays. With an average delay of 202 days for development projects on the NCS (Mohus, 2018), schedule delays clearly have ramifications on production shortfalls. As this thesis aims to evaluate the performance of production forecasts in isolation, a process of eliminating the effect of schedule delays is conducted. This entails time shifting the data to the point of actual production start, i.e. setting the time of first oil to year zero, which enables estimated production for year i after estimated production start to be compared to actual production for year i after actual production start. By virtue of this operation, the total number of viable fields is reduced to 54, translating to a substantial reduction in total number of production years from 602 to 548. Figure 3.1 shows the effect of time shifting the actual production data.

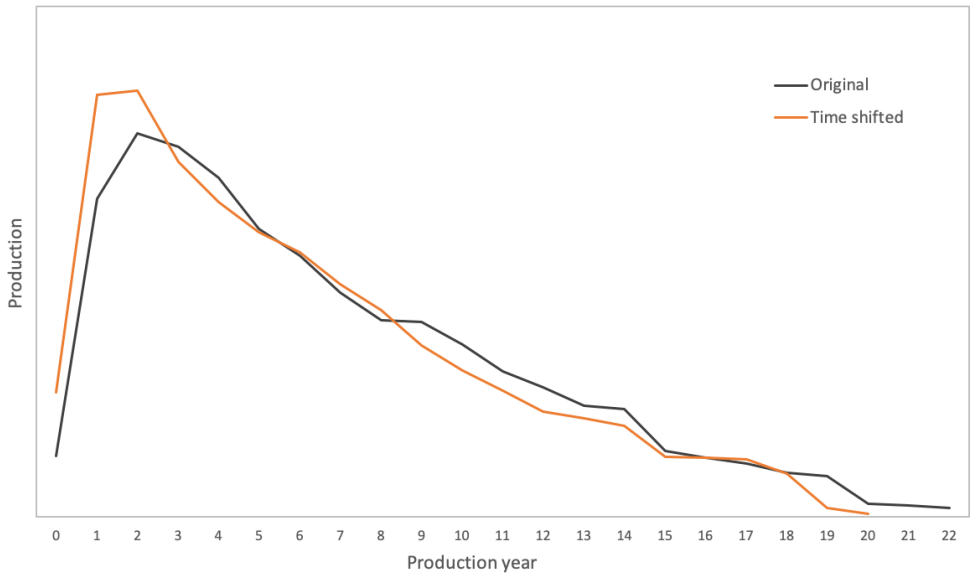


Figure 3.1: *Original versus time shifted actual production data*

Actual production for the 56 fields from original data is represented by the black line in Figure 3.1, while corresponding data for the remaining 54 fields after being time shifted to actual production start is presented by the orange one. As expected, the time shifting procedure yields a smaller tail production and a larger total production for the first 3 to 4 years compared to the original production profile. This can be explained by an earlier encounter of plateau production for fields whose production was time shifted, resulting in a larger portion of total oil production occurring at earlier stages in the production cycle. Figure 3.2 compares the yearly total production for all fields on the NCS to their corresponding estimates made at the time of FID, after being time shifted.

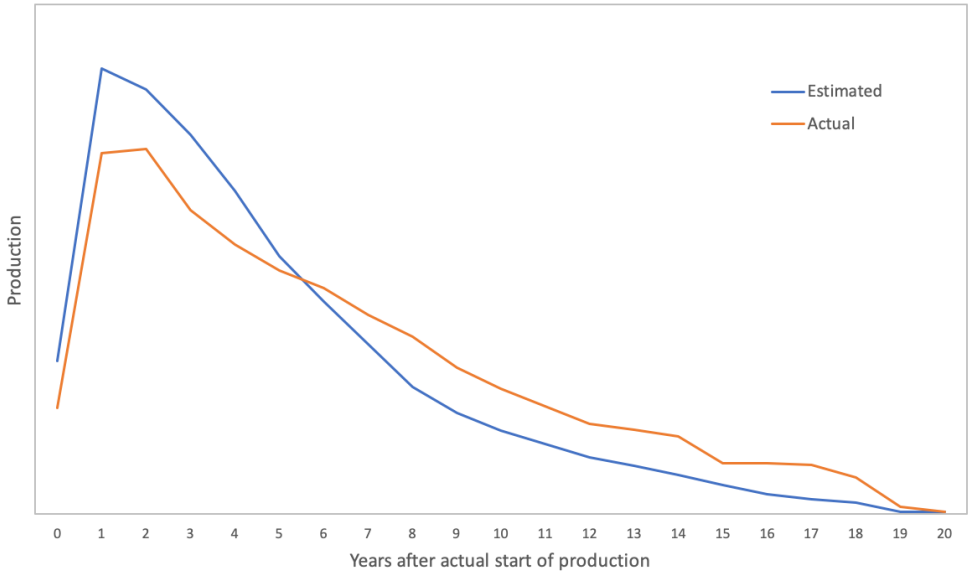


Figure 3.2: *Comparison of estimated and actual production after time shifting data to actual production start*

Figure 3.2 points to a clear trend of actual production falling short of estimated mean production in the initial years, even after eliminating the effect of schedule delays. After about 6 years, however, a shift occurs and actual production surpasses the estimates. From a cumulative perspective, shown in Figure 3.3, the total actual production exceeds the total estimated recovery from year 15 to 20.

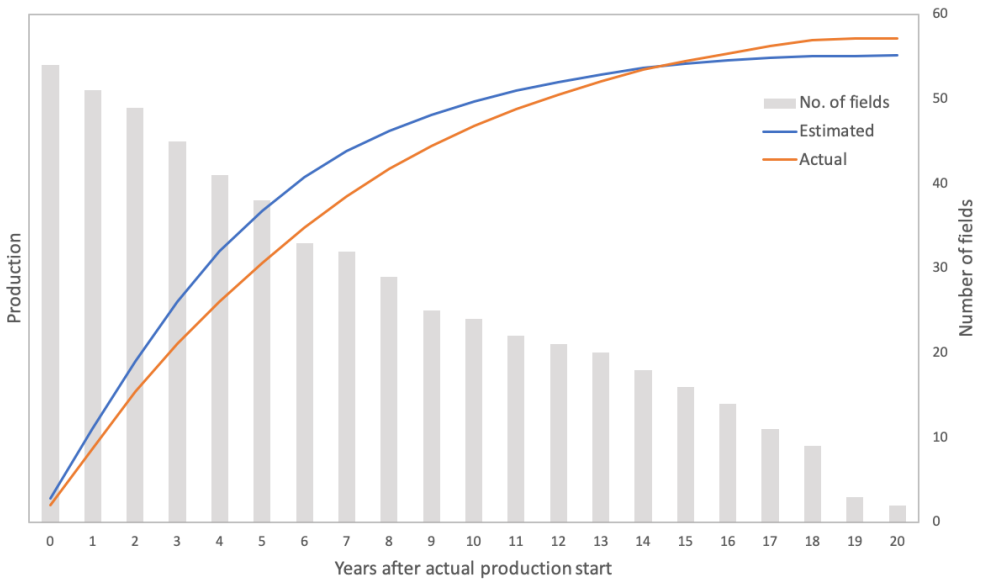


Figure 3.3: *Cumulative estimated and actual production after time shifting data to actual production start*

3.1.2 Data scrubbing

After time shifting production to actual production start, the dataset contains up to 20 years of reported production. However, attention is directed to a limited time period for several reasons. First of all, because of the time value of money discussed in Section 2.2, initial years of production carries the most economic impact on the project NPV. Consequently, from a pure economic perspective, well-informed production forecasts are of more importance for the first years after production start. Secondly, fields are commonly subject to redevelopments where operators initiate reinvestments with intentions of increasing recovery, e.g. through new technology or by implementing methods for enhanced oil recovery. Comparing estimates made at the time of FID with production volumes after additional and often unforeseen capital investments is misleading and gives an unfair edge towards the ultimate recoverable reserves. Thus, when focusing solely on production forecasts reflecting the initial conditions, years with reinvestments are undesirable. As the first instance of redevelopment for fields in the dataset is reported in year 8 (Bratvold et al., 2020), the period constrained by all prior years is a feasible starting point. Moreover, from Figure 3.2, it can be observed that actual production falls short of the estimates from year 0 to year 5, before the annual actual production proceeds to exceed the estimates. Rather than covering all years, this thesis therefore directs its attention to this time period, which will from this point be denoted as the first six years (F6Y) of production. For this time frame, the time shifted dataset comprise 278 production years and up to 54 fields.

Forecast performance for each of the F6Y for fields on the NCS is directly evaluated by comparing P90, mean and P10 production estimates on field-level with the reported actual production. This process is graphically illustrated by the scatter plot in Figure 3.4, where actual production is plotted against the mean estimate for all 54 fields in year 0. The blue dots represent the mean estimate, while the 80% confidence range defined by the P90 and P10 estimates for each field is illustrated by error bars. Further, the orange 45-degree line reflects all points for which actual production exactly equals estimated production and acts as a reference for evaluating forecast performance. To simplify interpretation of the cluster of fields in the lower left corner, Figure 3.4 is supplemented by Figure 3.5, showing a similar representation for fields with estimates below 1 million Sm^3.

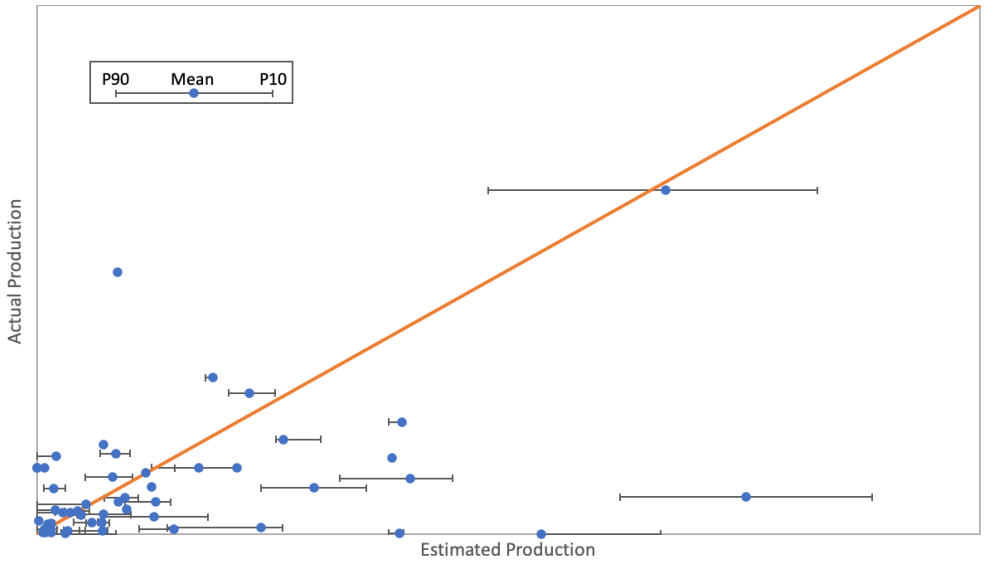


Figure 3.4: *Scatter plot year 0 for all fields*

Comparing Figures 3.4 and 3.5 provides no clear indication of differences between small and large fields in terms of biased estimates. This is further strengthened by a sensitivity analysis on field size with regard to optimism bias for the F4Y performed by Bratvold et al. (2020), for which the results are provided in Figure 3.6. This graph shows the fraction of fields whose actual production is less than or equal to the P50 and P10 production forecasts on the vertical axis and field size on the horizontal axis. Note that Bratvold et al. described the low estimate by a P10 fractile, while this thesis follows the NPD guidelines and therefore denotes the low estimate as a P90 value. As the results clearly show, they found no correlation between bias and field size for these years. It is reasonable to assume that the same holds for the F6Y.

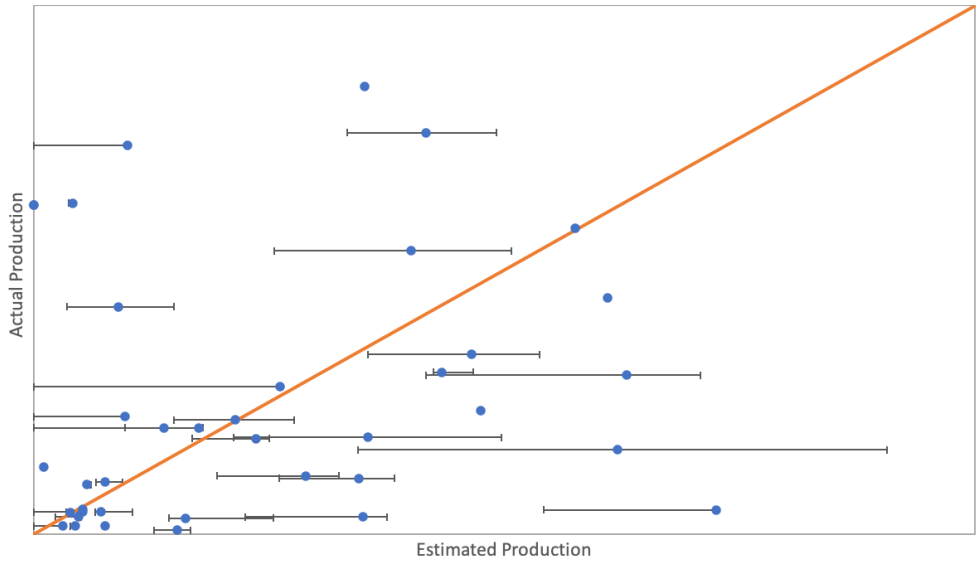


Figure 3.5: *Scatter plot year 0 for fields with estimated production less than 1 million* Sm^3

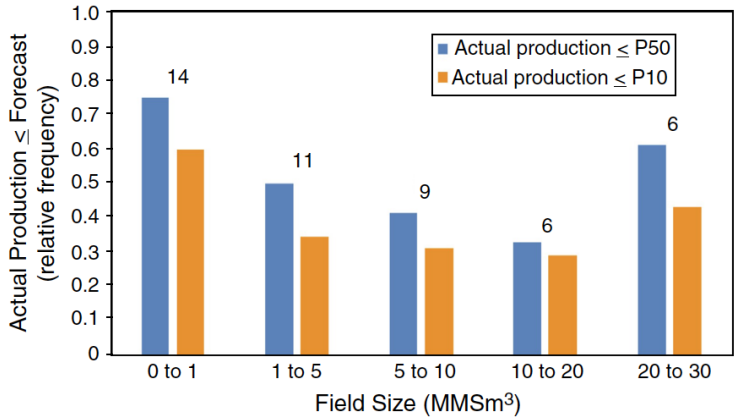


Figure 3.6: *Sensitivity analysis on field size with regard to optimism bias performed by Bratvold et al. (2020)*

For each of the F6Y, forecasts are evaluated in relation to the characteristics of unbiased forecasts presented in Section 3.1. If forecasts are well-calibrated, unbiased and consistent with the knowledge provided by the forecasters, approximately 80% of actual production outcomes should lie within the forecasted P90/P10 range. This means that 80% of the error bars in Figure 3.4 should touch the orange line. If not, the forecasts are overconfident. Moreover, the average reported actual production should be approximately equal to the average mean estimate. Put in other terms, 50% of the blue dots should lie to the left of the orange line and the other 50% should lie to the right. If this is not the case, the forecasts are either optimistic or pessimistic (Bratvold et al., 2020). Forecast calibration is therefore evaluated by determining the fraction of fields whose actual production lie inside the 80%

confidence interval defined by the P90 and P10 estimates and the fraction of fields whose actual production exceed the P90, mean and P10 production estimates.

Table 3.1 summarises the annual calibration statistics after time shifting the entire dataset for the F6Y, and provides the characteristics of unbiased forecast in the rightmost column for comparison. For year 0, only 11% of the actual production data lie within the P90/P10 interval. Summarising the other statistics for year 0, 51% of the actual observations exceed the P90 estimate, 39% exceed the mean estimate and 40% exceed the P10 estimate. Noticeably, no particular year meet the well-calibrated criteria. Moreover, deviations from the well-calibrated characteristics stating that 80% of the observations should lie between the P90 and P10 estimate is most prominent for year 0 and, after that, diminishes with time. The same is true for observations exceeding the P90 estimate. For the two other criteria, covering the number of observations exceeding the mean and P10 estimates, no clear relationship between forecast performance and year is found. Proceeding to study all observations in the period of interest, only 33% of actual observations in the F6Y fall inside the 80% confidence interval defined by the forecasts.

Table 3.1: *Overview of the annual calibration statistics for the time shifted original data, compared to unbiased characteristics provided in the rightmost column*

	Calibration Statistics for the Original Data							
Actual Production	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	F6Y	**Unbiased**
Inside [P90:P10]	11%	25%	39%	44%	44%	45%	33%	**80%**
Over P90	51%	53%	63%	60%	63%	79%	60%	**90%**
Over mean	39%	31%	39%	36%	37%	42%	37%	**50%**
Over P10	40%	27%	24%	16%	20%	34%	27%	**10%**

A thorough study of the provided forecasts for each of the F6Y, the low and high estimates given for some of the fields seem more or less arbitrarily chosen without adherence to a distribution. This can also, to some extent, be seen from Figure 3.4, in which some of the blue dots coincide with either the P90 or P10 estimate, whereas other points totally lack a specification of uncertainty. As these shortcomings represent clear sources of limitations of the production data at hand, data points within the dataset are neglected if:

- the mean estimate is lower than the P90 estimate
- the mean estimate is higher than the P10 estimate
- the P90 and P10 estimates are equal

After time shifting and data scrubbing for the F6Y to exclude missing or inconsistent data, which, in this work, be understood as data that fail to comply with the fundamental principles of probabilistic distributions. The final set of data comprise 237 production years for up to 45 fields and will further be referred to as the "reliable" set of data. Table 3.2 summarises the extent of the dataset after time shifting and data scrubbing for each of the first six years of production.

Table 3.2: *Summary of how the data scrubbing process reduced the extent of the dataset*

	Number of Fields						
Data	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	F6Y
Original	56	54	51	49	45	41	296
Time Shifted	54	51	49	45	41	38	278
Reliable	35	43	45	42	39	33	237

4 Fitting production estimates to a distribution

As part of this work requires the data to be fully described by a distribution, the reliable forecasting data for fields on the NCS are first fitted to metalog distributions. This section aims to provide an introduction to continuous distribution functions, the metalog distribution, and the evolutionary solver utilised for fitting the provided data to distributions. Finally, a detailed description of the metalog fitting procedure is presented.

4.1 Framework of data processing tools

4.1.1 Continuous distribution functions

Continuous distribution functions such as the PDF and the CDF for the estimated field data is necessary to acquire a description of the distributions related to the NCS dataset. Figure 4.1 provides an illustration of typical PDFs and CDFs. The PDF is a function that describes the relative likelihood for a random variable X to take on a given value x (Haslwanter, 2015). The random variable in upcoming operations are estimated field data which are required to exhibit a probabilistic value. Concurrently, there is no likelihood of taking a value less than zero. Thus, the properties of a PDF become:

$$PDF\,(x) \geq 0 \;\forall\; x \in R \quad , \qquad \int_{-\infty}^{\infty} PDF\,(x)\,dx = 1$$

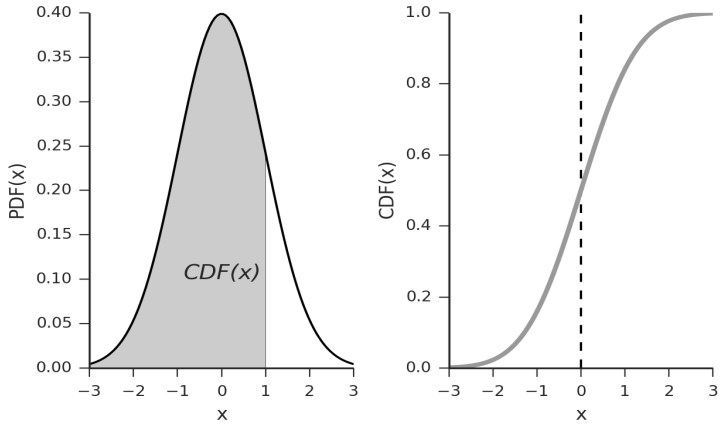


Figure 4.1: *Probability Density Function (left) and Cumulative distribution function (right) of a normal distribution (Haslwanter, 2015)*

The CDF of a random variable X, evaluated at x, is the probability that the random variable will take on a value less or equal to that of x. In scalar continuous distributions, this represents the area under the PDF from $-\infty$ to x (Arora, 2016).

$$P(a \leq x \leq b) = \int_{-\infty}^{\infty} PDF\,(x)\,dx = CDF(b) - CDF(a)$$

4.1.2 The metalog distribution

Data utilised in this work is fitted to a metalog distribution because it, compared to other distributions such as Pearson, Johnson, and other more traditional data display methods, offers almost unlimited shape flexibility through a system consisting of bounded, semi-bounded and unbounded distributions. Further, the metalog quantile functions and PDFs have simple closed-form expressions that are quantile-parameterized linearly by CDF data (Keelin, 2016), making it especially convenient for decision analysis. The theoretical framework from which the CDFs and PDFs for the three sets of bounds are generated is presented in Appendix A. For ease of application, the metalog family is also implemented in two separate pre-programmed Excel sheets – the "SPT metalog" sheet and the "metalog" sheet – that are both downloadable from; *metalogdistributions.com*.

