Field Features Do Not Explain Greenfield Production Forecasting Bias

E. Nesvold[1]* and R. B. Bratvold[1]

[1]UiS

Summary

Over the past years, it has become clear that greenfield oil production forecasts are subject to strong optimism and overprecision biases: Significant early production shortfalls are the rule rather than the exception and the elicited uncertainty range is generally too narrow. This has large negative consequences for the net present valuation of such investments. A study from 2011 based on post-factum evaluation of greenfield production forecasts suggests that there is a causal relationship between certain project/field characteristics and production attainment (i.e., optimism bias). However, while self-reported causes of failure may provide interesting insights, such analyses are subject to cognitive hindsight bias. It is therefore necessary to test such claims more rigorously. Research on megaprojects in other industries suggests that forecasting bias is omnipresent but is stronger in certain circumstances (e.g., information and communication technology projects are subject to larger cost overruns than road construction projects). An important question is therefore whether there are combinations of field characteristics/features which can be measured objectively, such as field size, reservoir complexity, oil prices, and lack of drillstem tests (DSTs), etc., and which can be shown to have predictive power of overly optimistic and overconfident production forecasts.

The data set in this study consists of 71 greenfield oil production forecasts at project sanction on the Norwegian Continental Shelf (NCS), with production starting between 1995 and 2020. Each forecast consists of a triplet of production curves which represent the statistical p_{10}, the expectation, and the statistical p_{90}. The forecasts are compared with actual production data. Metadata about the fields gathered from the Norwegian Petroleum Directorate (NPD) are used to establish 16 informative field features, from field reserves to the number of appraisal wells per unit area. These features are tested for predictive power, both individually and simultaneously, of optimism bias and of a general forecast quality metric.

First, we show that value erosion caused by time overruns and production shortfalls are both significant, but that the relative importance of effects after production start is higher. Second, none of the tested machine learning models show any predictive power of forecasting bias. Because of this systematic presence of bias in the production forecasts, we argue that oil and gas companies need to make important changes to their decision-making workflows to take into account well-documented research findings on cognitive and organizational bias from the past decades, instead of the ever-increasing model complexity. Illustratively, as a final point, we show that a no-skills-involved reference class forecast based on empirical production curves from abandoned fields outperforms operators' own greenfield forecasts. This approach may perhaps serve as a useful benchmark for future forecasts.

Introduction

Forecasting experts in the oil and gas industry often come up with explanations for why specific forecasts are poor. Reservoir engineers involved in production forecasting are often quick to point out field-related challenges. Other often-cited reasons range from "lack of data," "poor data," and "wrong data" to poor uncertainty modeling in other parts of the organization and external forces outside the control of the organizations. Such reasoning often suffers from hindsight bias. An important question is whether it is possible to relate forecast bias to particular field features. An often-cited study by Nandurdikar and Wallace (2011) is based on self-reported causes by project teams and suggests that appraisal strategy and hydrocarbon API gravity are two examples of features that can be statistically related to low production attainment. However, these claims are based on aggregated data and do not show statistically convincing results. An important goal of this study is to assess whether such explanations are statistically meaningful.

It is the inherent uncertainty in exploration and production projects that creates investment opportunities and competitive advantages. Clearly, for realistic valuation at the time of project sanction, it is important that the uncertainty quantification models used to support sanctioning decisions are unbiased. If we can identify features which are predictive of bias, it would help us avoid investing in poor projects and the ensuing erosion of value. There is a long list of well-known biases which have been identified in behavioral economics, project management (Flyvbjerg 2021), and cognitive psychology (Kahneman 2011). In this work, we focus on two types of bias: *optimism bias* (the expected value is too high) and *overprecision bias* (Hoffrage 2016) (after adjusting for optimism bias, the forecasted range of possible outcomes is too narrow). First, empirical evidence shows that one tends to assign higher probabilities to favorable outcomes and to underweight risk factors in project planning (De Reyck et al. 2017). According to Kahneman et al. (2011), optimism bias may be considered as the most significant form of human bias. Second, people and organizations are usually more confident in their judgment than the underlying facts allow for (Griffin and Tversky 1992). Research shows that quantitative and regular feedback on forecast performance is necessary to learn from and to calibrate forecasts. The saying about expert judgment is often fitting: "Often wrong, but rarely in doubt" (Griffin and Tversky 1992). **Fig. 1** illustrates the general problem in terms of probability densities.