The metalog sheet allows for up to 10 000 input parameters that can either be assigned specific probabilities or defined as equally likely. In this sheet, the user can specify boundedness and the number of terms used to generate the CDF and PDF. The SPT (Symmetric Percentile Triplet) metalog sheet represents a special case of the metalog sheet that is limited to 3-term metalogs, and takes a median as well as a low and high estimate for a specified confidence level as inputs. In both sheets, lower and upper bounds may naturally be specified to reflect the nature of the parameter being analysed. The metalog distributions impose certain requirements on the input parameters to constitute the model. They must:

1. lie within the interval defined by the lower and upper bounds of the distribution (if specified)
2. be strictly increasing
3. be probabilistically defined

4.1.3 Evolutionary Solver

The Evolutionary Solved add-in in Excel aids in the metalog fitting process to be performed in Section 4.2. Evolutionary Solver uses an algorithm based on theory of natural selection and is more likely to find globally optimum solutions for nonlinear equations than its counterpart GRG-nonlinear. The Evolutionary solver algorithm is graphically illustrated in Figure 4.2 and is constructed as follows (Yound, 2020):

1. It starts with a random ”population” of sets of input values that are each plugged into a model, from which a set of output values are retrieved.
2. Next, the selection of values whose output is closest to that of the target value are selected to create a second set of ”offspring” values. These offspring values are essentially ”mutations” of the values retrieved in step 1.
3. The values retrieved in step 2 are then evaluated, and a ”winner” is once again chosen to create a third population.
4. This process is repeated until no better solution for the objective function can be found from one population to the next.

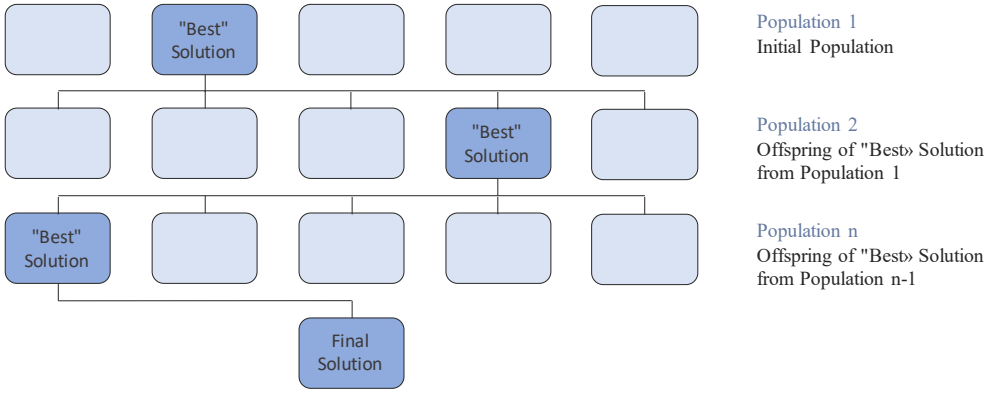


Figure 4.2: *Illustration of the evolutionary solver algorithm (Yound, 2020) (Modified)*

4.2 Metalog distribution fitting

Estimated production data for each of the F6Y is, in its current state, presented by P90, mean and P10 production forecasts at field level, indicating an underlying distribution of outcomes for each field. However, no further information beyond these three values are provided. The first objective post time shifting and data scrubbing, is therefore to mathematically retrieve distributions that describe the provided data. This is addressed by fitting the estimated production data to metalog distributions by utilising the SPT metalog Excel sheet introduced in Section 4.1.2.

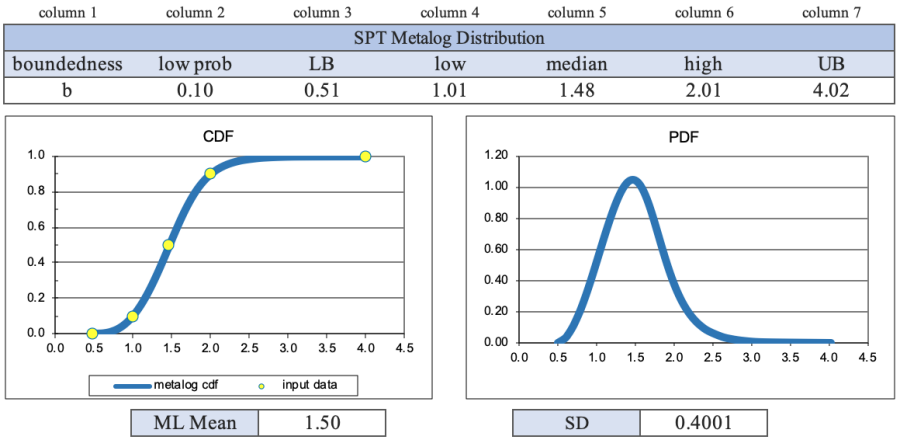
Feasible metalog distributions can only be generated for fields whose estimates comply with the criteria presented in Section 4.1.2. On rare occasions, the P90 estimate is reported as zero. Since it is impossible to produce a negative volume of oil, the lower boundary of the distribution can as a minimum be set to zero. In turn, a positive nonzero P90 estimate is required for adherence to criterion 2. Furthermore, some years include fields where either actual or mean estimated production is reported as zero, imposing problems on the upcoming normalisation process to be performed in Section 5. For these fields, estimated mean production is compared to an actual production of zero, or vice versa. Subsequently, normalised production becomes either zero or undefined. Including these data points would violate criteria 1 and 2. The final selection of fields after excluding fields whose P90, mean estimate or actual production is reported as zero is provided in Table 4.1 and will further be referred to as the "ML consistent" set of data. To further enable comparison with previously performed processes, results from the time shifting and data scrubbing processes performed in Section 3 are also included.

Table 4.1: *Extent of the dataset after data scrubbing and ensuring adherence of the metalog distribution requirements*

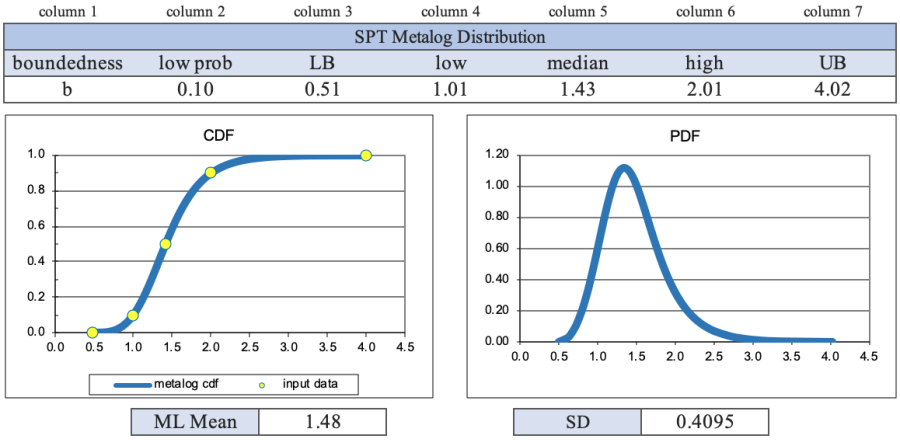
	Number of Fields						
Data	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	F6Y
Original	56	54	51	49	45	41	296
Time Shifted	54	51	49	45	41	38	278
Reliable	35	43	45	42	39	33	237
ML consistent	35	43	45	42	37	31	233

Proceeding to describe estimated production at field level for each of the F6Y, arguments provided in Section 6.1.3 and 6.1.4 favour the use of bounded metalog distributions with 3 terms. Thus, the bounded member of the SPT metalog sheet, built on Equations A.5 and A.6 for $n = 3$ terms, is utilised. Recalling from Section 4.1.2, this model takes the median as well as low and high estimates for a specified confidence level as inputs. Estimated P90 and P10 values for each field can, thus, be used directly as the 10th and 90th percentiles, respectively. However, as the provided dataset contains no information about the P50 percentile, this is determined through use of the Solver add-in in Excel. Solver is configured to let the mean from the metalog distribution (ML mean) converge to the mean estimate given in the dataset through an evolutionary genetic algorithm (see Section 4.1.3) that varies the metalog P50 percentile until a best match is obtained. The rationale of excluding no or infinite production is exercised by setting lower and upper distribution boundaries fixed at $0.5 \cdot$ P90 and $2 \cdot$ P10, respectively. These particular bounds also appear reasonable for capturing the minimum and maximum production capability of each field, considering the associated probabilities for current estimation data.

Figure 4.3 illustrates a metalog fitting operation with synthetic data utilising the SPT metalog sheet and Evolutionary Solver. For this example, reported P90, mean and P10 production estimates are 1.01, 1.48 and 2.01 million Sm^3, respectively. Setup follows by assigning a probability for the low estimate in column 2, and directly inserting the P90 and P10 production estimates into columns 4 and 6 in Figure 4.3a. As the median specified in column 5 is expected to equilibrate at a value close to the distribution mean, this is temporarily set equal to the original mean estimate. Running Solver with the configurations specified above results in the output illustrated in Figure 4.3b, returning a metalog distribution that is consistent with the original P90 and P10 production estimates, and the P50 percentile for which the distribution mean matches the closest adjacent value to the original mean estimate. As shown in this example, the ML mean converges to a value of 1.48, exactly matching the original mean estimate with 2 decimals of accuracy. To capture marginal variations that may occasionally occur when utilising Evolutionary Solver, the process is repeated three times for each field per year through a self-constructed Excel Visual Basic for Application (VBA) program.



(a) *Before Evolutionary Solver is run*



(b) *After Evolutionary Solver is run*

Figure 4.3: *Mean matching operation through Evolutionary Solver in the SPT bounded-metalog sheet with fixed lower and upper boundaries of* $0.5 \cdot low$ *and* $2 \cdot high$*, respectively*

As illustrated in Figure 4.3, the above steps provides a metalog distribution described by a CDF profile containing data points for P90, P50, P10, restricted by the lower bound (LB) and upper bound (UB), and a PDF with associated mean and standard deviation. Ideally, the resulting metalog distribution mean equals the mean estimate provided in the original dataset. However, the degree to which the ML mean converged to the mean estimate varies among the different fields. This may point to an inconsistent relationship between the three different estimates used as input for the fitting process which, in turn, indicate differences in quality of the original distributions from which the P90, mean and P10 production estimates are retrieved. For some fields, the mean estimate is heavily skewed towards either the P90 or P10 value, which may result in difficulties when attempting to generate a suitable metalog distribution. The relative error between the original mean estimate and the metalog distribution mean, expressed by Equation 4.1, acts as an indicator of how well

the generated metalog distribution represents original data. Table 4.2 provides an overview of the number of ML consistent fields satisfying different limits for relative mean error for each of the F6Y.

$$Relative\ mean\ error = \frac{ML\ mean - Mean\ estimate}{Mean\ estimate} \tag{4.1}$$

Table 4.2: *Number of fields for different relative mean errors for the generated metalog distributions with fixed boundaries*

	Number of Fields					
Relative error	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
1%	19	27	29	27	28	25
2%	22	29	31	29	32	25
3%	22	30	31	30	33	26
5%	27	32	35	34	33	27
10%	31	36	41	39	35	30
No limit	35	43	45	42	37	31

Naturally, the number of ML consistent fields increases with acceptable relative mean error. Table 4.2 illustrates that, for all relative errors in the mean, the number of ML consistent fields is largest in year 2 and smallest in year 0 or 5. The year with the lowest number of ML consistent fields is restricting in terms of statistical significance of the reference class. Thus, when evaluating the trade-off between relative mean error and number of fields, the point of initial enquiry falls on the year with the lowest number of included fields. In Figure 4.4, the minimum number of ML consistent fields for all years in the F6Y is plotted against relative mean error. For the restricting year, 30 fields have an ML mean that deviates less than 10% from the original mean estimate. Lowering the acceptable relative error from this point induces a progressive reduction in number of fields until an acceptable relative mean error of 1% is reached, leaving a selection of only 19 fields.

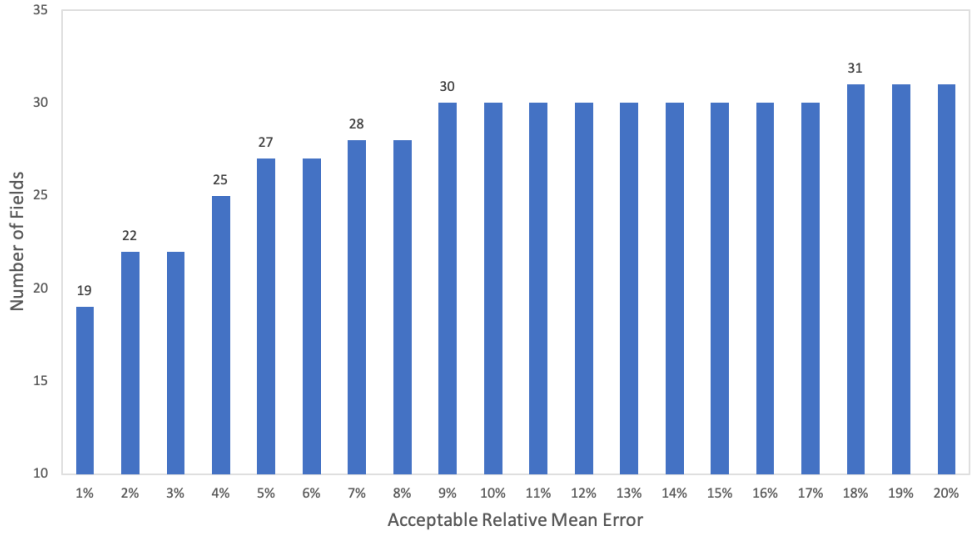


Figure 4.4: *Number of fields plotted against acceptable relative mean error for the generated metalog distributions with fixed boundaries*

Intending to enhance the metalog distribution's ability to match the provided mean estimate and, in turn, reduce the relative error between the metalog and original means, flexible boundaries are introduced to the metalog fitting process. Rather than letting ML mean converge towards the mean estimate by only varying the median, Evolutionary Solver is now additionally allowed to change the lower and upper bounds of the distribution. This is achieved by introducing the following boundary constraints:

$$0 \leq \text{LB} \leq 0.5 \cdot \text{P90}$$
$$\text{P10} \leq \text{UB} \leq 5 \cdot \text{P10}$$

Once again, Evolutionary Solver is run three times for each field. Because solutions found when using fixed boundaries are still valid after introducing more relaxed constraints, a resulting total of 6 distributions are retrieved for each field. The distribution that best reflects the mean estimate, i.e. has the lowest calculated relative error in the mean, is chosen. An updated overview of the effect of acceptable mean error on the number of ML consistent fields is provided in Table 4.3. It can be seen that the lowest number of ML consistent fields for the various levels of relative mean errors is still constrained by year 0 and year 5.

Table 4.3: *Number of fields for different relative mean errors for the generated metalog distributions with flexible boundaries*

	Number of Fields					
Relative error	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
1%	25	33	35	31	30	28
2%	29	33	37	34	33	28
3%	29	33	37	35	33	28
5%	32	36	38	36	33	28
10%	34	38	43	40	36	30
No limit	35	43	45	42	37	31

Figure 4.5 shows the relationship between the minimum number of fields and acceptable relative mean error for the metalog fitting process with flexible boundaries. The dark blue columns represent the number of ML consistent fields for distributions with fixed boundaries and corresponds to that of Figure 4.4, while the light blue columns represent the additional number of ML consistent fields as a result of introducing flexible boundaries. As illustrated, introducing more flexibility to the distribution by relaxing the boundary constraints increases its ability to match the mean estimate. Furthermore, the number of ML consistent fields is less affected by relative error in the mean. Similar to the distributions with fixed boundaries, it is observed that the field count corresponding to a relative mean error of 10% is still 30. However, reducing the acceptable relative error to 1% only reduces the minimum number of fields to 25, compared to 19 when boundaries were held fixed.

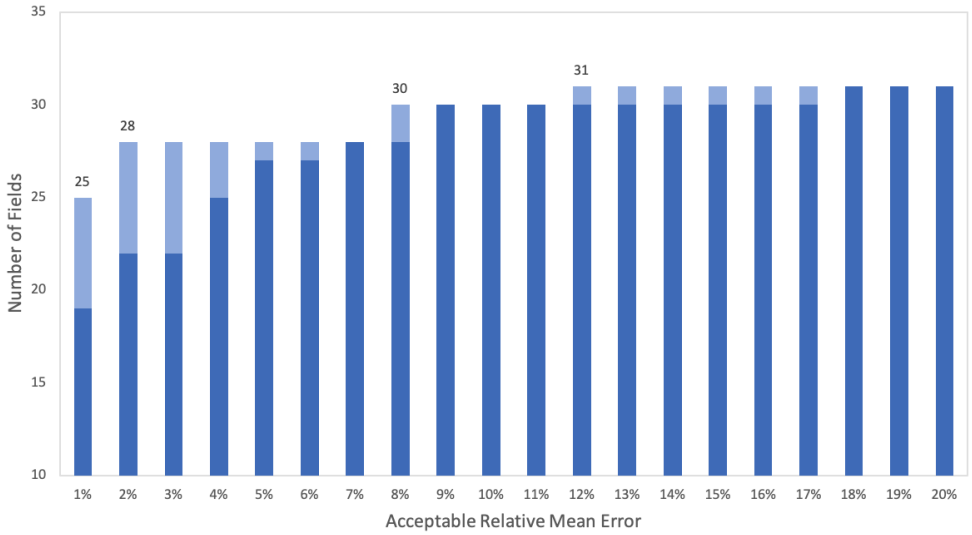


Figure 4.5: *Number of fields plotted against acceptable relative mean error for the generated metalog distributions with flexible boundaries*

Summarised, introducing flexible boundaries for the production distributions at field level generally reduces the relative mean error. This is seen by an increase in the number of fields satisfying a given relative error, which, in turn, yields enhanced statistical significance. Moreover, flexible boundaries enable the resulting distributions to better reflect variations regarding field specific ranges for production capacity. The analyses performed in this thesis therefore utilises the metalog distributions generated with flexible boundaries. As the distributions are used as input for RCF, the choice of an acceptable relative error between the metalog mean and the original mean comes down to an evaluation of the quality of the resulting reference class. From discussion provided on this topic in Section 6.1.5, up to 2% relative mean error is accepted.

5 Debiasing production forecasts through RCF

Once distributions at field level for each of the F6Y are generated for all viable fields, RCF is performed to correct the provided estimates by reducing or removing biases. This method is generally utilised for the purpose of predicting future performance of a project by gathering knowledge about actual performance from a collection of projects with similar characteristics (Leleur et al., 2015). The ruling concept of RCF is that the project at hand is expected to exhibit similarities to the projects in the reference class. For this to be true, it is important that the reference class is well defined in the sense that it is 1) broad enough to be statistically meaningful and, at the same time, 2) narrow enough to be representative for the considered project (Flyvbjerg, 2006). Some of the uncertainty aspects related to the current project performance may, then, be retrieved by studying past performance of the projects that constitute the reference class. This is referred to as "taking an outside view" on the project being forecasted (Flyvbjerg, 2007a). For development projects on the NCS, such a reference class may be constructed using the processed metalog consistent estimation data for each of the F6Y.

This section thoroughly describes the methodology of applying RCF with intentions of correcting the original production forecasts for overconfidence and optimism bias to provide a better-informed decision making basis. The distributions from Section 4.2 are used to generate suitable reference classes and, in turn, retrieve adjustments required to debias original data. The results are evaluated by comparing model calibration of forecast performance before and after correction, and through in-sample and out-of-sample tests performed by applying the results to various test groups. This process is illustrated by the dotted line in Figure 5.1.

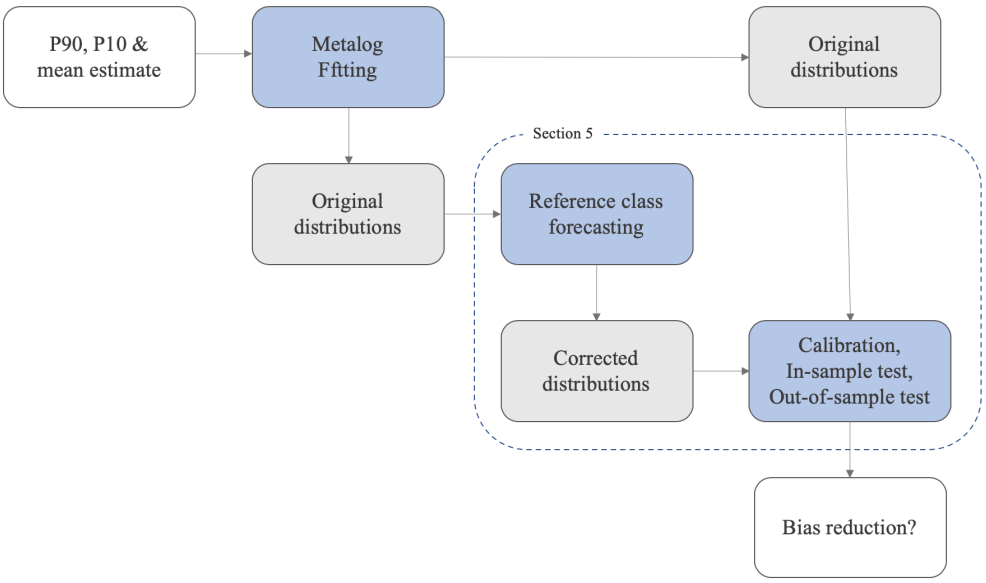


Figure 5.1: *Overview of the RCF process and tests performed to evaluate the resulting bias reduction*

5.1 General methodology description

The RCF methodology developed and implemented in this thesis utilises normalised production data to generate distributions of forecast performance for each of the F6Y, from which the required adjustment to debias the original production forecasts are retrived.

5.1.1 Normalising the production data

Both estimated and actual production is reported in units of million Sm^3. For the purpose of this work, production is normalised following Equation 5.1. Here, *Estimated Production* is the chosen base estimate, i.e. either the P90, mean, or P10 estimate provided in the dataset.

$$Normalised\ Production = \frac{Actual\ Production}{Estimated\ Production} \tag{5.1}$$

This data normalisation process provides not only a common scale for evaluating the relationship between actual and estimated production, which amounts to the forecast performance, but also allows for direct comparison among different fields. A normalised production of 1 translates to an exact match between the actual production and the base estimate, whereas a normalised production greater than or less than 1 implies that actual production is higher or lower than the base estimate, respectively.

5.1.2 Generating normalised annual distributions

Production is normalised for all fields to be included in the reference class. This results in a list of normalised production outputs for each year and essentially constitutes the basis for generating distributions for normalised production on a year-to-year basis for each of the F6Y. However, before annual distributions of outcomes can be obtained, the data is required consistent with the metalog distribution input requirements described in Section 4.1.2. This is achieved by sorting the normalised annual field data in ascending order and defining all input parameters as equally likely. To limit the distribution to positive values only, semi-bounded distributions with a lower bound of zero are utilised. A metalog distribution can then be constructed using the metalog Excel sheet presented in section 4.1.2, yielding CDF and PDF curves representative of the forecast performance for a given year after actual production start. The CDF curve, as originally defined by the metalog distribution, is presented with normalised production on the x-axis and the corresponding cumulative probability, $P(X \leq x)$, on the y-axis. Figure 5.2 shows the CDF for a reference class constructed from an arbitrary selection of ML consistent fields from the dataset.

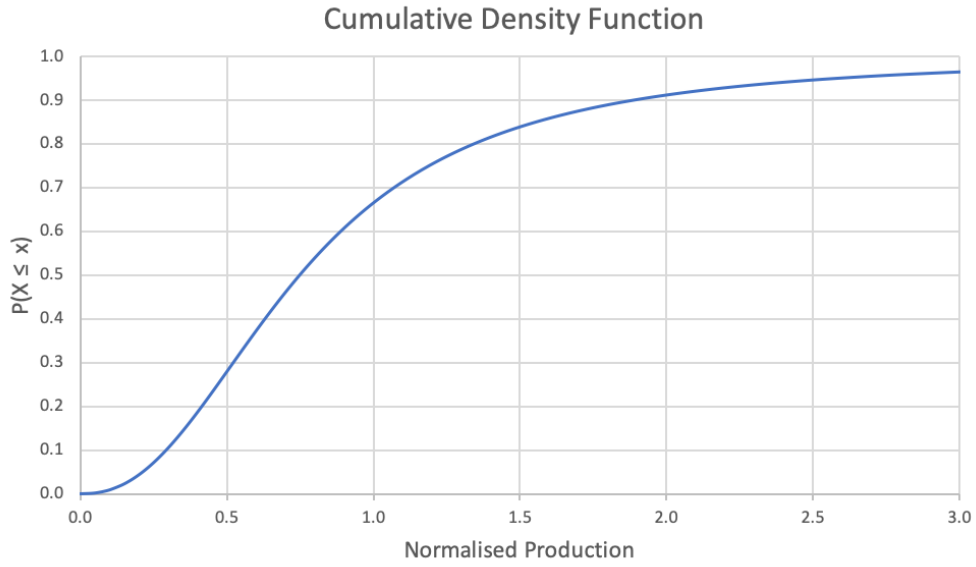


Figure 5.2: *CDF curve for a reference class based on a random selection of ML consistent fields. Normalised production is on the y-axis and the corresponding probabilities is on the x-axis*

To make interpretation in line with the NPD guidelines of the P90, P50, and P10 percentiles, the probabilities are inverted. This results in a survival function (SF), expressing the probability of normalised production being larger than or equal to a given value, i.e. $P(X \geq x)$. Next, the axes are flipped, creating an inverse survival function (ISF) with the survival probability, $P(X \geq x)$, on the x-axis and normalised production on the y-axis (Haslwanter, 2015). The ISF derived from the CDF in Figure 5.2 is shown in Figure 5.3, for which the names of the axes are best explained through an example. Studying Figure 5.3 shows that a survival probability of, say, 10% corresponds to a normalised production of about 1.9, implying that 10% of all normalised productions on the NCS exceed 1.9. Moreover, because normalised production equals actual production divided by the base estimate, 10% of all reported data for actual production is, in this example, 1.9 times greater than their corresponding base estimate. Multiplying all base estimates with 1.9, therefore, gives a normalised production larger than 1.0 for 10% of the fields. In effect, to be 10% confident that the production forecast is met, a correction factor equal to the normalised production corresponding to a probability of 10% from Figure 5.3 had to be applied to the base estimate. Thus, the ISF gives the required adjustment (or correction factor) to be applied to the base estimate to achieve a certain confidence of meeting the forecast. This translates to having the probability of meeting the forecast on the x-axis and the required adjustment on the y-axis.

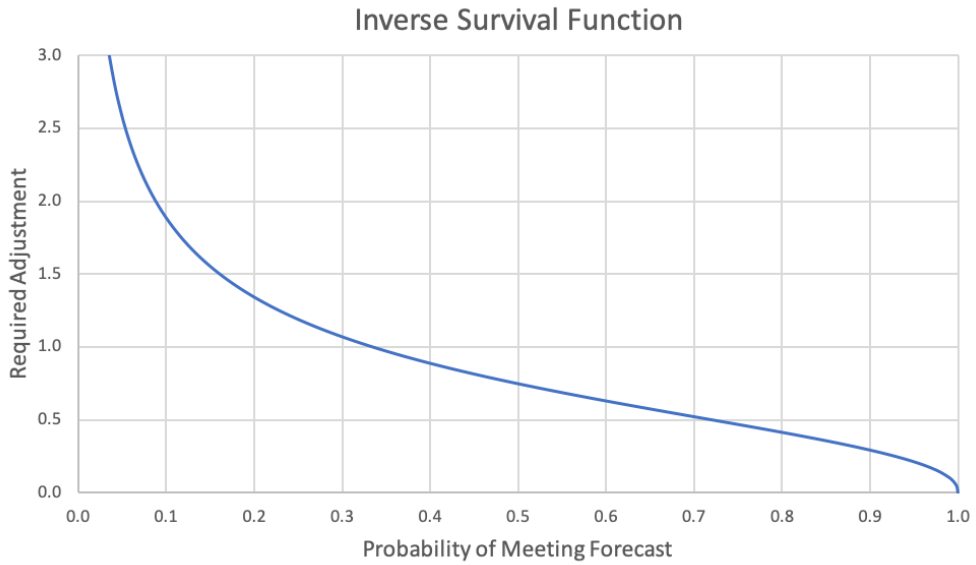


Figure 5.3: *ISF curve for a reference class based on a random selection of ML consistent fields. Normalised production is on the y-axis and the corresponding probabilities is on the x-axis*

5.1.3 Performing correction

The required adjustment, i.e. correction factor to be applied to the base estimate, for obtaining a specific probability of meeting the production forecast is found directly from Figure 5.3. Here, "meeting the production forecast" implies that the observed actual production is equal to or greater than the forecasted production. It can be seen that to be 90% confident that the production forecast will be met, a correction factor of approximately 0.32 needs to be applied to the base estimate. While this returns the P90 correction factor, P50 and P10 correction factors are found in a similar manner. Production estimates for development projects on the NCS can then be corrected using these correction factors. This is achieved by multiplying the correction factors for each year with the base estimates for the corresponding year to find corrected P90, P50 and P10 production forecasts for the project at hand.

5.2 Applying RCF

The above description gives a basic outline of how RCF will be performed with purpose of improving forecast performance for development projects on the NCS. Attention is first directed towards the number represented by the mean estimate. However, to be consistent with the generated metalog distributions, the ML mean is used as base estimate in the normalisation process. With a minimum of 28 ML consistent fields for each year, a variety of reference classes can be chosen. Fields to be included in the reference class can be filtered out based on a specified set of criteria or conditions. When evaluating the forecast performance of fields on the NCS, such criteria may be based on the technology used to generate forecasts, type of depositional environment, how the forecasts are generated or effort put into generating the forecasts. Other possible criteria include year of PDO approval, actual production start, year of actual production, production volumes, or more technical conditions such as reservoir size, depth, pressure or temperature. Due to limited information about the fields in the provided dataset, they can only be distinguished by year of PDO approval, actual production start or based on timing of actual production. As we wish the reference classes to reflect the available information for RCF performed in a given year, this thesis makes distinctions with respect to timing of actual production after production start. Furthermore, because production is normalised, forecast performance is presented by a common scale based on the relationship between actual and estimated production, rather than in terms of production excess or shortfall volumes.

5.2.1 Two different reference classes

For the purpose of studying the degree of consistency in results for reference classes built on different sets of data, two possible methods for filtering out fields to be included in a reference class is presented. These reference classes represent two thought scenarios where a certain amount of historic data is available from previously performed similar projects. Reference class 1 is the equivalent of facing an investment decision in year 2010 and basing the uncertainty analysis on similar projects performed from 1997 to 2009. Reference class 2 is the thought scenario of utilising historic data from projects performed in the time frame between 1997 to 2014 to aid decision making for a development project in 2015.

Reference class 1: Performing RCF in 2010

The first reference class (RC 1) to be considered representative for production forecasts on the NCS is found by including all fields who initiated production in years prior to 2010 and only considering production data for these fields up until 2009. These fields are retrieved from the ML consistent set of data, arrived at in Section 4.2, whose relative error between the metalog mean and the original mean is less than 2%. Figure 5.4 provides an overview of the number of fields included in reference class 1 for each year.

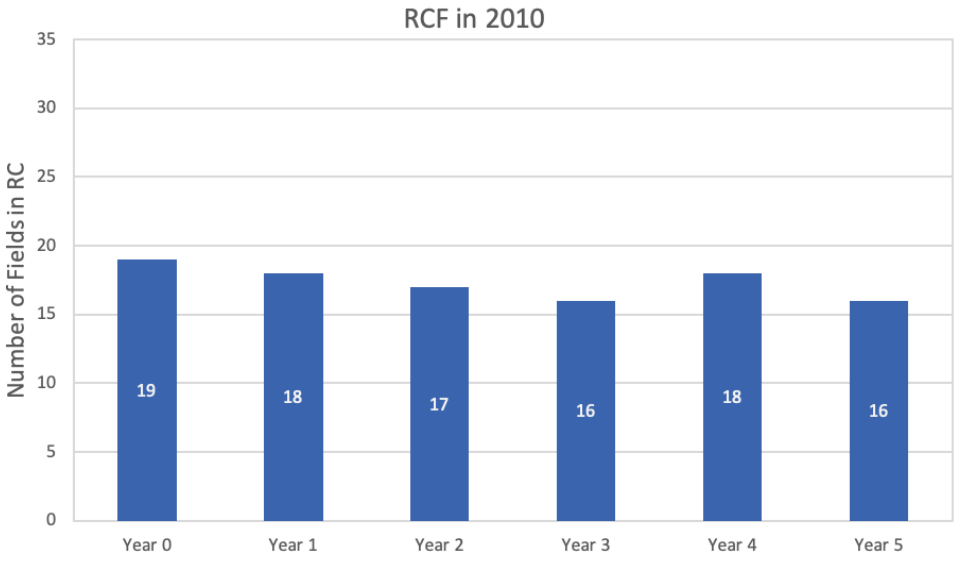


Figure 5.4: *Number of fields in reference class 1*

Proceeding to perform reference class forecasting for each of the first six years of production, actual production is normalised by the ML mean according to the procedure in Section 5.1.1. The sorted values for each year are, then, used as input in the semi-bounded member of the metalog Excel sheet, with 3 terms and a lower bound of zero. From the resulting metalog ISF for each year, the P90, P50 and P10 percentiles are retrieved. As described in Section 5.1.3, these percentiles represent the required adjustment of the ML mean to achieve 90, 50 and 10 percent confidence of meeting the production forecast, respectively. Table 5.1 presents the obtained correction factors for each of the first six years of production for RC 1.

Table 5.1: *Yearly correction factors retrieved from reference class 1*

	Correction Factors for RC 1					
Percentile	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
P90	0.05	0.45	0.54	0.51	0.44	0.43
P50	0.69	0.78	0.81	0.78	0.74	0.94
P10	1.47	1.08	1.17	1.03	1.48	2.13

Reference class 2: RCF in 2015

Another possible reference class that can be extracted from the ML consistent set of data is found by including fields with actual production start from 1997 to 2014 and, for these fields, only including production data reported prior to 2015. Studying the time frame for which this reference class is defined shows a significant overlap with reference class 1 for historic observations from 1997 to 2009. Production data for years 2010 to 2014, however, is only included in reference class 2, providing a larger selection of historical observations for this reference class. An overview of the number of fields for each year for this specific reference class can be found in Figure 5.5. Moreover, P90, P50 and P10 correction factors are listed in Table 5.2, again representing required correction to achieve 90, 50 and 10 percent confident of meeting the forecasted production. These are obtained following the same methodology as described in detail for reference class 1.

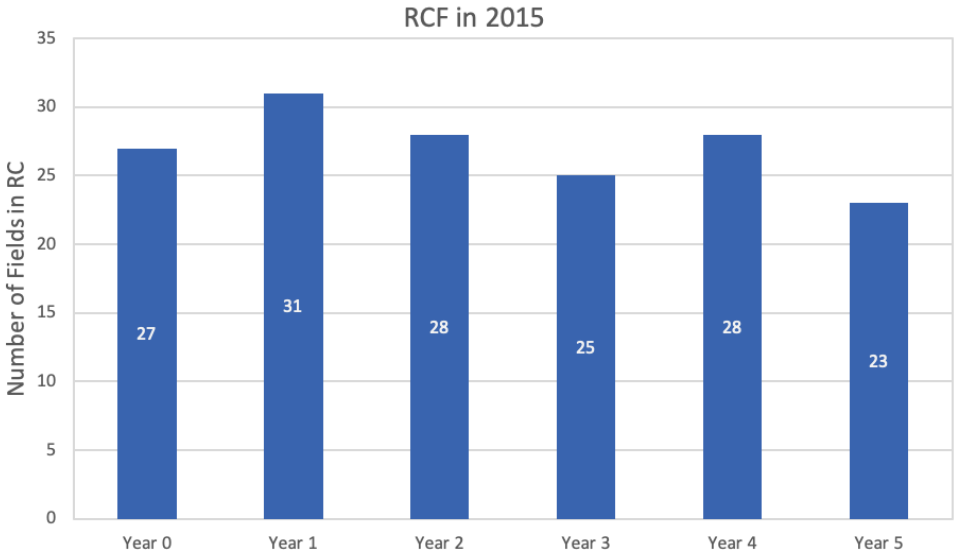


Figure 5.5: *Number of fields in reference class 2*

Table 5.2: *Yearly correction factors retrieved from reference class 2*

	Correction Factors for RC 2					
Percentile	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
P90	0.06	0.38	0.41	0.56	0.47	0.33
P50	0.61	0.81	0.95	0.91	0.86	0.90
P10	1.80	1.40	1.65	1.66	1.79	2.05

Comparing RC 1 and RC 2

Studying Tables 5.1 and 5.2 demonstrates the implications of choosing different reference classes. If the same project was to be evaluated based on both reference class 1 and 2, neither the corrected P90, P50 nor the P10 estimate would coincide. This is graphically illustrated in Figure 5.6, providing a side-by-side representation of the annual correction factors obtained from both reference classes. The blue and orange solid lines represent the P50 correction factor for RC 1 and RC 2, respectively. To investigate variations in the 80% confidence interval, these are further accompanied by their respective P90 and P10 correction factors, represented by the dotted lines. While only small variations can be observed for the P90 and P50 correction factors, the P10 correction differs significantly between the two reference classes. Reference class forecasting performed for a given project would, thus, yield different results depending on whether the project was under development in 2010 or in 2015. RC 2 is seen to generally result in a broader P90/P10 confidence interval. Moreover, the larger P10 would assumably contribute to a larger mean estimate (or expected production) for RCF performed on the basis of RC 2. Further, assuming that the expected present value originating from production volumes are calculated based on the mean estimate, applying

RCF based on RC 2 results in a higher present value compared to RCF based on RC 1. As this present value is key in final investment decisions for development projects in the petroleum industry, using the correction factors from RC 2 may result in a higher probability of project acceptance.

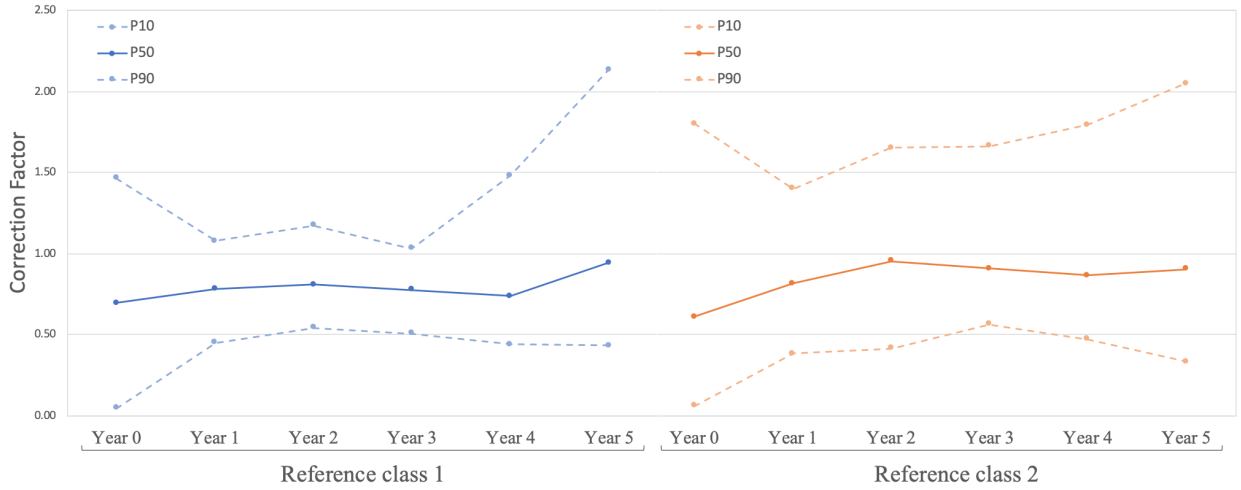


Figure 5.6: *Side-by-side representation of the P90, P50 and P10 correction factors resulting from RC 1 and RC 2*

For both reference classes, the retrieved P90 and P10 correction factors for year 0 are markedly different compared to the other five years. This is assumed to be caused by monthly schedule delays that were not accounted for in the time shifting procedure performed in Section 3.1.1. Further elaboration of this notion is provided in Section 6.1.1.

5.2.2 Progressive RCF

Differences in the correction factors retrieved from the two reference classes defined in the previous section emphasise that the results from RCF are susceptible to variations depending on what projects the analysis is based on. In turn, so is the project's estimated present value of cash inflows. This led to an interest in how the foundation for RCF has developed through time. After the earliest production start is reported in 1997, new fields are continuously put in production up until 2017. The effect of a larger selection of forecast performance observations as progressively more historic data becomes available is investigated by performing progressive annual RCF from 1998 to 2018. Reference class forecasting performed in a given year is, then, restricted to the selection of fields with actual production start before this year. Furthermore, to properly ensure that the reference class only includes data that would actually be available in that specific year, constraints are also put on production year. For example, when performing RCF in 2005, one will naturally only have access to data for fields with actual production start before 2005, i.e. from 1997 to 2004. Moreover, actual production data for these fields are only reported up until 2004. Implementing these constraints, thus, enables RCF to be performed for all years from 1998 to 2018 based on the available information in the year of interest. Note that, for this operation, 2-term metalog distributions are utilised in instances where the 3-term metalog fails to provide a feasible distribution. The results for year 1 are illustrated in Figure 5.7, where the P90, P50 and P10 correction factors are represented by the green, orange and blue lines, and the gray bars report the number of fields included in the reference class. Similar results for the remaining years can be found in Figures B.1 and B.2 in Appendix B.

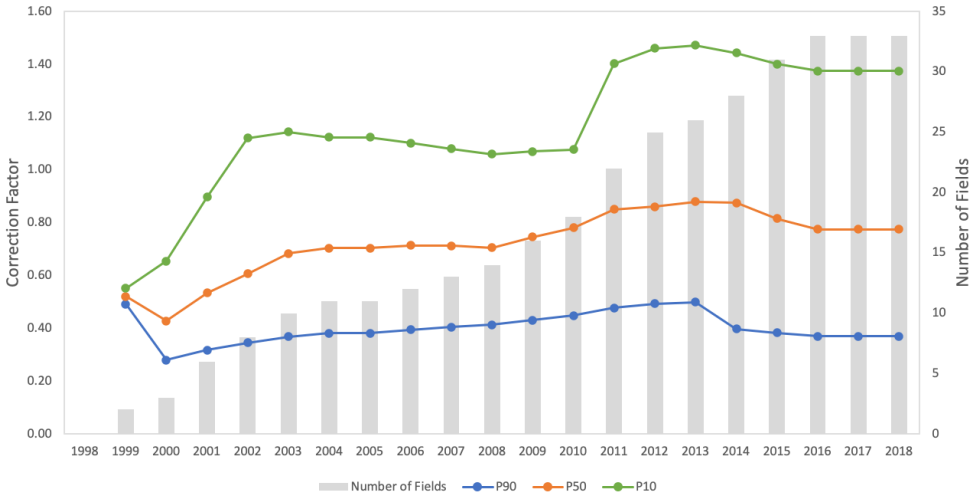


Figure 5.7: *P90, P50 and P10 correction factors resulting from progressive RCF from 1998 to 2018 for year 1*

Progressive RCF shows that all correction factors experience variations through time. The P10 correction factor is proven to exhibit the most prominent variations and, although remaining periodically stable, generally increases as the database of historic observations grows larger. While both the P90 and P50 correction factors are subject to less significant variations, they too fail to remain stable throughout the 20 year time period from 1998 to 2018. In an attempt to capture all possible reference classes and the corresponding variations in the different correction factors for each year, this work is not limited to one single reference class. Instead, iterative random sampling of reference classes is performed.