It is a well-known fact that megaprojects are subject to overruns with respect to cost and time (Flyvbjerg 2014; Morris 1990). This is no less true for oil and gas projects, as shown in several studies (e.g., Esmaeili and Kashani 2022; Oglend and Osmundsen 2016). On the other hand, the effect of benefit shortfalls have been much less scrutinized. To these authors' best knowledge, only three studies on the performance of field-level production forecasts have been published. A study by Nandurdikar and Wallace (2011) using data from Independent Project Analysis Inc. shows that after 4 years of production, cumulative production attainment is on average approximately 80% of planned production. Similarly, a study of probabilistic production forecasts from 32 fields on the NCS by Bratvold et al. (2020)

*Corresponding author; email: erik.nesvold@gmail.com

Original SPE manuscript received for review 1 June 2022. Revised manuscript received for review 4 October 2022. Paper (SPE 212834) peer approved 13 October 2022.

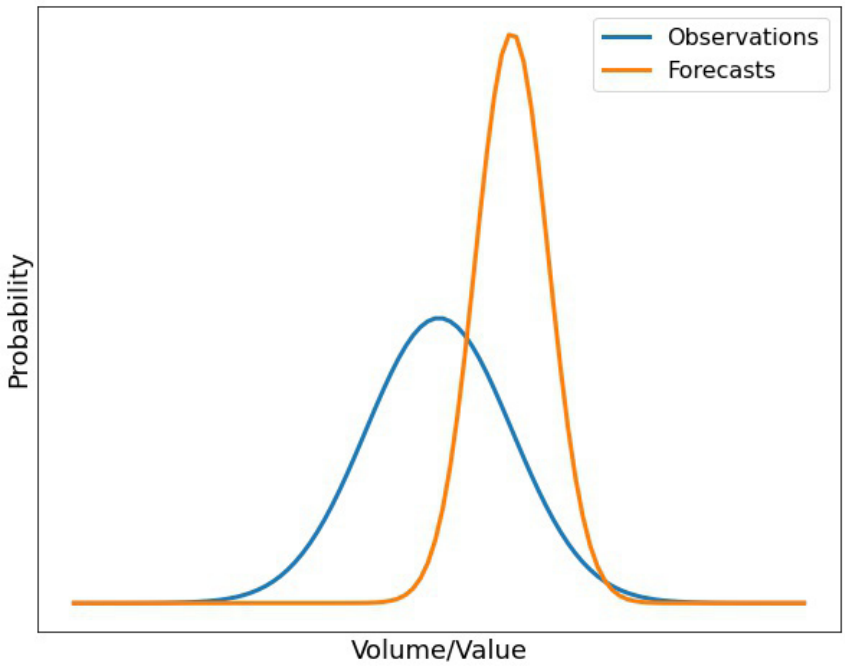

Fig. 1—Illustration of the general problem with probabilistic forecasts of production and revenue streams: too optimistic and too confident.

shows that 59% of the fields had produced below the operators' p_{10} estimate (i.e., 10% estimated probability of being below) after 4 years, and attainment relative to expected production was in accordance with Nandurdikar and Wallace's estimate. Nesvold and Bratvold (2022) extend this analysis and suggest statistical calibration of future forecasts based on empirical data. In all three aforementioned studies, the results are based on forecasted production start being aligned with actual production start. When delays in production start are also accounted for, the benefit shortfalls are significantly higher—as shown in this article. The consequences are obvious and negative for the valuation of investments in oil and gas fields.

In this work, we focus on cumulative production after 4 calendar years—both because of the importance of ramping up production quickly to maximize the value of the discounted revenue stream and because oil fields usually receive additional investments after a few years of production—so the original forecasts from the time of the final investment decision (FID) are no longer directly comparable to the production data after more than 5 years. The data set in this study consists of forecasts from the time of the FID for greenfields on the NCS between 1995 and 2020, production data and metadata about the oil fields. The data are further described in the next section.