5.2.3 Random sampling of reference classes

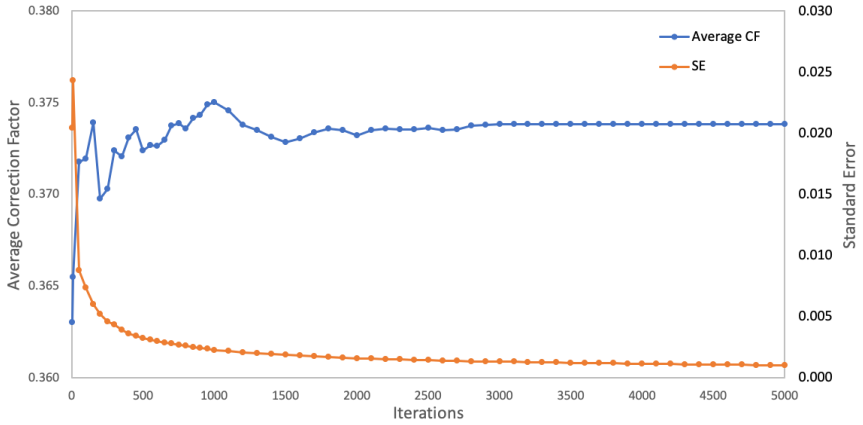
Through programming in Excel VBA, random samples are drawn from the selection of normalised production data for ML consistent fields until a desired reference class size is obtained. For each iteration, the randomly chosen reference class is used as input to generate a metalog distribution following the description given in Section 5.1.2. Next, P90, P50 and P10 correction factors are retrieved from the ISF curve as described in Section 5.1.3. The random sampling process takes reference class size and number of iterations as arguments. A natural initial point of inquiry is determining the number of iterations required for producing robust results that can be consistently reproduced.

Determining the number of iterations

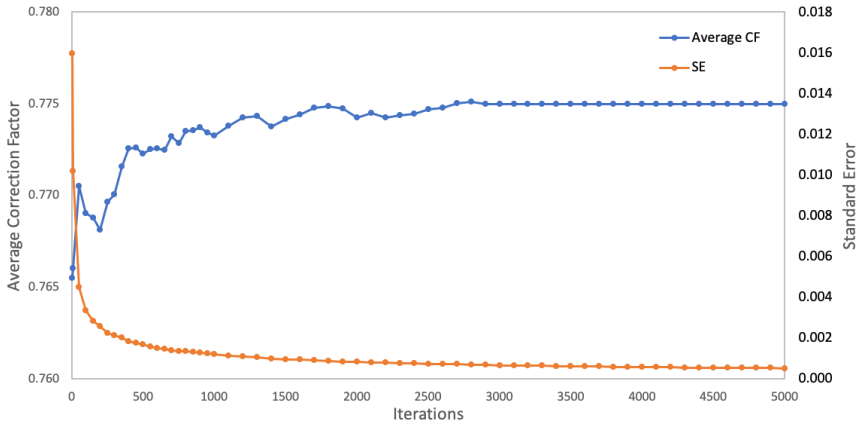
To determine the necessary (and sufficient) number of iterations, the random sampling simulation is initially run for a varying number of different reference classes. For a reference class size of 80% of the total number of ML consistent fields, the simulation is run with 5, 10, 50 and 100 iterations, from which point 100 iterations are added for each run until a maximum of 5000 different reference classes is reached. Average correction factors are determined for each run. Furthermore, to quantify the variations in the distributions represented by the mean of the three different correction factors for a given year, the standard error of the mean is calculated according to Equation 5.2, where σ is the standard deviation and n is the number of iterations. This is a measure of how well the sample mean represents the data, providing a measure of the spread (Kenton, Will, 2020). A smaller standard error signifies a more representative mean. From the inverse nature of this relationship, a low standard error is desired.

$$SE = \frac{\sigma}{\sqrt{n}}\,, \qquad where \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}} \tag{5.2}$$

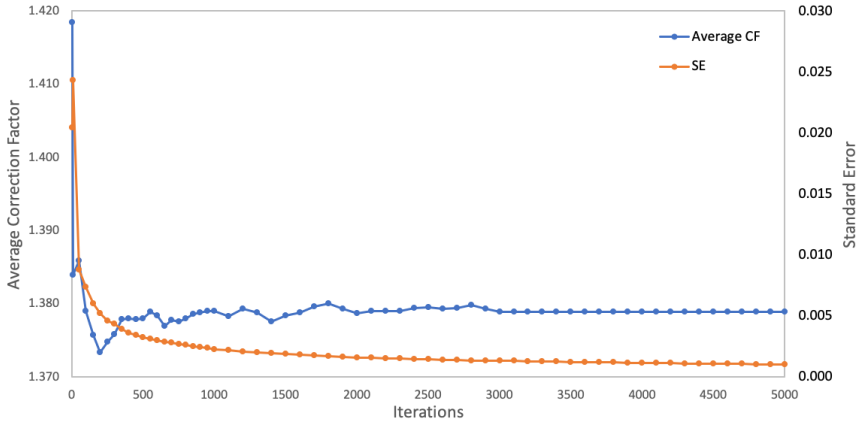
The results are plotted against the number of iterations in Figure 5.8 to find a possible value of convergence for the correction factor means and to study how the related standard error is affected by the number of iterations used in the random sampling of reference classes.



(a) *P90 Correction Factor*



(b) *P50 Correction Factor*



(c) *P10 Correction Factor*

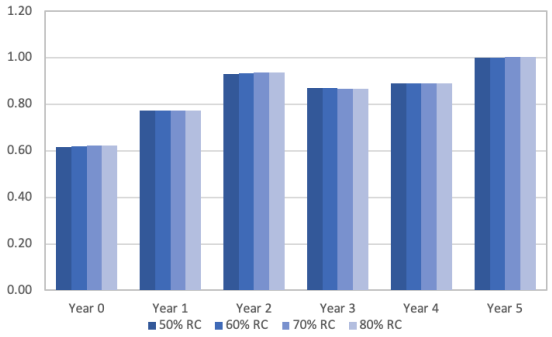
Figure 5.8: *Number of iterations plotted against the mean of each correction factor and their related standard error*

For runs with less than about 2500 different reference classes, significant variations are observed for all three correction factors. Further increasing the number of iterations from this point, however, seems to have a negligible effect on the average, pointing to a clear trend of convergence as the number of iterations increases beyond 2500. As for the standard error, this reduces continuously as the number of iterations increases. Moreover, the reduction is greatest for a smaller number of iterations and flattens out as a more sufficient number of different reference classes is reached. This implies a more robust mean estimate as the number of iterations increases. Based on these findings, 3000 iterations are deemed sufficient for the purpose of this analysis.

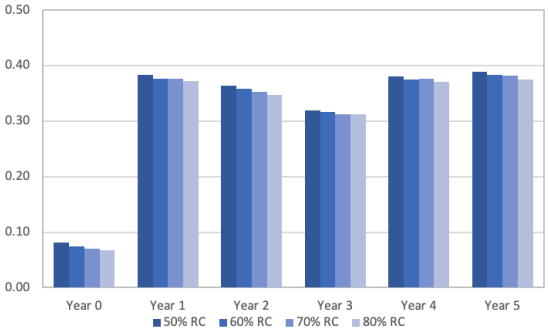
Determining the reference class size

Recalling the definition of RCF provided in Section 5, the reference class should be broad enough to be statistically meaningful but also sufficiently narrow to truly represent the specific project. Including either 50, 60, 70 or 80% of all ML consistent fields in the reference class, a smaller selection of fields in each reference class yields more possible unique and different reference classes. However, this excludes a corresponding amount of relevant historic data for each iteration. Hence, the second step of method development becomes determining the number of fields to be included in each of the randomly sampled reference classes. To achieve this, 3000 new iterations are run for randomly sampled reference classes comprising 50, 60, 70 and 80% of the total number of ML consistent fields. The results are shown in Figure 5.9.

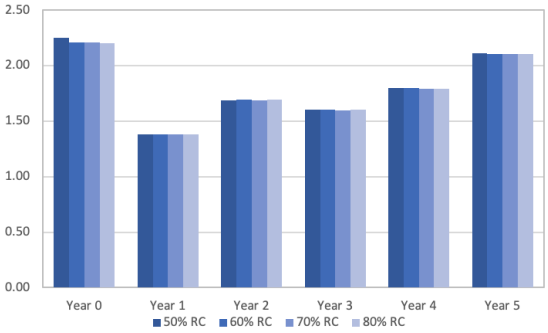
These three figures indicate that the final results, i.e. the average P90, P50 and P10 correction factors, are close to independent of the number of fields included in the reference class. Comparing these results to those obtained through progressive RCF (see Appendix B), random sampling reduces variations related to the size of the reference class. The correction factor with the most prominent variations − the P90 correction factor − only experiences minor differences in the magnitude of 0.01 at most. For this correction factor, an evident trend of reductions in the mean when the amount of fields included in each reference class increases is observed. Because there is no major differences in the results, the choice of RC size is made considering the principles of RCF. Because a broad reference class is desired, including as many fields as possible while still leaving room for random sampling of different sets of reference classes is a natural approach. Moreover, because a lower P90 yields a wider 90% confidence interval which, in turn, increases the probability of covering unobserved actual production, a reference class size of 80% is chosen.



(a) *P50 Correction Factor*



(b) *P90 Correction Factor*



(c) *P10 Correction Factor*

Figure 5.9: *P90, P50 and P10 correction factors as a function of reference class size*

From the above, reference class forecasting is performed by randomly sampling 3000 different reference classes comprising 80% of the total number of ML consistent fields. Following the justification provided in Section 6.2.2, the mean correction factors resulting from these iterations are retrieved. Correction factors for each of the F6Y are given in Table 5.3. Compared to the results obtained through progressive RCF performed in Section 5.2.2, the correction factors are seen to coincide with those found when performing RCF in 2018, in which all available data is used. For year 0, 50% confident of meeting the forecast requires a correction factor of 0.62 to be applied to the mean estimate. Put in other terms, this implies that the observed actual production, on average, falls short of the mean estimate by 48%. Similar to the results obtained from performing RCF with the two reference classes defined in Section 5.2.1, year 0 is observed to be an anomaly also for the random sampling of reference classes. This strengthens the suspicion that monthly schedule delays are present.

Table 5.3: *P90, P50 and P10 correction factors for each of the F6Y for ML mean-based RCF*

	Correction Factors					
Percentile	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
P90	0.07	0.37	0.35	0.31	0.37	0.38
P50	0.62	0.77	0.94	0.87	0.89	1.00
P10	2.20	1.38	1.69	1.60	1.79	2.10

For each field, these correction factors can now be applied to the estimates for the corresponding year to generate corrected distributions of forecasted production. As the actual production was normalised by the ML Mean, the correction also has to be performed on this number. Note, however, that the metalog mean converged towards the original mean estimate when the distributions were generated. This justifies performing correction on the original mean estimate for fields who, for reasons described in Section 4.2, were not included in the metalog fitting process.

5.3 Corrected forecast performance

Validity of the correction factors retrieved through random sampling of reference classes in Section 5.2.3 is evaluated through a series of tests. First, a comparative model calibration for the original and corrected production forecasts is performed, for which the root mean squared error (RMSE) improvement related to a perfectly calibrated judge is retrieved. Next, in-sample tests are performed by applying the obtained yearly correction factors to all ML consistent fields that was included in the iterative random sampling of reference classes. Finally, the corrected model performance is evaluated through testing on independent fields that are not included in the reference class, through an out-of-sample test.

5.3.1 Forecast calibration

Model calibration is performed based on probability-probability plots, which is used to assess a CDF distribution related to a perfectly calibrated reference and removes scaling issues through the use of plotting positions (Gan et al., 1991). A perfectly calibrated estimation model may be defined as the model whose cumulative probabilities of the observed values, when sorted in ascending order, come from a plotting position formula (Cunnane, 1978). For a given year with J number of fields and, thus, J observed values, the plotting position for an observed actual production j is calculated from Equation 5.3. The parameter a can take on different values between 0 and 0.5 and is used to specify the probability distribution. This value is chosen based on the suggestion of Cunnane, who found that $a = 0.4$ is an unbiased

quantile estimator with minimum variance.

$$p_j = \frac{j - a}{1 + j - 2a} \tag{5.3}$$

The plotting position is calculated for ML consistent fields j to J. Next, the actual production percentile of the metalog distribution generated in Section 4.2, described by the P90, P50 and P10 percentiles together with the corresponding lower and upper bounds, is determined for each field. This process is illustrated in Figure 5.10, showing the cumulative density function for an arbitrary field and year within the F6Y. Note that this figure is unique for each field for a given year. For this particular field, an actual production of 4 Sm^3 corresponds to a cumulative probability of 0.2, while a production of 5 Sm^3 corresponds to a cumulative probability of 0.8. This is $P(actual \leq pp)$, and the value that eventually will be plotted against the plotting position as part of the model calibration. For fields whose actual production falls outside of the range defined by the lower and upper bounds of the distributions, no actual production percentile could be extracted. Thus, in scenarios where actual production exceeds the upper bound, the cumulative probability is set to 1. On the contrary, when actual production lies below the lower bound, the cumulative probability is set to 0.

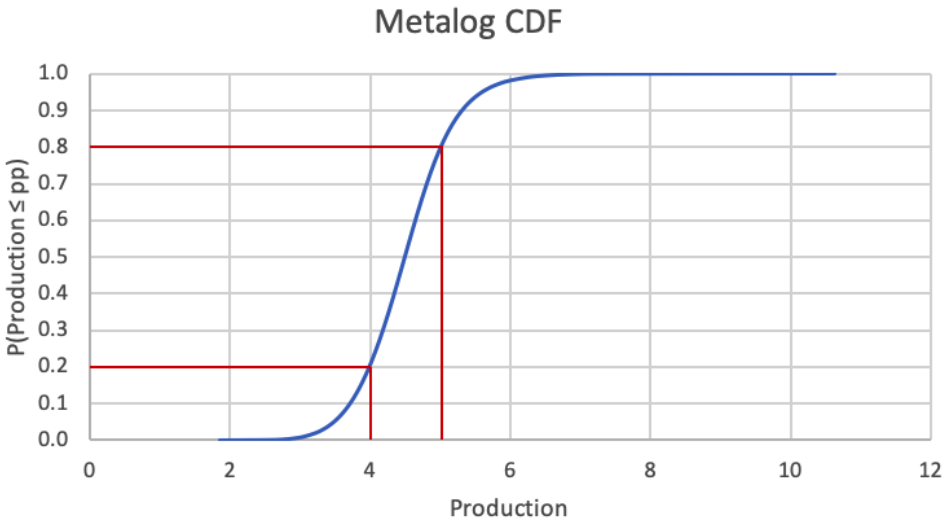


Figure 5.10: *Process of determining the actual production percentile from the metalog CDF*

For the calibration procedure, the lower and upper bounds of the metalog distribution for the original data were set equal to those determined in Section 4.2. For the corrected distributions, however, no information about the boundaries could be retrieved. Although the lower and upper bounds could have been obtained from correction factors corresponding to the P100 and P0 percentiles, respectively, this would fail to take into consideration the variation observed in the original distributions. Some distributions had lower and upper bounds closer to the P90 and P10 estimates than other. It is only natural that the corrected distributions also reflect these characteristics. Thus, the lower and upper bounds for the corrected distributions are determined by matching the ratio between the original P90/P10 estimates and the original lower and upper bounds. This is achieved following Equation 5.4 and 5.5. In these equations, LB and UB denotes the lower and upper bounds, and the subscripts i and $corr$ are used to distinguish between parameters for the original and corrected distributions, respectively.

$$LB_{corr} = \frac{LB_i}{P90_i} \cdot P90_{corr} \tag{5.4}$$

$$UB_{corr} = \frac{UB_i}{P10_i} \cdot P10_{corr} \tag{5.5}$$

After repeating the above process for all fields for a particular year, their actual production percentiles are sorted in ascending order. The sorted values are then plotted against the calculated plotting position, which is ascending by nature. Creating this plot for both the original and the corrected distributions, the two models can be compared to each other and to a perfectly calibrated judge. Figure 5.11 shows the resulting calibration plot for year 1. The blue and orange lines represent the model calibration of the original and corrected forecasts, respectively. Their ability to predict future production is measured by the degree to which they coincide with the perfectly calibrated judge for which the percentiles for actual production equals the plotting position. This judge is represented by the black line.

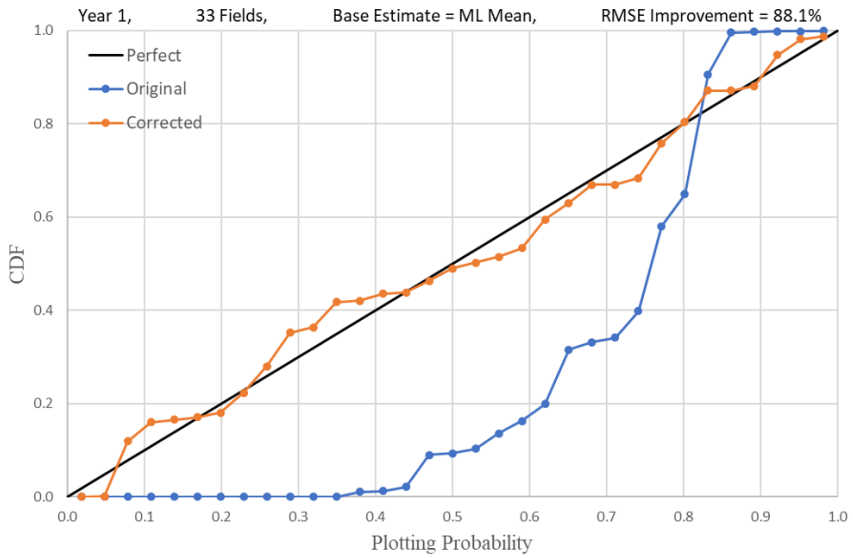


Figure 5.11: *Forecast calibration plot for year 1, showing the original (blue line) and corrected (orange line) forecasts in comparison to a perfectly calibrated judge*

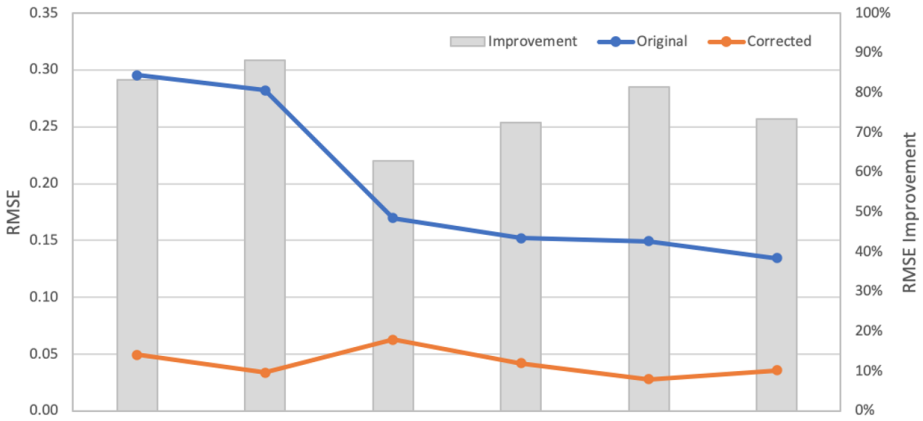
Interpreting Figure 5.11 gives a solid indication of model performance, where good performance is characterized by lines lying close to the perfectly calibrated judge. To mathematically express this model performance, the RMSE between the model and the judge is calculated following Equation 5.6 (obtained through modification of the general equation for RMSE (Kim et al., 2012)). An RMSE of zero indicates perfect calibration.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(pp - P(actual \leq pp))^2}{n}} \tag{5.6}$$

Calculating the RMSE for both models, i.e. before and after applying RCF, provides a quantitative measure of how forecast performance is improved during this process. Figure 5.12 presents the calculated RMSE for each year and for both models. RMSE before and RMSE after are the calculated RMSE before and after correction, respectively. Moreover, the table also includes an overview of the number of fields and RMSE Improvement. As this analysis consistently allows for a maximum of 2% deviation between the actual mean estimates and the mean of the generated metalog distributions, this field count coincides with that of Table 4.3 presented in Section 4.2. The RMSE improvement is calculated using Equation 5.7. It is important to note that the RMSE, because the values are squared, is restricted to positive values only. As a result, an RMSE improvement of 100% is only achievable if the corrected model is perfectly calibrated. Seen from Figure 5.12, the

correction procedure based on annual reference class forecasting yield significant improvements for each of the F6Y, ranging from 63% to 88%.

$$RMSE\ Improvement = \left(1 - \frac{RMSE\ After}{RMSE\ Before}\right) \cdot 100 \tag{5.7}$$



RMSE	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Original	0.295	0.282	0.170	0.152	0.149	0.134
Corrected	0.049	0.034	0.063	0.042	0.028	0.036
Improvement	83.3%	88.1%	62.9%	72.5%	81.4%	73.3%
	n = 29	*n = 33*	*n = 37*	*n = 34*	*n = 33*	*n = 28*

Figure 5.12: *Calibration results from ML Mean-based RCF*

Furthermore, field-by-field RMSE is calculated for the original and corrected models. Since n in Equation 5.6 then becomes 1, leaving only the numerator inside of the square root, this will further be referred to as the Root Squared Error (RSE). Next, RSE improvement is calculated following Equation 5.7, and the results are plotted in Figure 5.14. For fields whose original actual production percentile lies close to its plotting position, which imply an initial RSE close to zero, the corrected actual production percentile is likely to deviate more from the plotting position. This translates to a negative RSE improvement and a corrected production forecast that is worse than the original one. Although this is expected to occur for some fields, the very low initial RSE makes even minor changes largely affect the calculated RSE improvement. The field marked by the red circle in the calibration plot for year 5, presented in Figure 5.13, is a perfect example of this. For this field, the plotting position is 0.447, while the actual production percentile before and after correction is 0.443 and 0.387, respectively. This amounts to an RSE of 1.36 $\cdot 10^{-5}$ before correction and 3.52 $\cdot 10^{-3}$ after correction. Applying Equation 5.7 to this particular field yields a negative RSE improvement of about -25800%.