Data

Forecasts and Production Data. Operators on the NCS must annually submit production forecasts to the NPD for the Norwegian revised national budget (RNB) from the time of the FID until field abandonment. Due to the fiercely competitive nature of the oil and gas industry, production forecasts are usually highly confidential. When RNB data are shown in public, the forecasts are usually aggregated to the national level[i]. It is the RNB forecasts which are closest in time to the FID for each field which are used here. Obviously, there may be instances where the RNB forecasts submitted to the NPD are not identical to the forecasts used to support the development decision. However, for 12 fields where we have been able to compare the internal plans for development and production with the RNB data, the forecasts have been almost identical. Thus, we will assume that the RNB forecasts are the same as the ones used for decision-making within the companies.

Fig. 2 shows example forecast series which are submitted to the NPD every year for each field: p_{10}, the expectation, and p_{90}. Here, p_x is the probability x that production P is below a certain value. The data set includes both annual and cumulative rates, which have different uncertainty ranges. We use only cumulative production forecasts and will refer to the three time series as F_{10}, F_M, and F_{90}.

In contrast with the forecasts, actual production data are publicly available from NPD databases (NPD 2022). Production is broken down into oil, gas, natural gas liquids, and condensate. Thus, it is straightforward to compare production data with the forecasts. We refer to Forecast Year 1 as FY1, Production yYar 1 as PY1, etc.

Field Metadata. What type of project characteristics can be related to organizational and individual bias? Research in cognitive psychology and behavioral economics shows that project management style, incentives, and politics usually play a central role (Flyvbjerg and Rasmussen 2021). However, such soft information is neither easy to tease out of reports nor to quantify. Reservoir geologists and engineers in the oil industry would perhaps emphasize the importance of expert knowledge, statistical methods and workflows—but this is also information which is not easily available.

In contrast, quite a lot of "hard" features, such as the historical oil price, the depositional environment, reservoir complexity, field size, reservoir drive mechanism, and the number of DST logs can be found for most fields. **Fig. 3** shows example geographical data that are publicly available on the NCS (NPD 2022). It is reasonable to hypothesize that many of these features may have a direct or indirect

[i]Field-level forecasts were generously made available by the NPD for the study, but are subject to a confidentiality agreement. Thus, all fields and operators are anonymized.

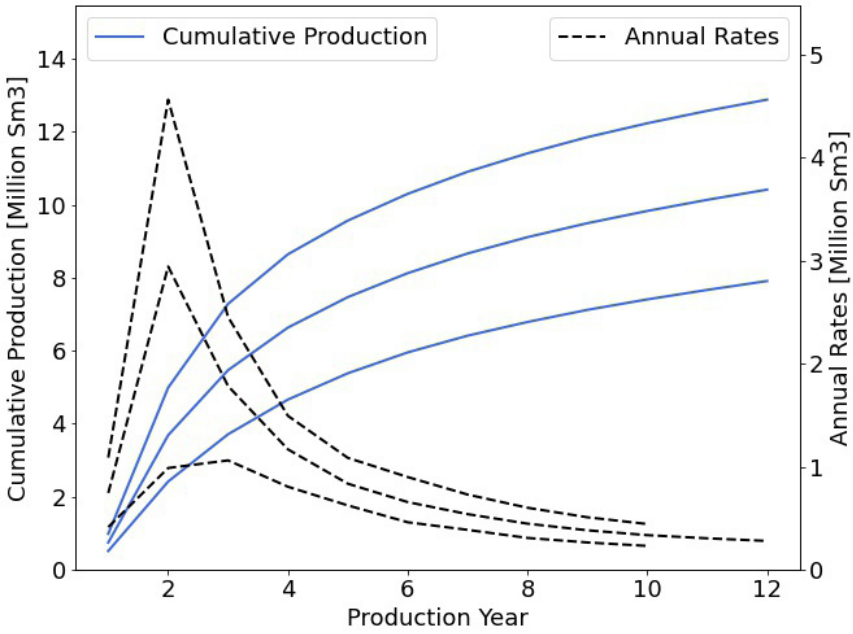

Fig. 2—Example forecasts of cumulative production and annual rates. Note that the cumulative sum of the annual forecasts are not (or should not be) equal to the cumulative forecasts because of auto-correlation over time. The low, central, and high curves represent the statistical p_{10}, the expectation, and the statistical p_{90}, respectively, against production time.