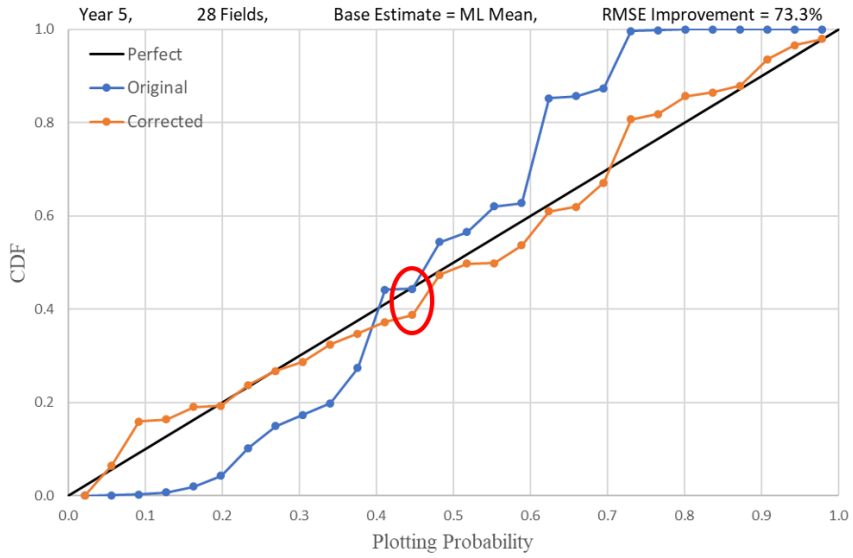


Figure 5.13: *Forecast calibration plot for year 5, showing the original (blue line) and corrected (orange line) forecasts in comparison to a perfectly calibrated judge*

Although this is a rather extreme case, similar behaviour is observed for several fields. As including such values in the graph makes for difficult interpretation, Figure 5.14 is restricted to RSE improvements between -100% and 100%, enhancing illustration of the range of interest constrained by the two orange lines, i.e. how many fields had positive RSE improvement. However, all fields are included when calculating the fraction of fields with a positive RSE improvement. For year 1, 97% of all fields experienced a positive RSE improvement. Figures showing yearly calibration for the other 5 years and year-by-year RSE improvement related to these are presented in Figures B.3 and B.4 in Appendix B. Ralted statistics are presented in Table 5.4.

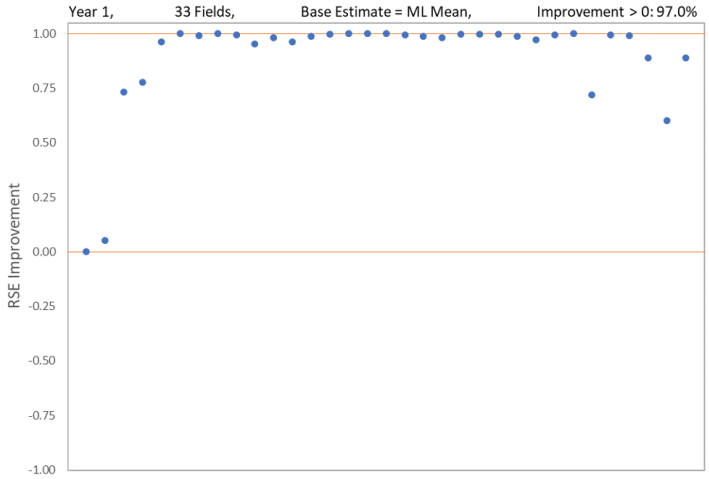


Figure 5.14: *Field-by-field RSE improvement through the process of RCF for year 1*

Table 5.4: *Overview of the total number of fields and the number of fields with a positive and no or negative RSE improvement when applying RCF*

	Field-by-field RSE Improvement					
Fields	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Total	29	33	37	34	33	28
Improvement > 0	27	32	32	30	31	24
Improvement ≤ 0	2	1	5	4	2	4

5.3.2 In-sample testing

Results from the calibration procedure indicate a significant improvement after correcting the original estimates using the correction factors found from random sampling of reference classes. To test the correction factors on the entire set of available data, correction is now performed on all fields in the dataset – including those not found ML consistent. If sufficiently defined, the correction factors derived from the random sampling procedure will be representative of the forecast performance for all fields in the dataset. Applying the determined correction factors to the entire set of data should, then, act to improve the overall forecast performance of this selection of fields. Recalling from the correction procedure provided in Section 5.1.3, correction is performed on the same estimate for which production is normalised. Because the generated ML mean was used in the normalisation process, correction is therefore performed by applying the correction factors to the mean estimate provided for each field. Next, the forecast performance resulting from this correction can be expressed through the calculation of certain calibration statistics. For each field, actual production is compared to the original and corrected production forecast

distributions by determining the calibration statistics presented in Section 3.1.2. Instead of comparing actual production to the mean, like for the calibration statistics of the original distributions, actual production is compared to the P50 value of the corrected distributions. This is due to the lack of information about the corrected mean for fields that are not ML consistent.

Calibration statistics for the corrected distribution are presented in Table 5.5. Studying the results in relation to those obtained for the original distributions, provided in Table 3.1, it is evident that the corrected production estimates exhibit characteristics that are more closely aligned with the definitions of a well-calibrated (unbiased) forecast. This implies enhanced forecast performance for the entire dataset after applying the correction factors to the mean estimates.

Table 5.5: *Overview of the annual calibration statistics for the corrected data, compared to unbiased forecast characteristics provided in the rightmost column*

	Calibration Statistics for the Corrected Data							
Actual Production	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	F6Y	**Unbiased**
Inside [P90:P10]	76%	76%	80%	78%	78%	74%	77%	**80%**
Over P90	93%	90%	92%	84%	88%	87%	89%	**90%**
Over P50	57%	53%	39%	44%	46%	39%	47%	**50%**
Over P10	17%	14%	12%	7%	10%	13%	12%	**10%**

5.3.3 Out-of-sample testing

Above is considered how forecast performance can be improved for the whole selection of fields by applying reference class forecasting. However, a substantial number of these fields were also utilised in the process of generating the correction factors. To truly investigate the validity of the determined correction factors, tests should be performed on independent fields. Such tests are performed by implementing out-of-sample tests in the random sampling simulation. For each iteration in the random sampling of reference classes, fields that are not included in the reference class forms a test group. For this test group, the calibration statistics presented in Section 3.1.2 are calculated based on their original distributions. Next, correction factors found representative for the reference class are then applied to the mean estimate of each field in the test group to generate corrected distributions. Finally, calibration statistics are calculated for the corrected distributions. Determining the average for 3000 iterations enables comparison of forecast performance before and after correction. The results from the out-of-sample test are provided in Table 5.6.

Table 5.6: *Average calibration statistics for test groups before and after ML mean-based RCF, compared to unbiased forecast characteristics presented in the rightmost column*

	Average Calibration Statistics for TG Before ML mean-based RCF						
Actual Production	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	**Unbiased**
Inside [P90:P10]	14%	31%	44%	50%	42%	50%	**80%**
Over P90	41%	51%	65%	67%	63%	79%	**90%**
Over P50	34%	27%	35%	46%	45%	53%	**50%**
Over P10	27%	21%	22%	17%	21%	28%	**10%**
	Average Calibration Statistics for TG After ML mean-based RCF						
Actual Production	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	**Unbiased**
Inside [P90:P10]	81%	82%	77%	77%	80%	79%	**80%**
Over P90	89%	92%	92%	85%	92%	91%	**90%**
Over P50	51%	47%	41%	51%	48%	47%	**50%**
Over P10	8%	11%	15%	8%	12%	12%	**10%**

Recalling the characteristics of well-calibrated estimation models, 80% of actual observations should lie within the 80% confidence interval confined by the P90 and P10 estimates, and 90, 50 and 10% of actual observations should exceed the P90, P50 and P10 estimates, respectively. The out-of-sample test shows that, before correction, production forecasts for the test group are not satisfying these characteristics. Moreover, the probabilities experience significant variations among the different years. For example, it can be seen that in year 5, 50% of all fields in the test group lie inside the 80% interval. In year 0, this number is as low as 14%. Regardless of these variations, applying RCF to the test groups yields consistent improvements for all years. After correction, the calibration statistics is seen to be closely aligned with those of well-calibrated forecasts. In effect, this translates to an overall reduced overconfidence and optimism bias.

This can be even further illustrated by determining, for a particular year, the normalised calibration statistics for the original and corrected distributions for each field. Normalisation is performed by dividing each calibration statistic with its well-calibrated percentage. Subsequently, each calibration statistic is equal to 1 if the production forecast is well-calibrated. When evaluating production forecasts, values lower than 1 indicate over estimation and, consequently, room for improvement. Values larger than 1 imply that operators on the NCS, on average, produce more than what is estimated. Figure 5.15 graphically illustrates the normalised calibration statistics for year 1. The blue and orange lines represent original and corrected production forecasts, respectively, while the black line is the characteristics of well-calibrated forecasts. Evident from this graph, calibration statistics for the corrected production forecasts are more closely aligned to the

well-calibrated characteristics than the original production forecasts. Similar interpretations can be made for the other years, for which identical plots are presented in Figures B.1 and B.2 in Appendix B.

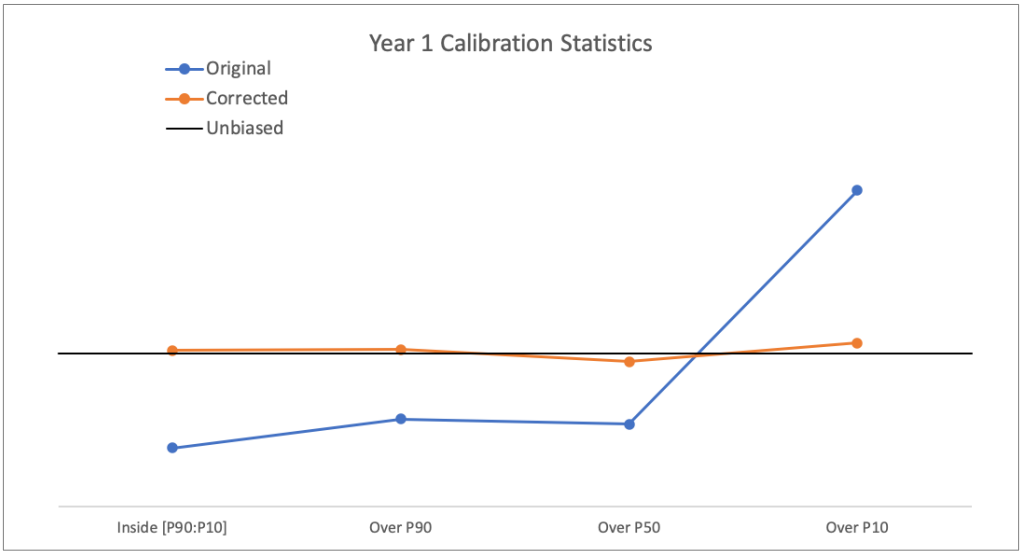


Figure 5.15: *Normalised calibration statistics for year 1*

5.4 Evaluating the low and high estimates

The above sections focus solely on the mean estimate in the process of correcting the production forecasts. To discover possible inconsistencies in the relationships between the low, medium and high estimates, the low and high estimates are now evaluated separately. This means that actual production is normalised based on the P90 and P10 estimates rather than on the ML Mean. New annual distributions that now express the forecast performance with respect to the low and high estimates can, then, be generated. As before, reference classes are constructed from random sampling of 80% of all fields, and 3000 different reference classes are selected for both sets of distributions. This results in new histograms like those presented in Figure 6.4 for the three correction factors, from which the average correction factors are extracted. The correction factors retrieved from P90- and P10-based random sampling of reference classes are summarized in Table 5.7 and 5.8. As before, the P90, P50 and P10 correction factors are the multipliers needed to be applied to the estimated value to ensure 90, 50 or 10 percent confidence of meeting the production forecast. Instead of applying these correction factors to the mean estimate to obtain the corrected distributions described by P90, P50, and P10 percentiles, correction is now performed on the original P90 and P10 estimates.

Table 5.7: *P90, P50 and P10 correction factors for each of the F6Y for P90-based RCF*

	P90-Based Correction Factors					
Percentile	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
P90	0.09	0.48	0.48	0.51	0.61	0.66
P50	0.81	1.11	1.35	1.30	1.32	1.51
P10	3.08	2.08	3.02	2.82	4.50	5.22

Table 5.8: *P90, P50 and P10 correction factors for each of the F6Y for P10-based RCF*

	P10-Based Correction Factors					
Percentile	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
P90	0.06	0.30	0.24	0.20	0.25	0.25
P50	0.52	0.65	0.76	0.70	0.65	0.70
P10	1.84	1.14	1.26	1.15	1.24	1.48

Next, the same calibration test from Section 5.3.1 is performed on the distributions generated by both the P90- and P10-based correction factors. These calibration plots and corresponding plots for field-by-field RSE improvement are presented in Figures B.6 to B.9 in Appendix B. The RMSE before correction is based on the original distributions and is thus independent of the method used to normalise actual production. However, correcting the original estimates using the three different sets of correction factors naturally results in three unique distributions. For a particular field, differences among these distributions will affect the actual production percentile retrieved from its respective CDF. Thus, comparison to the field's plotting position differs between the three sets of distributions which, ultimately, also affects the calculated RMSE. Figure 5.16 illustrates the above by comparing the RMSE improvement for the ML Mean-, P90-, and P10-based corrections. Despite the observable differences in RMSE improvement after performing correction with the three different base estimates, these are minor and no clear trend is present.

Figure 5.16: *RMSE improvement for the F6Y related to perfect calibration, resulting from ML Mean-, P90-, and P10-based RCF*

Out-of-sample tests performed for P90- and P10-based RCF are provided in Tables B.2 and B.4 in appendix B. These results show only small variations compared to those related to ML mean-based RCF. Thus, neither base estimate can be argued preferable over the other two, based on the results from this work. However, a significant number of production years were excluded from the analysis due to inconsistency. This indicates unreliable information in the P90 and P10 estimates, which will be further discussed in Section 6.4.

6 Discussion

6.1 Data processing and distribution fitting

6.1.1 Elimination of schedule delays

In its original form, many of the fields in the NCS dataset was subject to schedule delays. As a result, estimated production was compared to an actual production of zero until production of first oil. Aiming to eliminate the effects of schedule delays, the time shifting procedure described in section 3.1.1 was performed. Although this improved the basis for evaluating current forecast performance, completely removing the effect of schedule delays is impossible without more information. RCF performed in Section 5.2 show year 0 to be an anomaly, indicating that monthly (or daily) schedule delays are still present. This can be observed by studying the correction factors in Table 5.3, showing that year 0 has the lowest P50 correction factor. Moreover, the P90 and P10 correction factors for year 0 results in a considerably larger confidence interval for the corrected distribution compared to other years. While P90 correction factors for year 1 to year 5 all lie within the range of 0.3 to 0.4, the P90 correction factor for year 0 is as low as 0.07. As uncertainty of production forecasts delivered at the time of FID, in general, is expected to be larger for years later in the production cycle, this behaviour is more likely to be explained by schedule delays. A monthly schedule delay of, say, 6 months for a field with estimated production start in January, will not initiate production before in July. As this is not detected by the yearly time shifting performed in this work, actual production from July to December is compared to estimated production from January to December. Potential production shortfalls are then contributed to poor forecast performance, rather that the delayed production start. This is a perfect example of how monthly schedule delays may result in a poor indication of forecast performance for year 0. Consequently, development projects that deliver on time may be subject to excessive correction if applying the correction factors determined through RCF performed in this work. Although the effect of schedule delays is naturally also transferred to all following production years, it is only year 0 that is prone to the initial months of zero production. Thus, the effect of monthly schedule delays is significantly smaller for the other 5 years.

6.1.2 Choice of FnY

Correction factors can, if desired, be found for each year for which historic data exists. Although the dataset contains up to 20 years of reported production, this work restricts its attention to the F6Y following the argumentation provided in Section 3.1.2. One natural question that arises is whether the number of aggregation years (FnY) affects the results related to total performance over the first n years. Figure 6.1 shows the results from a sensitivity analysis performed by Bratvold et al. (2020) with intentions of answering this. Because their work is based on the same set of data that is utilised in this thesis, their results can be directly used to assist in this discussion. The sensitivity analysis shows the percentage of fields whose cumulative actual production over the FnY does not exceed their respective P90 and P50 estimates, denoted as $Actual\ q \leq P90$ and $Actual\ q \leq P50$. Evident from from Figure 6.1, these percentages are not very sensitive to the number of aggregation years. This argumentation is not important for year-on-year results presented for the F6Y but is highly relevant for results related to the cumulative forecast performance.

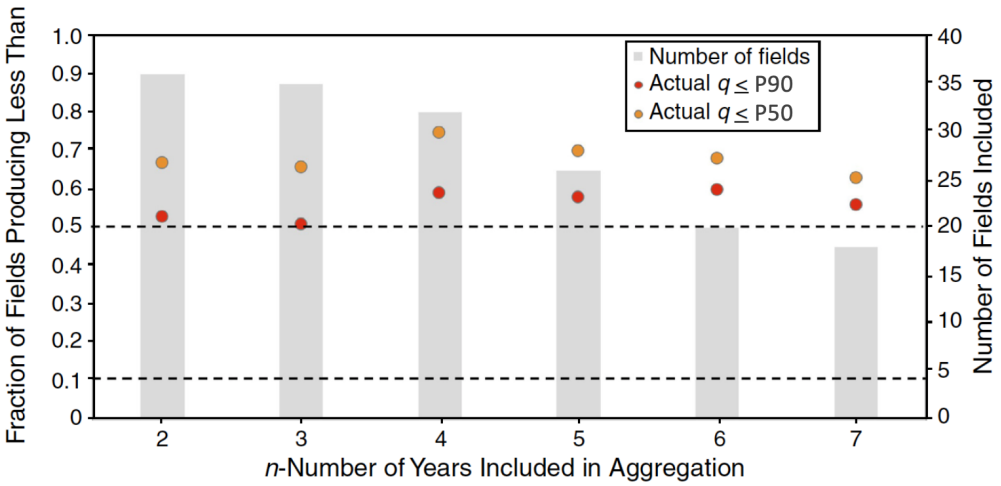


Figure 6.1: *Sensitivity analysis of how the number of aggregation years affects percentage of fields whose cumulative actual production does not exceed their cumulative forecasted P90 or P50 (Bratvold et al., 2020) (modified)*

6.1.3 Choice of n-term metalog

The number of metalog terms is chosen so that the distribution inherently imposes flexibility beyond the traditional distributions, but at the same time does not include too many terms, as this may lead to overfitting. In general, the use of 2 terms limits the metalog distribution to a logistic distribution, which gives it enough flexibility to match the mean and standard deviation, but not skewness or kurtosis. With 3 terms, the distribution can be additionally described in terms of skewness, whereas 4 terms are required for kurtosis to be included (Keelin, 2016).

The forecasts are probabilistic and, for each year, estimates production using three terms; forecasted P90, mean and P10. The metalog distribution is related to the data through a set of linear equations, which are solved using the input parameters. With m input parameters the resulting metalog distribution can be described by $n \leq m$ terms (Keelin, 2016). Moreover, the input parameters must be assigned a probability. Although the mean estimate lacks a corresponding probability, this is used to find a feasible P50 percentile for the metalog distribution by utilising the evolutionary algorithm in Excel. Thus, utilising the three terms provided in the dataset, this work is restricted to the use of either 2- or 3-term metalog distributions. Moreover, Keelin (2016) suggests that, for application in decision analysis with three assessed data points, a 3-term metalog should be chosen. The resulting CDF then passes through all three data points exactly, providing a complete representation of the dataset at hand. Thus, a 3-term metalog distribution was chosen for the process of obtaining field-by-field distributions representative of production estimates provided by the NPD. For consistency, the 3-term metalog distribution was also chosen for generating annual distributions as part of the random sampling of reference classes.

6.1.4 Choice of metalog boundedness

The metalog family consists of distributions with three different specifications for boundedness; unbounded, semi-bounded, and bounded. Because oil production cannot be less than zero, the unbounded and semi-bounded (upper) distributions are not feasible for this application. With intentions of avoiding subjective definitions of lower and upper bounds for each distribution, preliminary testing was performed with semi-bounded (lower) distributions with a lower bound of zero. This restricts production to positive values only. However, the shape of the resulting CDF and PDF, then, essentially depends on the magnitude of the production estimates relative to 0. This is illustrated in Figure 6.2. As this behaviour is not desired, the bounded metalog distribution is chosen to enable the definition of field-specific constraints for the lower and upper bounds. Implementation of lower and upper bounds can also be justified in the sense that they, if properly defined, better reflect the range of possible outcomes indicated by the P90 and P10 production estimates. For the random sampling procedure, where actual production is normalised by a base estimate, the semi-bounded metalog distribution with a lower bound of zero is utilised.