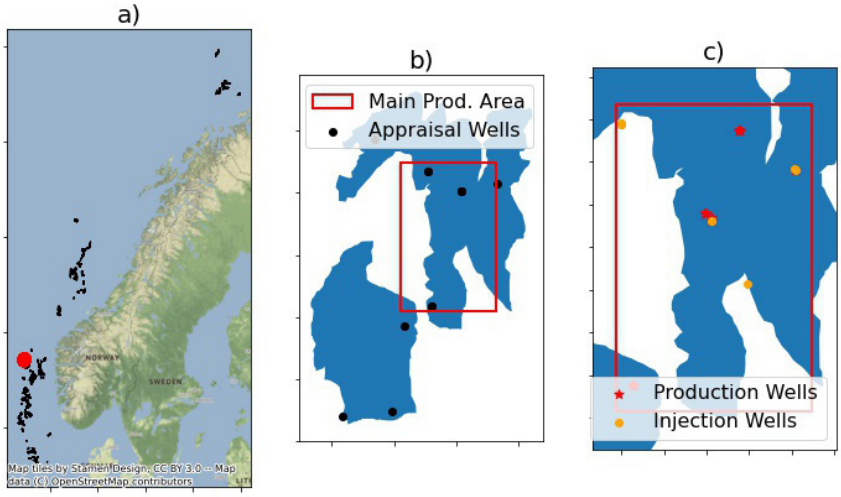

Fig. 3—Example data about the Vigdis field: (a) all NCS oil and gas fields (in black) and the field location (in red); (b) area of field license from the NPD (shaded blue), drilled appraisal wells and main area of the reservoir; and (c) current injection and production wells.

relationship with certain types of bias. For instance, it seems plausible that large fields have more experienced project managers than smaller satellite developments, which should entail "better" forecasts. It is also seems plausible that high oil prices may contribute to overly optimistic forecasts. A statistically significant relationship has been found between cyclical oil price changes and cost overruns in petroleum projects on the NCS (Dahl et al. 2017), but we know of no such analysis related to benefit shortfalls. Regarding nomenclature, field characteristics, parameters, and features will be used interchangeably in this text—all refer to possible predictors in a statistical model.

Table 1 shows field features x_1–x_{16} used as potential predictors of bias in this study. All the data are public and have been collected from public databases. Geophysical surveys are only available with spatial coordinates after 2009, so it was not possible to quantify the amount or quality of this information for this study. While much of the table should be self-explanatory, some of the parameters are explained in further detail here:

Feature Number	Description	Unit
x_1	Project sanction/FID year	year
x_2	Annual average of oil price 2 years before FID (inflation adjusted)	USD
x_3	Oil reserves by PY4	Sm^3
x_4	Mean forecast of ultimate recovered volume at time of FID	Sm^3
x_5	Average water depth to seafloor	m
x_6	Approximate depth from seafloor to main reservoir	m
x_7	Total investments (inflation adjusted) by PY4 per unit mean volumetric forecast at time of FID	NOK/Sm^3
x_8	Recovery method (1–3)	–
x_9	Reservoir quality minimum (1–5)	–
x_{10}	Reservoir quality maximum (1–5)	–
x_{11}	Deep marine deposition	True/false
x_{12}	Satellite field	True/false
x_{13}	High reservoir complexity	True/false
x_{14}	Number of development wells per unit reservoir area by PY4	$1/km^2$
x_{15}	Number of appraisal wells per unit reservoir area by PY4	$1/km^2$
x_{16}	Number of DST logs per unit mean volumetric forecast at time of FID	$1/Sm^3$

Table 1—Field features used for prediction.