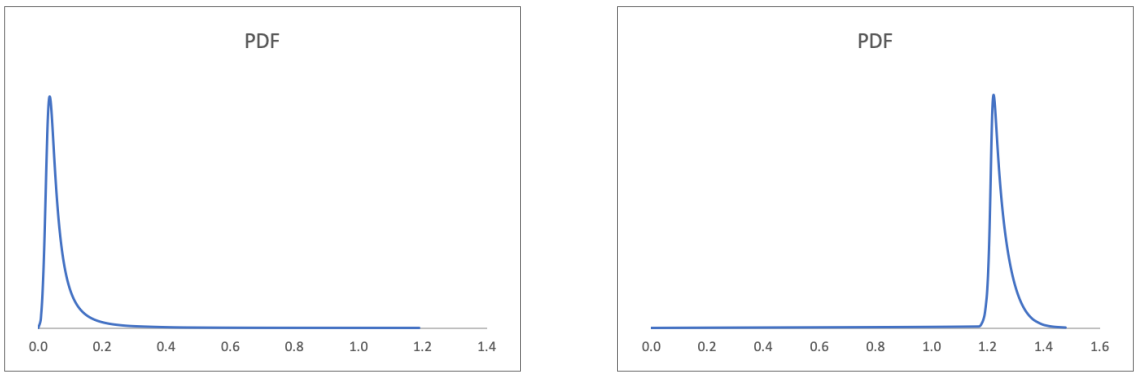


Figure 6.2: *PDFs for two arbitrary sets of data with a lower bound of zero*

6.1.5 Choice of acceptable relative mean error

Proceeding to study the choice of an acceptable relative error between the metalog mean and the original mean estimate, reference is made to the theory of RCF described in Section 5. Higher relative mean errors allow for more freedom in terms of the metalog distribution's ability to describe the mean estimate provided in the dataset. Thus, the selection of fields satisfying a low relative mean error is more representative of the dataset than selections constrained by a higher limit. This provides an argument for choosing a low limit for acceptable relative error in the mean. However, seen from Figure 4.5, a lower limit is accompanied by a smaller number of fields in the resulting reference class. Hence, the choice reduces to an evident trade-off between the statistical significance of the resulting

selection of fields and the degree to which they are representative of the original data. The relative mean error deemed acceptable may therefore affect the quality of the reference class. Studying Figure 4.5, the metalog mean for 25 fields deviates less than 1% from the original mean for the year with the lowest number of fields. For 2 to 7%, the minimum number of fields included in the reference classes for each year increases to 28. The value of increased statistical significance obtained from 3 additional fields for the year with lowest field count is perceived greater than the loss of comparability resulting from a 1% higher relative error in the mean for these fields. Thus, the analyses performed in this work allows for up to 2% error in the mean.

6.2 Reference class forecasting

6.2.1 The resulting reference class size distribution

The program used for random sampling of reference classes does not yield reference classes of a fixed size. Instead, the desired size is specified as a percentage of the total number of ML consistent fields. Next, for each iteration, all ML consistent fields are assigned a random number between 0 and 1. Fields whose assigned value is less than the desired reference class size will be included in the reference class. With the chosen reference class size of 80%, fields whose assigned random value is less than 0.8 are included in the reference class. The remaining fields form the test group. This means that, after 3000 iterations, the average reference class size will equal the specified value of 80%. For each single iteration, however, the reference class size may be smaller or larger than 80%. This functionality contributes to the random sampling procedure in terms of increasing the number of unique reference classes that can be defined for each year. The reference class size distribution observed for the ML mean-based random sampling of reference classes is provided in Figure 6.3. This histogram contains data for all first six years of production, amounting to 18 000 random samples.

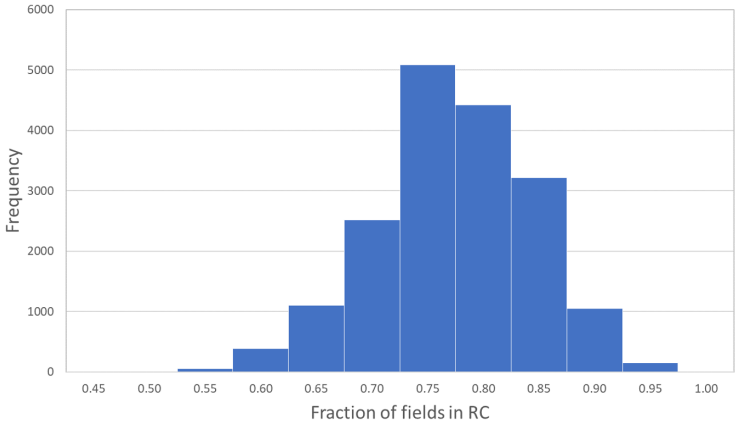
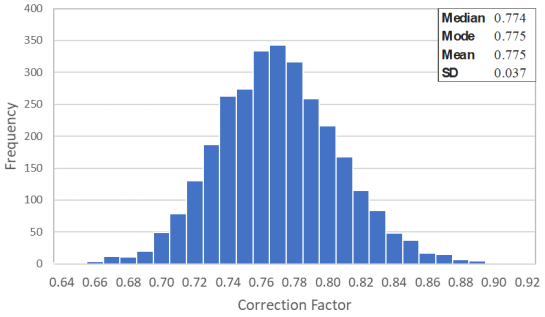


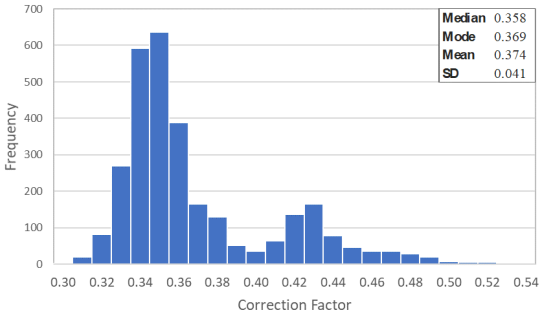
Figure 6.3: *Reference class size distribution for the F6Y*

6.2.2 Validity of using the mean correction factor

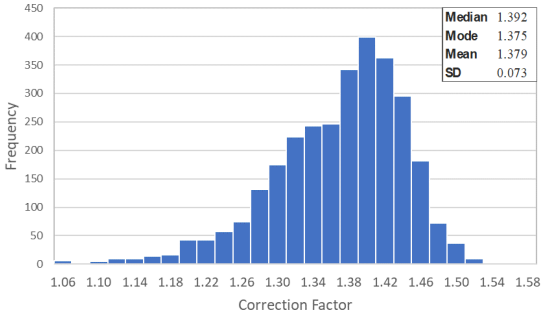
The random sampling of reference classes performed in Section 5.2.3 yields histograms for P90, P50 and P10 correction factors with a corresponding mean and standard deviation. Figure 6.4 shows the results for year 1.



(a) *P50 Correction Factor*



(b) *P90 Correction Factor*



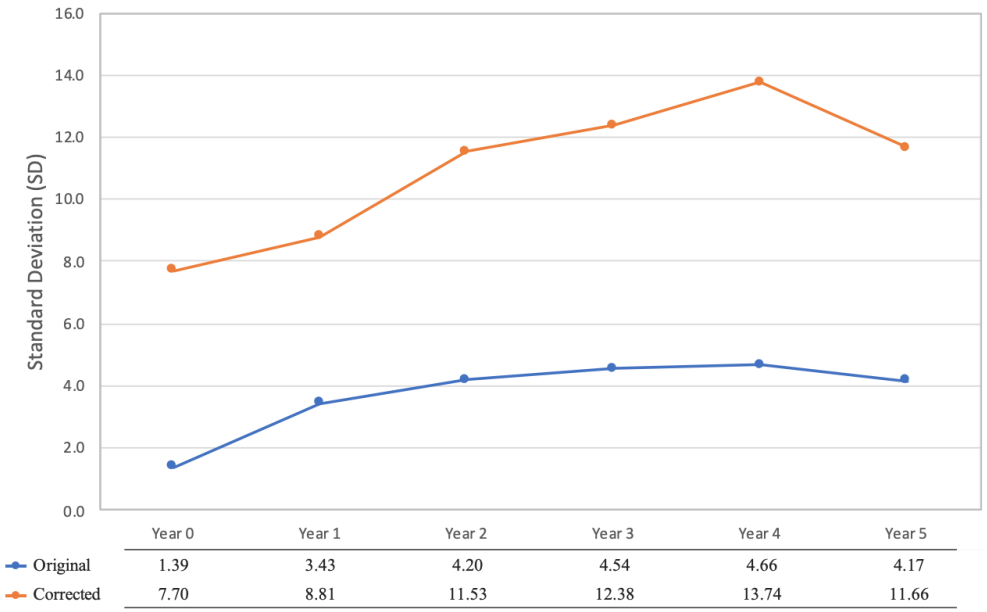
(c) *P10 Correction Factor*

Figure 6.4: *Histogram showing the distribution of P90, P50 and P10 correction factors for year 1 when running 3000 iterations with a reference class size of 80%*

Studying these histograms, the P50 correction factor is approximately normally distributed while the P90 and P10 correction factors are not symmetric. This is also demonstrated by looking at the relationships between the calculated median, mode and mean. One property of normally distributed data (and any other symmetrical distribution) is that the median, mode and mean are equal. In Figure 6.4b, illustrating the statistics for the P50 correction factor, little to no differences among the median, mode and mean are observed, strongly confirming normal distribution. Calculated statistics for the P90 and P10 correction factors, summarised in 6.4a and 6.4c, show that neither of these correction factors are normally distributed among the 3000 different iterations. These characteristics are common for all years in the F6Y. Although not identical for all years, the trend is clear; only the P50 correction factors can be characterized as normally distributed. Regardless, the differences between the median, mode and mean are minor. In turn, these findings justify the retrieval of the mean correction factors as robust and representative correction factors.

6.3 Corrected forecast calibration

The feasibility of adopting an outside view by applying RCF is tested through a series of tests. Results from the calibration procedure imply significantly enhanced model performance when correction factors are applied to the base estimate. Moreover, these results are also consistent with those retrieved through in-sample and out-of-sample testing. These tests provide evidence that RCF significantly enhances forecast performance and that the corrected distributions exhibit characteristics that are closely aligned with the characteristics of unbiased forecasts. Moreover, the annual standard deviation for the original and corrected distributions, obtained by summing the variance for each field and then taking the square root, is shown in Figure 6.5. This clearly illustrates the lack of regard given to uncertainty in the original production forecasts. Through assessing uncertainty of forecasted output with a broader perspective, which is demonstrated to be the result from RCF, development plans that are robust over a wide range of outcomes may be constructed. It is important to note that the sum of the variance of all fields is only equal to the annual variance if all forecasts are independent. However, production forecasts may be correlated with the oil price, and forecasts for different fields may be generated by the same forecaster and/or by the same software models. Due to lack of information on these concerns, we treat forecasts as independent in this work.



	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Original	1.39	3.43	4.20	4.54	4.66	4.17
Corrected	7.70	8.81	11.53	12.38	13.74	11.66

Figure 6.5: *Annual standard deviation for the original and corrected distributions*

Although providing promising results, it is important to note that the correction procedure performed in this thesis is performed with intentions of improving the overall forecast performance of development projects on the NCS. As it is impossible to know in advance which fields will successfully meet the original production forecasts and which fields will not, some development projects on the NCS may be subject to excessive correction. For these development projects, RCF results in corrected production forecasts that are lower than what will actually be produced and may, in turn, result in potentially profitable projects not getting accepted. However, looking at development projects on the NCS as a whole, or at an operator's portfolio of projects, the RCF methodology implemented in this thesis provides an overall better-informed basis for decision making and, thus, increases capital efficiency.

6.4 Base estimate sensitivity

Section 5.4 successfully illustrated that RCF yields approximately equal result whether it is based on the ML mean or on the P90 or P10 estimates provided in the dataset. However, these results are not representative of the dataset in its entirety. Recalling the results from the data scrubbing process performed in Section 3.1.2, the extent of the dataset was significantly reduced by 45 production years for the F6Y due to either inconsistent or missing data. Some fields had P90 and P10 estimates equal to the mean, which is supposed to reflect the expected value. For other fields, both P90 and P10 were reported as zero. Although removing inconsistent data was necessary to proceed with the analysis, the result is that important information may be ignored. The results would undoubtedly be different if no regard was given to the statistical reliability of the data.

The above cases essentially imply a lack of probabilistic production forecasts and a FID taken solely on the basis of the mean estimate with no regard to uncertainty ranges. Moreover, despite the flexibility of the metalog distribution, only 194 out of the 278 production years for the time shifted dataset were consistent with metalog distributions with an acceptable relative mean error of 2%. This may suggest that, for instances with limited available information, forecasters tend to rely on the rationalist approach, in which most of the effort is devoted to generating a base case in terms of a mean estimate. Although P90 and P10 production forecasts are also provided, they seem to be based on poor information. In fact, Figure 6.5 imply that the original uncertainty reflections, even for fields that were successfully fitted to metalog distributions, are too narrow. This signifies the need for stricter requirements for the probabilistic forecasts that are reported to the national authorities at the time of project sanction. Uncertainty assessments of

forecasts should be encouraged to be realistic (neither optimistic nor pessimistic) given the knowledge of the forecaster. Furthermore, top management should encourage increased uncertainty ranges related to production forecasts.

7 Conclusion

In this thesis, we have demonstrated that over the past 22 years, operators on the NCS exhibit significant optimism and overconfidence biases in their production forecasts. Analysing production data for 56 fields that were approved for development in this time period, we find that the forecasts provided at the time of FID are, as a general rule, both optimistic and overconfident. For the first six years of production, only 33% of actual observations fall inside the 80% confidence interval defined by the forecasted P90 and P10 fractiles, while 37% of actual observations exceed the P50 fractile − even after time shifting the data to reduce the impact of schedule delays.

To debias the original production estimates, an outside view is implemented by applying reference class forecasting. Correction factors for each of the F6Y are generated through random sampling of 3000 different reference classes. RCF forecast evaluation relative to perfect calibration shows that applying these correction factors to the original forecasts reduces the RMSE by up to 88%. Furthermore, in-sample and out-of-sample tests provide evidence that the corrected distributions are close to perfectly aligned with the characteristics of unbiased and well-calibrated production forecasts. Compared to only 33% for the original distributions, 77% of actual observations in the F6Y fall inside the 80% interval defined by corrected P90 and P10 production estimates. Moreover, 47% of the reported production data for the first six years of production exceed the corrected P50 estimate. Thus, the methodology developed and implemented in this thesis significantly reduces both optimism and overconfidence bias related to production forecasts for new development projects on the NCS.

As poorly informed production forecasts that lead to suboptimal decisions are commonly occurring in the oil and gas industry, this topic requires increased attention. For further studies, elimination of monthly schedule delays before performing RCF is recommended. Instead of issuing monthly production forecasts from the operators − which is a rather demanding request − one possible approach is to assume a trend (for example linear) in production for each year. Data of startup month for each field used in combination with interpolation of yearly actual production data may, then, further reduce the effect of schedule delays. Moreover, as this work has demonstrated that the required adjustment of forecasts exhibits annual variations, it is recommended to continue studying annual forecasts.

References

Arora, J. S. (2016). Introduction to Optimum Design, 2012. *Google Scholar*, pages 1–64.

Bentley, M. (2016). Modelling for comfort? *Petroleum Geoscience*, 22(1):3–10.

Bentley, M. and Smith, S. (2008). Scenario-Based Reservoir Modelling: The Need for More Determinism and Less Anchoring. *Geological Society, London, Special Publications*, 309(1):145–159.

Bratvold, R. B., Begg, S. H., Rasheva, S., et al. (2010). A New Approach to Uncertainty Quantification for Decision Making. In *SPE Hydrocarbon Economics and Evaluation Symposium*. Society of Petroleum Engineers.

Bratvold, R. B., Mohus, E., Petutschnig, D., Bickel, E., et al. (2020). Production Forecasting: Optimistic and Overconfident—Over and Over Again. *SPE Reservoir Evaluation & Engineering*.

Chen, James (2020). Time Value of Money (TVM). https://www.investopedia.com/terms/t/timevalueofmoney.asp. Accessed: 14. April 2020.

Cunnane, C. (1978). Unbiased Plotting Positions—a Review. *Journal of hydrology*, 37(3-4):205–222.

Flyvbjerg, B. (2006). From Nobel Prize to Project Management: Getting Risks Right. *Project management journal*, 37(3):5–15.

Flyvbjerg, B. (2007a). Eliminating Bias in Early Project Development Through Reference Class Forecasting and Good Governance. *KJ Sunnev rag, ed. Beslutninger pa svakt informasjonsgrunnlag. Trondheim, Norway*, pages 90–110.

Flyvbjerg, B. (2007b). Policy and Planning for Large-Infrastructure Projects: Problems, Causes, Cures. *Environment and Planning B: planning and design*, 34(4):578–597.

Flyvbjerg, B., Garbuio, M., and Lovallo, D. (2009). Delusion and Deception in Large Infrastructure Projects: Two Models for Explaining and Preventing Executive Disaster. *California management review*, 51(2):170–194.

Gan, F., Koehler, K. J., and Thompson, J. C. (1991). Probability Plots and Distribution Curves for Assessing the Fit of Probability Models. *The American Statistician*, 45(1):14–21.

Haslwanter, T. (2015). Characterizing a Distribution. http://work.thaslwanter.at/Stats/html/statsDistributions.html?fbclid=IwAR1-QzKurTlkw54vi7JoRH7eNOWioDyr9OMmD-NxtC-6ycFcoyUoedWxfWw.

Kahneman, D. (1979). Prospect Theory: An Analysis of Decisions Under Risk. *Econometrica*, 47:278.

Kahneman, D. and Tversky, A. (1977). Intuitive Prediction: Biases and Corrective Procedures. Technical report, Decisions and Designs Inc Mclean Va.

Keelin, T. W. (2016). The Metalog Distributions. *Decision Analysis*, 13(4):243–277.

Kenton, Will (2020). Standard Error. https://www.investopedia.com/terms/s/standard-error.asp. Accessed: 01. June 2020.

Kim, S., Shin, H., Joo, K., and Heo, J.-H. (2012). Development of plotting position for the general extreme value distribution. *Journal of Hydrology*, 475:259–269.

Kongsnes, E. (2015). Goliat er Forsinket av Dårlig Vær. https://www.aftenbladet.no/aenergi/i/gLMOJ/goliat-er-forsinket-av-darlig-vr.

Leleur, S., Salling, K. B., Pilkauskiene, I., and Nicolaisen, M. S. (2015). Combining Reference Class Forecasting With Overconfidence Theory for Better Risk Assessment of Transport Infrastructure Investments. *European Journal of Transport and Infrastructure Research*, 15(3):362–375.

Meddaugh, W., Meddaugh, W., McCray, B., et al. (2017). Quantitative Assessment of the Impact of Sparse Data and Decision Bias on Reservoir Recovery Forecasts. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Ministry of Petroleum and Energy (2010). Guidelines for PDO and PIO. Oslo, Norway: Ministry of Petroleum and Energy.

Mohus, E. (2018). Over Budget, Over Time, and Reduced Revenue, Over and Over Again-An Analysis of the Norwegian Petroleum Industry's Inability to Forecast Production. Master's thesis, University of Stavanger, Norway.

Nandurdikar, N. S., Wallace, L., et al. (2011). Failure to produce: An investigation of deficiencies in production attainment. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Norwegian Petroleum Directorate (2000). PDO/PIO Guidelines. Stavanger, Norway: Norwegian Petroleum Directorate.

Norwegian Petroleum Directorate (2011). Preliminary Production Figures for November 2011. Stavanger, Norway: Norwegian Petroleum Directorate.

Norwegian Petroleum Directorate (2017). Large Oil and Gas Deposits in Tight Reservoirs. Stavanger, Norway: Norwegian Petroleum Directorate.

Norwegian Petroleum Directorate (2019). 3- Resource Classification. Stavanger, Norway: Norwegian Petroleum Directorate.

Norwegian Petroleum Directorate (2020). Reporting to the Revised National Budget 2020. Stavanger, Norway: Norwegian Petroleum Directorate.

Oil and Gas Authority (2017). Lessons Learned from UKCS Oil and Gas Projects 2011-2016.

PetroWiki (2016). Probabilistic Verses Deterministic in Production Forecasting. https://petrowiki.org/Probabilistic_verses_deterministic_in_production_forecasting. Accessed: 19. February 2020.

PetroWiki (2020). Reservoir Simulation. https://en.wikipedia.org/wiki/Reservoir_simulation. Accessed: 11. April 2020.

Renard, P., Alcolea, A., and Gingsbourger, D. (2013). Stochastic versus deterministic approaches. In *Environmental Modelling: Finding Simplicity in Complexity, Second Edition (eds J. Wainwright and M. Mulligan)*, pages 133–149. Wiley Online Library.

Rey, S. J. (2015). Mathematical Models in Geography. In Wright, J. D., editor, *International Encyclopedia of the Social and Behavioral Sciences (Second Edition)*, pages 785 – 790. Elsevier, Oxford, second edition edition.

Skodje, M. and Steneberg, I. J. (2011). Værtrøbbel for Yme-plattformen. https://www.nrk.no/rogaland/vaertrobbel-for-yme-plattformen-1.7550005.

Stangeland, G. (2015). Total Utsetter Martin Linge med Ett År. https://sysla.no/?p=2247440.

Welsh, M. B., Begg, S. H., Bratvold, R. B., et al. (2007). Modelling the economic impact of common biases on oil and gas decisions. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Welsh, M. B., Begg, S. H., et al. (2010). Don't Let It Weigh You Down: How to Benefit From Anchoring. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Yeten, B., Hovadik, J. M., and Muron, P. (2015). System and Method for Forecasting Production From a Hydrocarbon Reservoir. US Patent 9,043,188.

Yound, C. (2020). Excel Solver: Which Method Should I Use. https://www.engineerexcel.com/excel-solver-solving-method-choose/.

Appendices

A The metalog distribution

Field-by-field metalog distributions are generated in Section 5.2 using the bounded metalog distribution, while random sampling of reference classes, and its related out-of-sample test, performed in Section 5.2.3, utilises the semi-bounded metalog distribution to generate annual distributions of normalised production. However, to simplify the derivation process, this section will also present the unbounded metalog distribution.