- Feature 8 is the recovery method, either pressure depletion (score: 1), water injection (score: 2), or water and gas injection (score: 3). It is translated into a score between 1 (less expensive) and 3 (more expensive). For example, the Gimle field produces with one water injection well and three producers (score: 2).
- Reservoir quality (Features 9 and 10) is described for most fields as poor (score: 1), moderate (score: 2), good (score: 3), very good (score: 4), excellent (score: 5), or a range between these. Thus, we set an integer score between 1 and 5.
- The depositional environment of the reservoir rock is not always known. However, there may be sufficient information for reservoir geologists to draw some conclusions. For instance, deep marine deposits are of a different nature than shallow-marine deposits and require other geomodeling methods and different expertise. Most fields on the NCS are assumed to be one or the other, so Feature 11 is Boolean.
- If a field is developed with a tieback to a larger field, it is classified as a satellite field (Feature 12). As an example, the Gimle field, which had production start in 2006, was developed from the C facility on the Gullfaks field (production from 1986).
- Although geophysical data and detailed reservoir models are not available, a textual description of the reservoir is provided by experts at the NPD. When conditions such as heavy faulting; high-pressure, high-temperature wells; or high compartmentalization are present, Feature 13 (high reservoir complexity) is set to score True. Otherwise, this Boolean feature is set to False.
- Feature 14 is the number of completed injection and production wells per unit oil reserves by PY4. The oil reserves estimate can vary considerably over time and with additional investment.
- Feature 16 is the number of completed DST logs per unit oil reserves at the time of the FID (i.e., a proxy for information quality and quantity about production potential.

Square root transforms are applied to some of the parameters before use in prediction models (see Appendix B for further details).

Quantification of Current Forecast Quality and Value Erosion

Production shortfalls and value erosion of new oil and gas fields are often blamed on delays in production start. Overruns with respect to time and cost are so common in all industries that it has become known as the "iron law of megaprojects" (Flyvbjerg 2014). Empirical results show that this also occurs across all oil- and gasfield types ([e.g., both subsea projects (Jergeas and Ruwanpura 2010) and oil sand projects (Bergli and Falk 2017)]. In hindsight, there are always possible explanations for why delays happen, such as unexpected supply chain problems and poor project management. To compare production data with forecasts, it is of interest to distinguish between the impact of such logistical issues and handling of subsurface uncertainty by decision makers. Detailed schedules of well development are not available for the oil fields in this data set. Hence, it is not possible to quantify this effect exactly. Nevertheless, as a proxy, it is possible to separate the impact of delays in production start from later production shortfalls. The forecasts are only given by calendar years, so the estimated month of production start is not known. Thus, a delay in production start by 3 months from November to February is rounded up to a year in the results below. However, forecasted production start months are assumed by the authors to be uniformly distributed through the calendar year, so this strikes both ways (a delay of 20 months would be rounded down to 1 year if estimated production start was in February). **Fig. 4** shows example cumulative production curves with and without adjustment for delays. In the former case, PY1 is set equal to FY1. In the next subsections, both are used to quantify bias and value erosion.

The Forecasts Show Clear Presence of Bias. The mean attainment ratio $\alpha_y = \mathbb{E}_y[P/F_M]$ [i.e., actual cumulative production (P) relative to expected cumulative production (F_M)] is computed for each production year y. To remove the effect of optimism bias, the forecast triplet for field i in year y is shifted down by a term $\delta_{y,i}$ computed as:

$$\delta_{i,y} = F_{M,i}(1 - \alpha_y). \quad (1)$$

Fig. 5 shows an example for a single field of how forecasts adjusted for optimism bias compare to the original FID forecasts. After adjustment, the expected attainment ratio for all fields is now one in each year. **Fig. 6** shows aggregated statistics for all fields for PY2–PY10, with PY1 and FY1 aligned, with and without adjustment for optimism bias. Strikingly, after 2 calendar years, almost 60% of fields have