Unbounded metalog distribution

Definition 1. metalog quantile function with n terms:

$$M_n(y; \mathbf{x}, \mathbf{y}) = \tag{A.1}$$

$$\begin{cases} a_1 + a_2 \ln\left(\frac{y}{1-y}\right) & n = 2 \\ a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3(y-0.5)\ln\left(\frac{y}{1-y}\right) & n = 3 \\ a_1 + a_2 \ln\left(\frac{y}{1-y}\right) + a_3(y-0.5)\ln\left(\frac{y}{1-y}\right) + a_4(y-0.5) & n = 4 \end{cases}$$

Terms beyond $n = 4$ can be added from the following:

$$\begin{cases} M_{n-1} + a_n(y-0.5)^{\frac{n-1}{2}} & \text{for odd} \\ M_{n-1} + a_n(y-0.5)^{\frac{n}{2}-1}\ln\left(\frac{y}{1-y}\right) & \text{for even} \end{cases}$$

The cumulative probability y ranges from $0 < y < 1$. The x and y coordinates of the CDF data are found in column vectors $\mathbf{x} = (x_1, ..., x_m)$ and $\mathbf{y} = (y_1, ..., y_m)$ for length $m \geq n$, where at least n of the y_i's are distinct. The a_i's are real constants that are utilised as a combination of parameter-substitutions and series expansion for adding more shape flexibility as the number of terms (n) increases (Keelin, 2016).

Definition 2. metalog PDF:

Differentiating Equation A.1 with respect to y and inverting the result yields the metalog probability density function (PDF):

$$m_n(y) = \tag{A.2}$$

$$\begin{cases} \dfrac{y(1-y)}{a_2} & n = 2 \\ \left[\dfrac{a_2}{y(1-y)} + a_3\Big(\dfrac{y-0.5}{y(1-y)} + \ln\big(\dfrac{y}{1-y}\big)\Big)\right]^{-1} & n = 3 \\ \left[\dfrac{a_2}{y(1-y)} + a_3\Big(\dfrac{y-0.5}{y(1-y)} + \ln\big(\dfrac{y}{1-y}\big)\Big) + a_4\right]^{-1} & n = 4 \end{cases}$$

Terms beyond $n = 4$ can be added from the following:

$$\begin{cases} \left[(m_{n-1}(y))^{-1} + a_n\Big(\dfrac{n-1}{2}\Big)(y - o.5)^{\frac{n-3}{2}}\right]^{-1} & \text{for odd} \\ \left[(m_{n-1}(y))^{-1} + a_n\left(\dfrac{(y-0.5)^{\frac{n}{2}-1}}{y(1-y)} + \Big(\dfrac{n}{2}-1\Big)(y-0.5)^{\frac{n}{2}-2}\ln\Big(\dfrac{y}{1-y}\Big)\right)\right]^{-1} & \text{for even} \end{cases}$$

It is noticable that the PDF $m_n(y)$ is expressed as a function of the cumulative probability y. In the customary operation of plotting the PDF, the metalog sheet by Keelin is arranged so that the horizontal axis presents production estimates $M_n(y)$ while the vertical axis presents the associated probability density, with y varying $\in (0, 1)$ to retrieve the corresponding values on both axes.

Semi-bounded metalog distribution

To retrieve the log metalog quantile function with n terms for the semi-bounded metalog distribution, one has to set a lower bound b_l for x. Suppose that $z = \ln(x - b_l)$ is metalog distributed according to Equation A.1. Setting $\ln(x - b_l)$ equal to Equation A.1 and solving for x gives the following expression for the log metalog quantile function:

$$M_n^{log}(y; \mathbf{x}, \mathbf{y}, b_l) = b_l + e^{M_n(y)} \qquad 0 < y < 1 \tag{A.3}$$

$\mathbf{z} = (ln(x_1 - b_l), ..., ln(x_m - b_l))$ is a column vector, where $\mathbf{x} = (x_1, ..., x_m)$, for all $m \geq n$. Further, each $x_i > b_l$, and $\mathbf{y} = (y_1, ..., y_m)$, where $0 < y_i < 1$ for each y_i, at least n of the y_i's are distinct. When $y = 0$, the expression retrieves b_l.

Differentiating Equation A.3 with respect to y and inverting the result yields the log metalog PDF:

$$m_n^{log}(y) = m_n(y)e^{-M_n(y)} \qquad 0 < y < 1 \tag{A.4}$$

where $m_n(y)$ and $M_n(y)$ is equations (A.2) (A.1), respectively. The log metalog feasibility condition is $m_n^{log}(y) > 0$ for all $y \in (0, 1)$. When $y = 0$, $m_n^{log}(y)$ equals 0.

Bounded metalog distribution

The logit metalog distribution to be utilised in the upcoming section, draws its benefits from the possibility to define a lower and upper bound for estimated annual production. These will be denoted by b_l and b_u, respectively, where $b_u > b_l$. The logit metalog distribution is the metalog transform that corresponds to $z = \text{logit}(x) = \ln\left(\frac{x - b_l}{b_u - x}\right)$ being metalog distributed. Setting $\ln\left(\frac{x - b_l}{b_u - x}\right)$ equal to equation (A.1) and solving for x yields the logit metalog quantile function with n terms:

$$M_n^{logit}(y; \mathbf{x}, \mathbf{y}, b_l, b_u) = \frac{b_l + b_u e^{M_n(y)}}{1 + e^{M_n(y)}} \qquad 0 < y < 1 \qquad \text{(A.5)}$$

$\mathbf{z} = \left(\ln\left(\frac{x_1 - b_l}{b_u - x_1}\right), ..., \ln\left(\frac{x_m - b_l}{b_u - x_m}\right) \right)$. Where $\mathbf{x} = (x_1, ..., x_m)$, $b_l < x_i < b_u$ for each x_i, and $\mathbf{y} = (y_1, ..., y_m)$, $0 < y_i < 1$ for each y_i. Differentiating Equation A.5 with respect to y and inverting the result yields the logit metalog PDF:

$$m_n^{logit}(y) = m_n(y) \frac{(1 + e^{M_n(y)})^2}{(b_u - b_l)e^{M_n(y)}} \qquad 0 < y < 1 \qquad \text{(A.6)}$$

where $m_n(y)$ is Equation A.2 and $M_n(y)$ is Equation A.1. The logit metalog feasibility condition is $m_n^{logit}(y) > 0$ for all y.

B Supplementary results

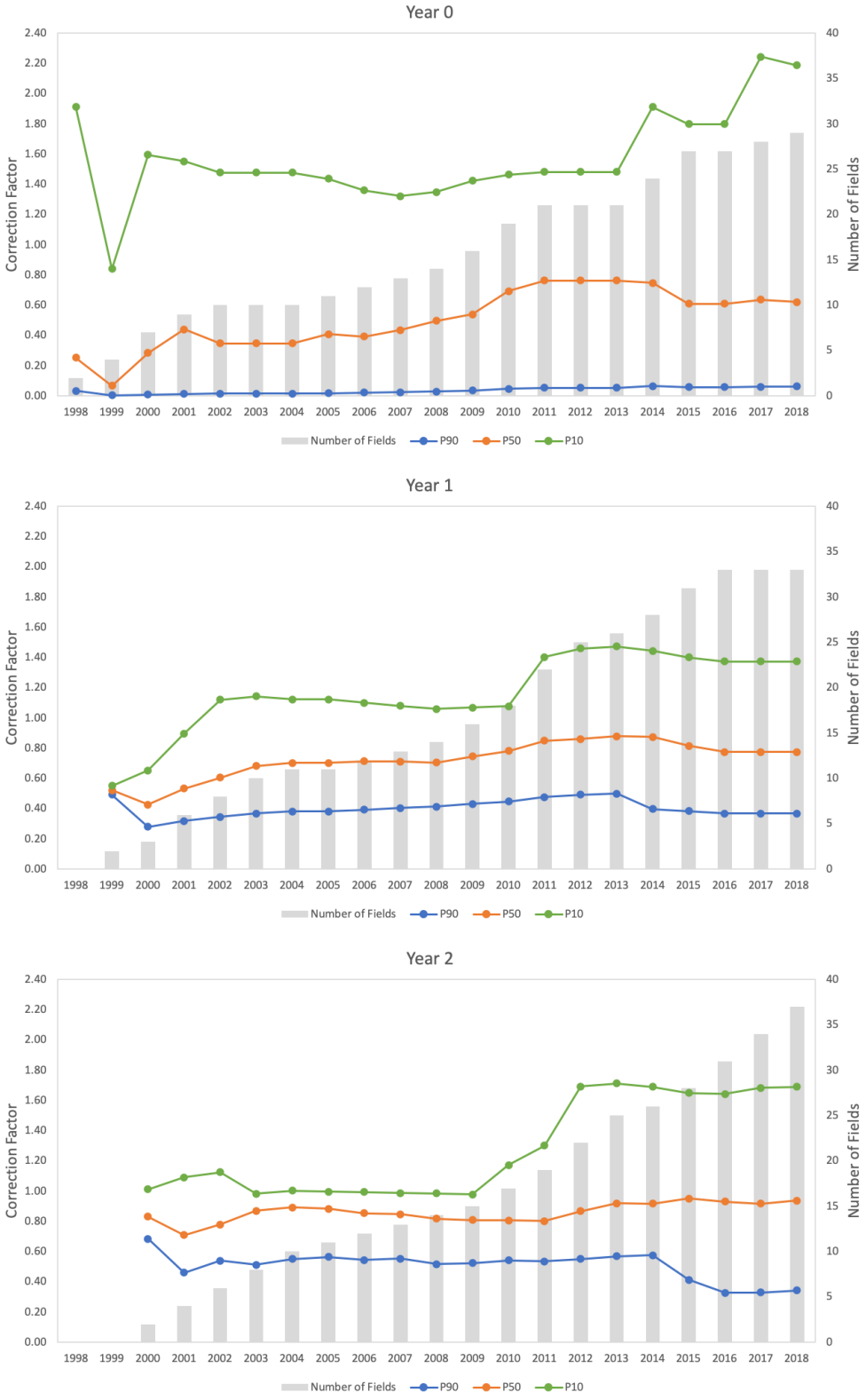


Figure B.1: *Correction factors retrieved from progressive reference class forecasting for year 0, 1 and 2*

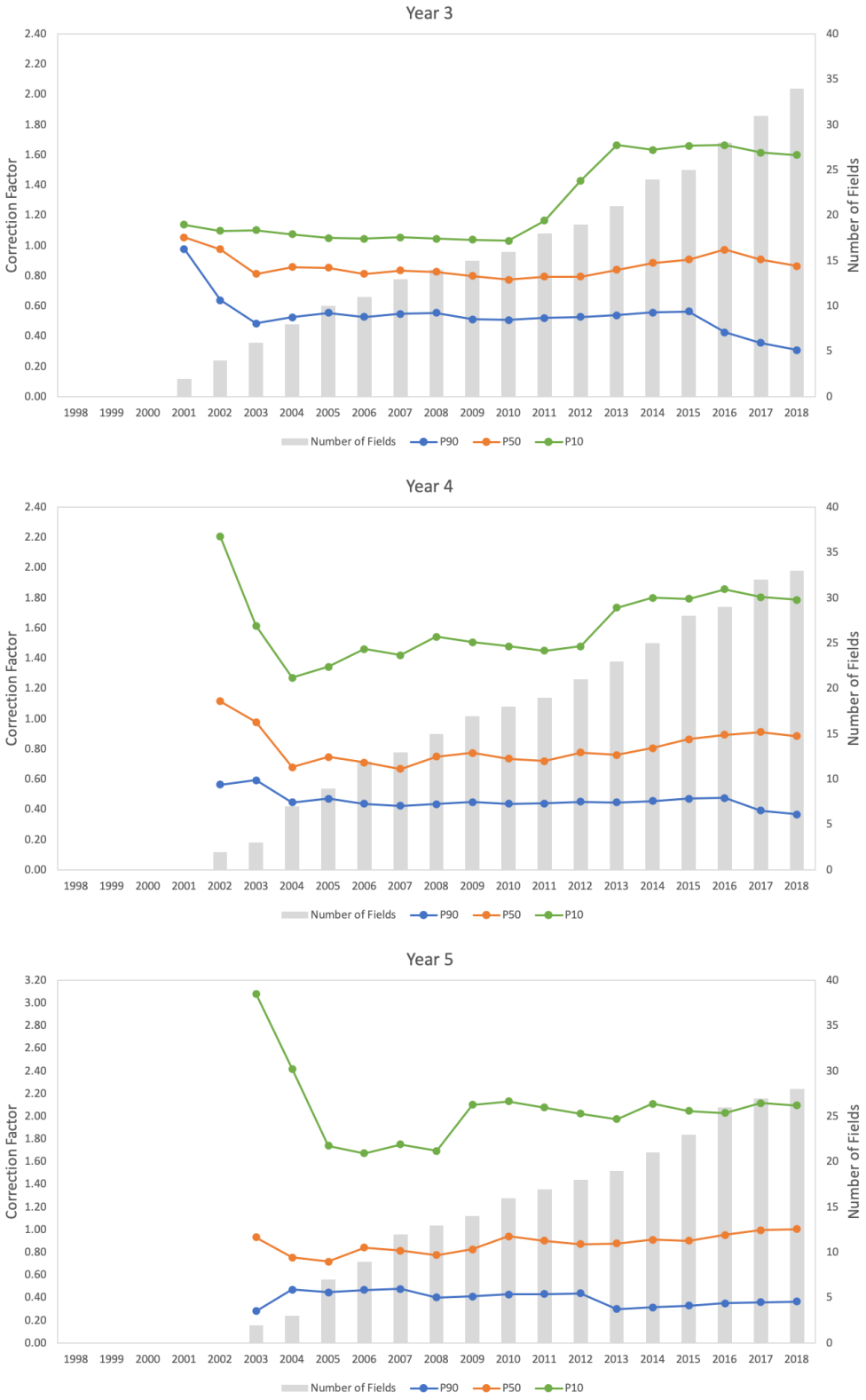


Figure B.2: *Correction factors retrieved from progressive reference class forecasting for year 3, 4 and 5*

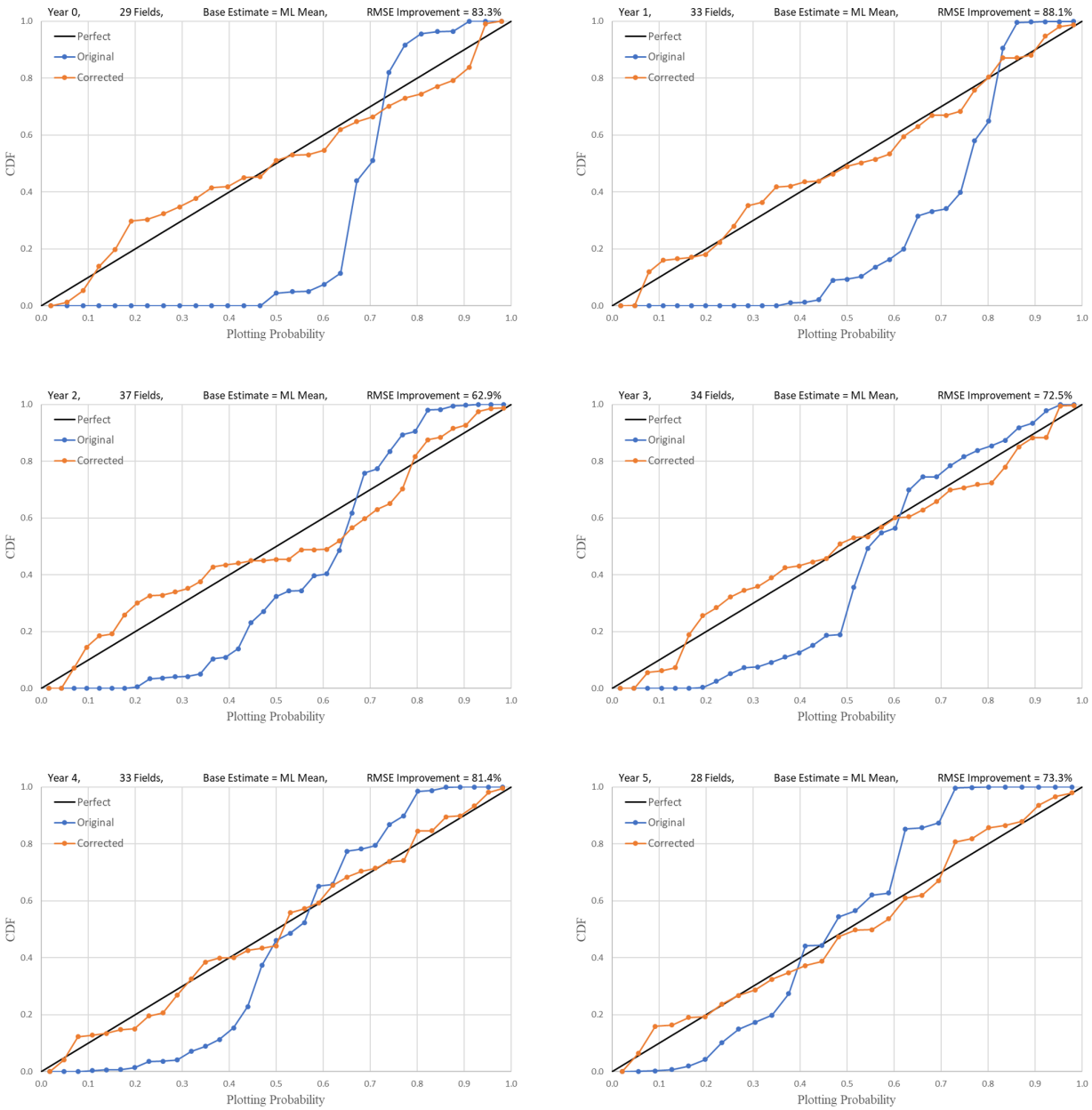


Figure B.3: *Calibration plots for ML mean-based RCF*

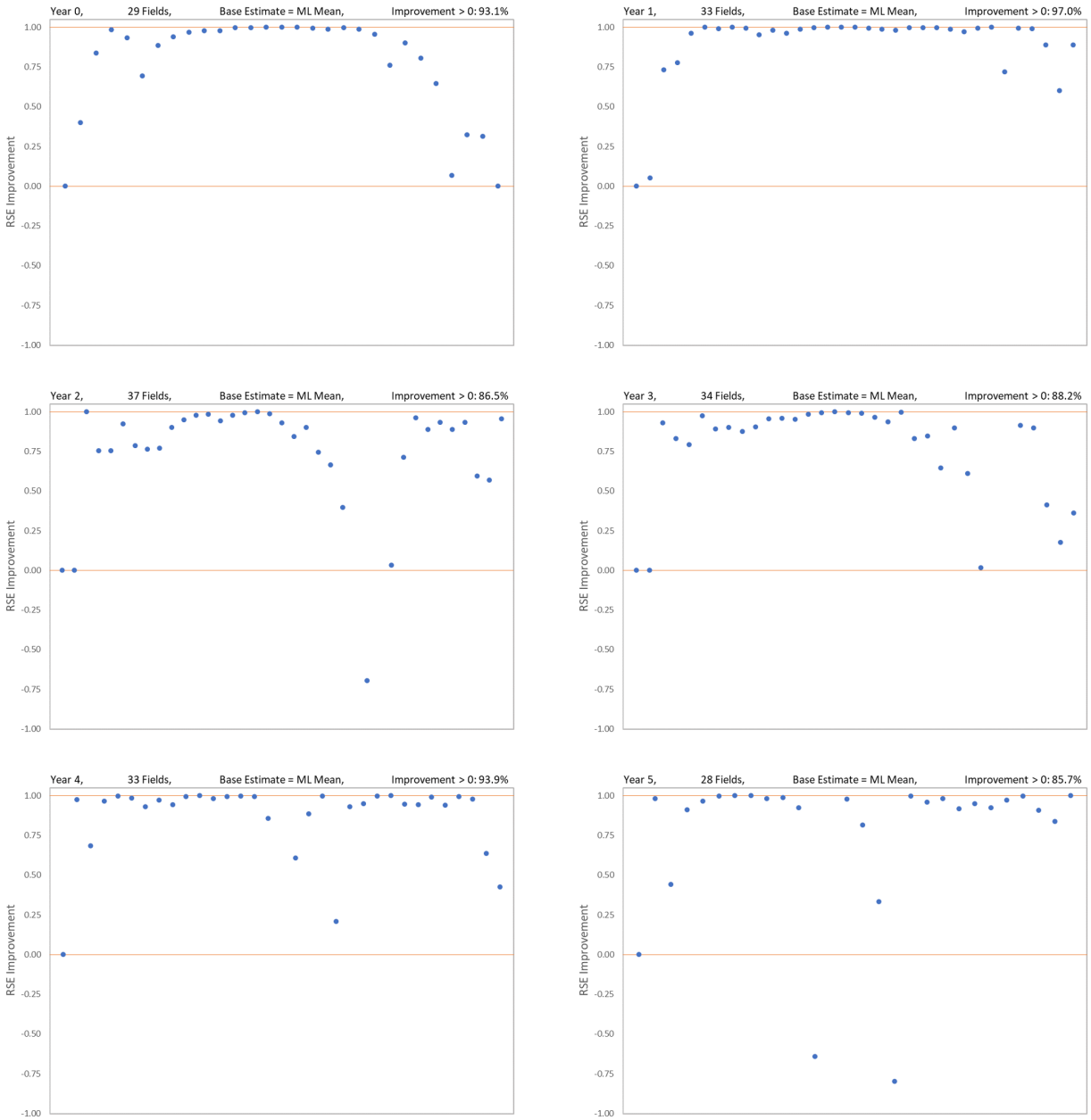


Figure B.4: *Field-by-field RSE improvement for ML mean-based RCF*

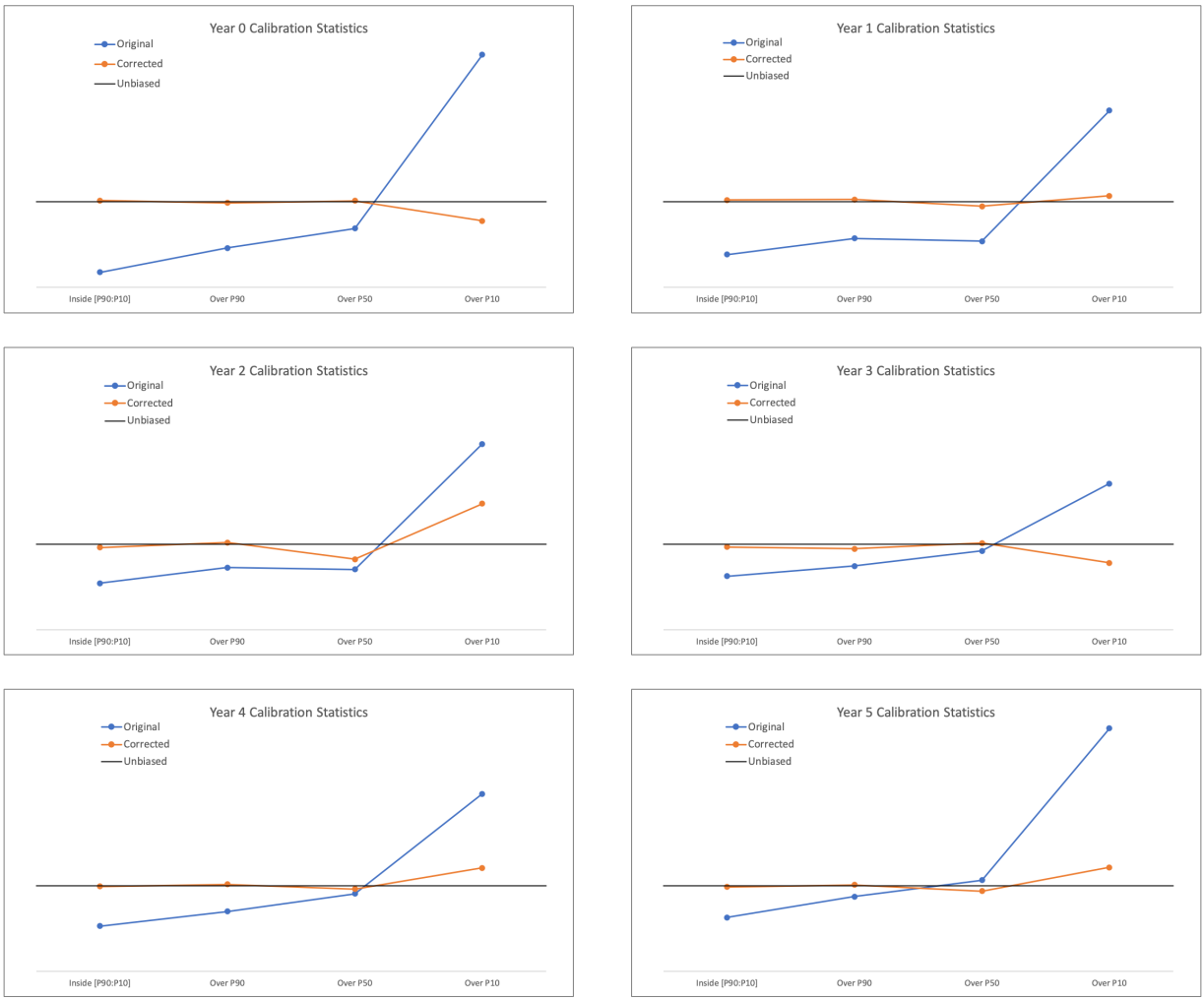


Figure B.5: *Results from out-of-sample test for ML mean-based RCF, showing calibration statistics normalised by the characteristics of unbiased forecasts*

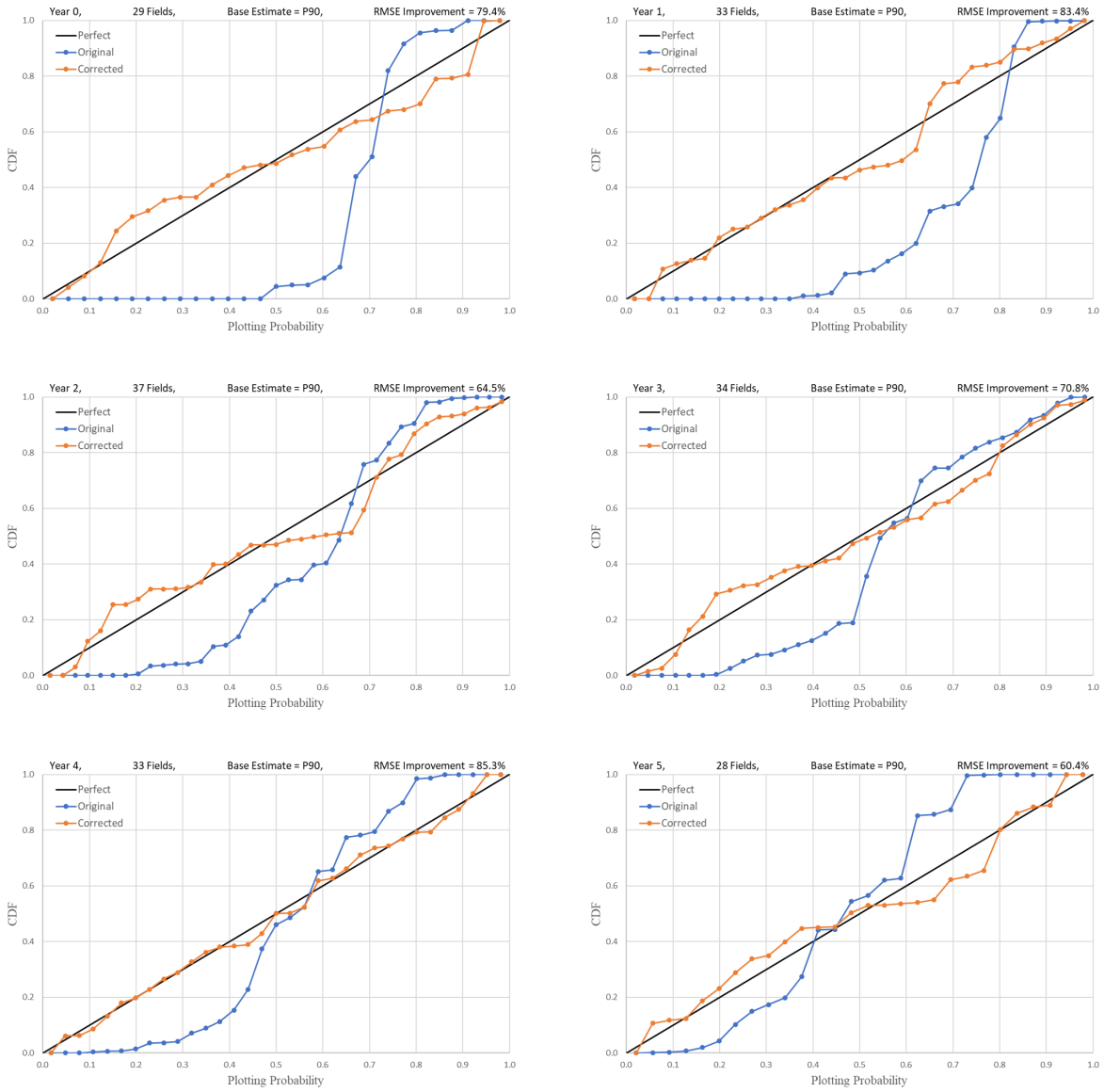


Figure B.6: *Calibration plots for P90-based RCF*

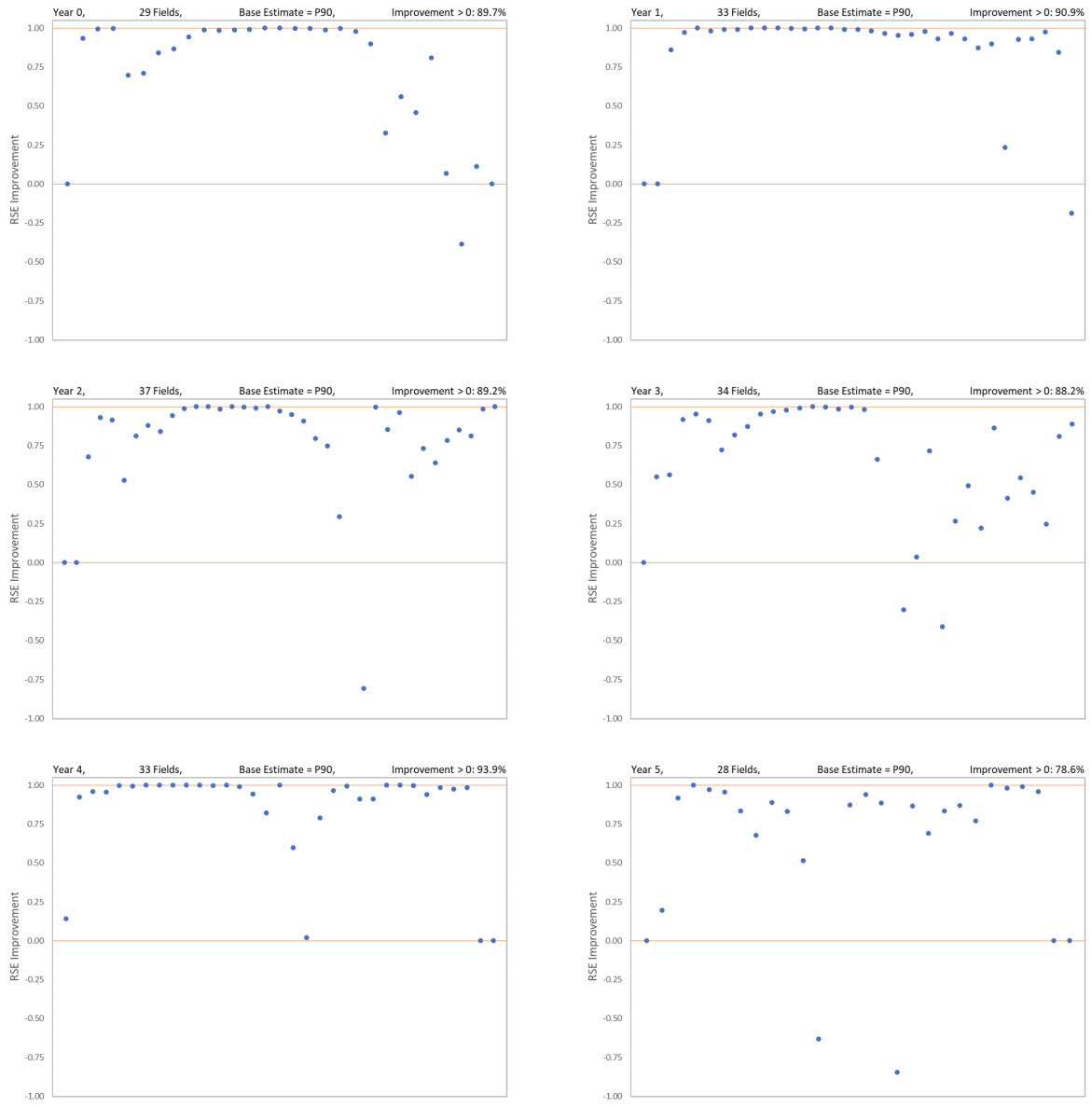


Figure B.7: *Field-by-field RSE improvement for P90-based RCF*

Table B.1: *Overview of the total number of fields and the number of fields with a positive and no or negative RSE improvement when applying P90-based RCF*

	Field-by-field RSE Improvement (P90-based RCF)					
Fields	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Total	29	33	37	34	33	28
Improvement > 0	26	30	33	30	31	22
Improvement ≤ 0	3	3	4	4	2	6

Table B.2: *Results from out-of-sample test for P90-based RCF, showing the average calibration statistics for the test group before and after correction*

	Average Calibration Statistics for TG Before P90-based RCF						
Actual Production	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	**Unbiased**
Inside [P90:P10]	13%	31%	44%	50%	43%	50%	**80%**
Over P90	42%	52%	65%	67%	64%	79%	**90%**
Over P50	35%	27%	35%	47%	46%	55%	**50%**
Over P10	28%	21%	21%	17%	21%	28%	**10%**
	Average Calibration Statistics for TG After P90-based RCF						
Actual Production	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	**Unbiased**
Inside [P90:P10]	83%	76%	72%	74%	78%	83%	**80%**
Over P90	90%	92%	91%	87%	88%	93%	**90%**
Over P50	49%	42%	42%	47%	48%	52%	**50%**
Over P10	8%	16%	18%	14%	10%	11%	**10%**

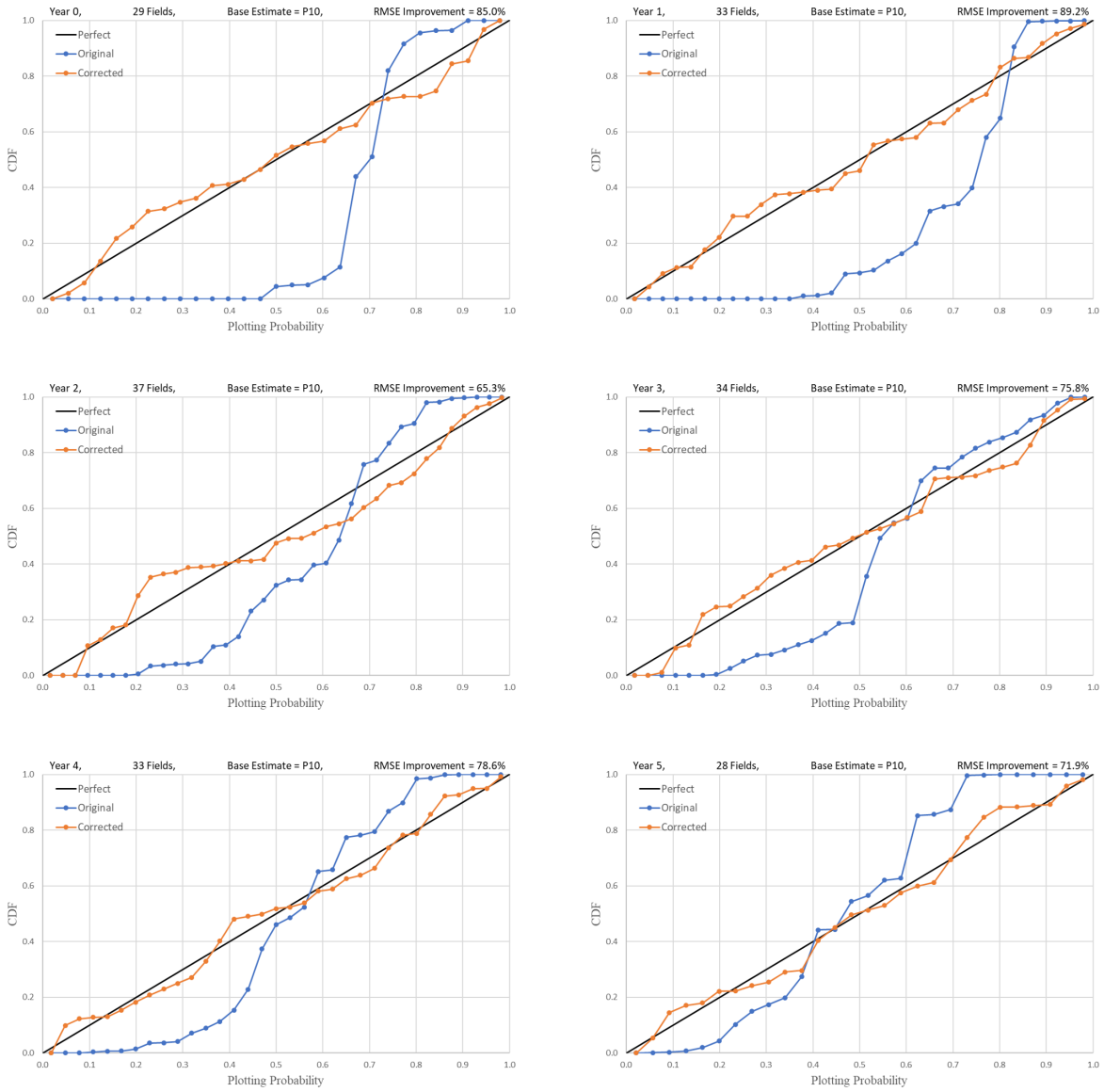


Figure B.8: *Calibration plots for P10-based RCF*

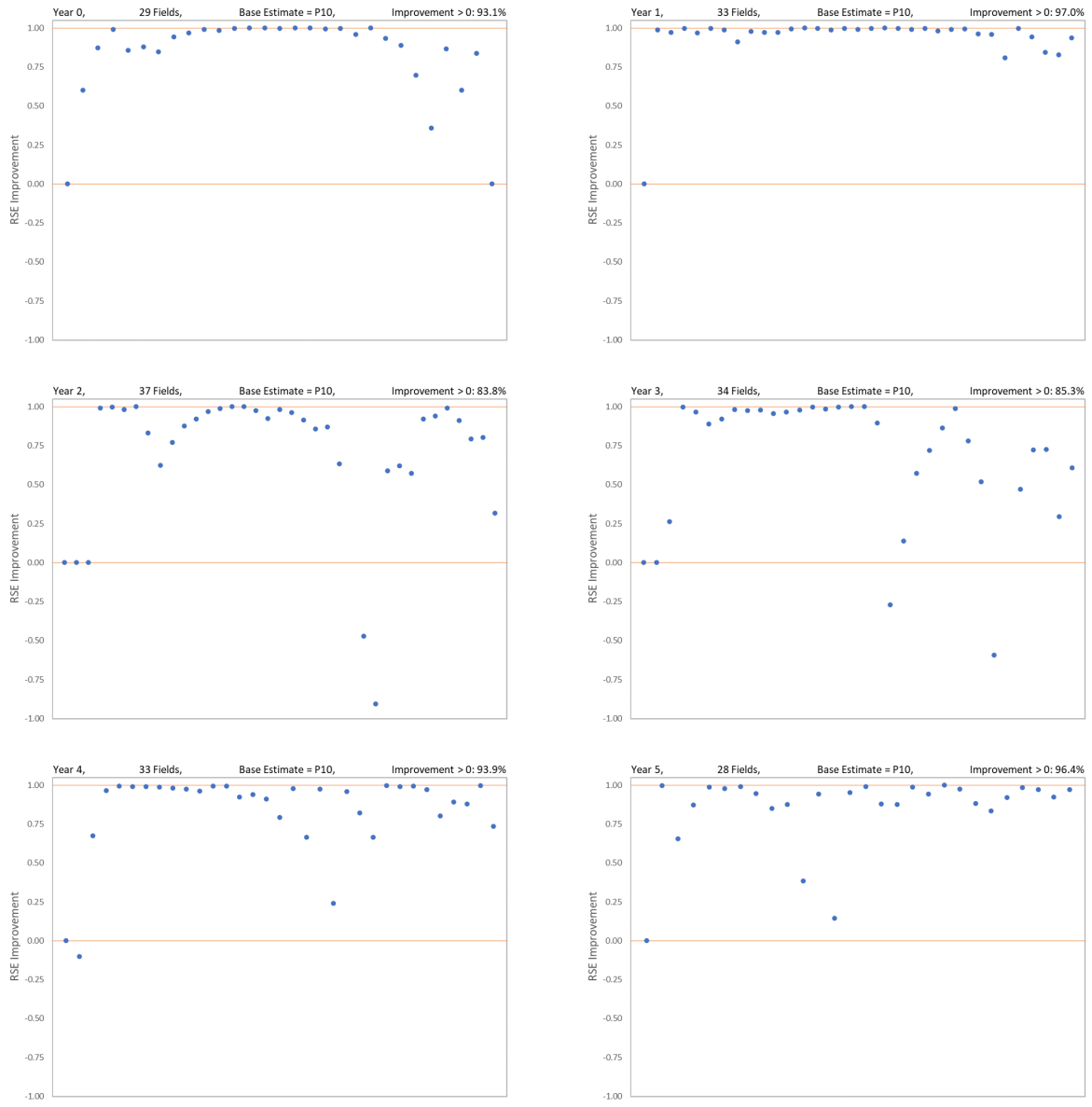


Figure B.9: *Field-by-field RSE improvement for P10-based RCF*

Table B.3: *Overview of the total number of fields and the number of fields with a positive and no or negative RSE improvement when applying P10-based RCF*

	Field-by-field RSE Improvement (P10-based RCF)					
Fields	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Total	29	33	37	34	33	28
Improvement > 0	27	32	31	29	31	27
Improvement ≤ 0	2	1	6	5	2	1

Table B.4: *Results from out-of-sample test for P10-based RCF, showing the average calibration statistics for the test group before and after correction*

	Average Calibration Statistics for TG Before P10-based RCF						
Actual Production	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	**Unbiased**
Inside [P90:P10]	14%	30%	43%	50%	42%	50%	**80%**
Over P90	42%	52%	65%	67%	64%	78%	**90%**
Over P50	35%	28%	35%	46%	46%	54%	**50%**
Over P10	28%	22%	22%	18%	22%	29%	**10%**
	Average Calibration Statistics for TG After P10-based RCF						
Actual Production	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5	**Unbiased**
Inside [P90:P10]	80%	78%	78%	77%	78%	79%	**80%**
Over P90	88%	91%	90%	89%	93%	91%	**90%**
Over P50	52%	49%	45%	50%	54%	50%	**50%**
Over P10	9%	13%	12%	12%	16%	12%	**10%**