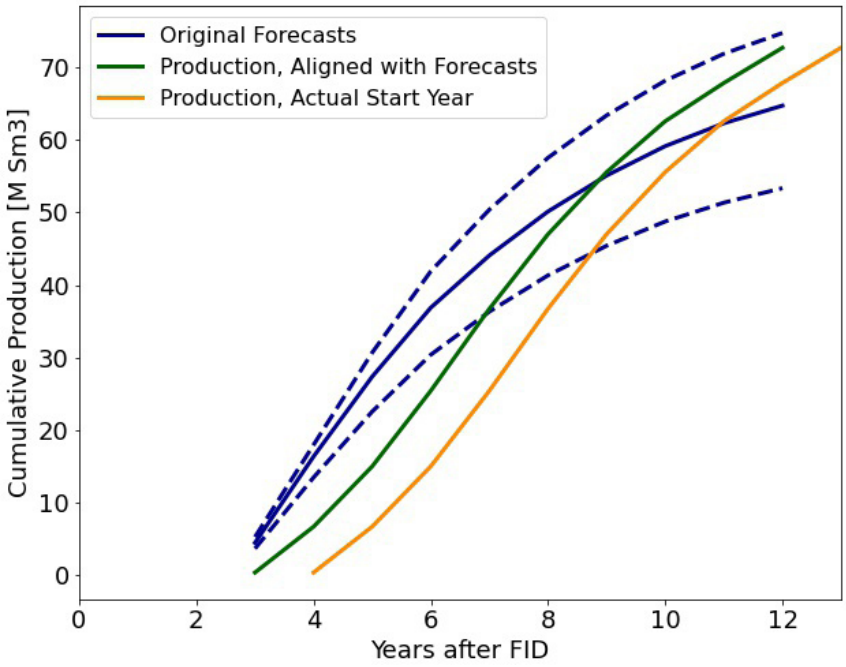

Fig. 4—Cumulative oil production relative to forecasts, with and without alignment of the production start year. The forecast curves are F_{10}, F_M, and F_{90}, from low to high.

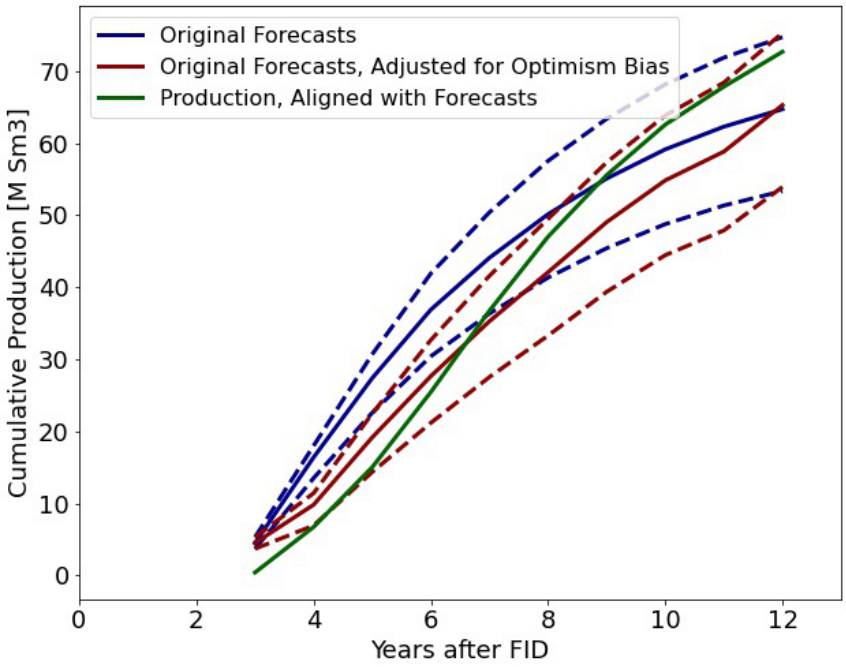

Fig. 5—Forecasts adjusted for optimism bias by Eq. 1. The forecast curves are F_{10}, F_M, and F_{90}, from low to high.

produced less than the F_{10} estimate and only about 25% of the outcomes are within the F_{10}–F_{90} range. These numbers are somewhat more favorable from the operators' point of view than found by Bratvold et al. (2020) (e.g., the fraction of outcomes below F_{10} after PY4 is closer to 40% than to 60%). Nonetheless, if the forecasts were unbiased, the fraction of outcomes below F_{10} and within F_{10}–F_{90} should be much closer to 10% and 80%, respectively. It is also noteworthy that fields which produce more than the F_{90} estimate are overrepresented. This means that a significant number of fields could probably produce more than they actually do; decisions about procurement of pipes and processing equipment are also based on production forecasts, and these may act as effective constraints on production (Bieker et al. 2007). Thus, overprecision in forecasts leads to value erosion on the high side as well.

Notwithstanding the early shortfalls, production attainment increases over time and is actually over 1 after 10 years of production. This differs somewhat from the results of Nandurdikar and Wallace (2011), who find that average production attainment stays almost constant at 80% from PY1 to PY4 (compared to "planned" production—it is not known whether the production forecasts they use are expected values, median values, or deterministic forecasts). In other words, we find that early production shortfalls have on average been eliminated

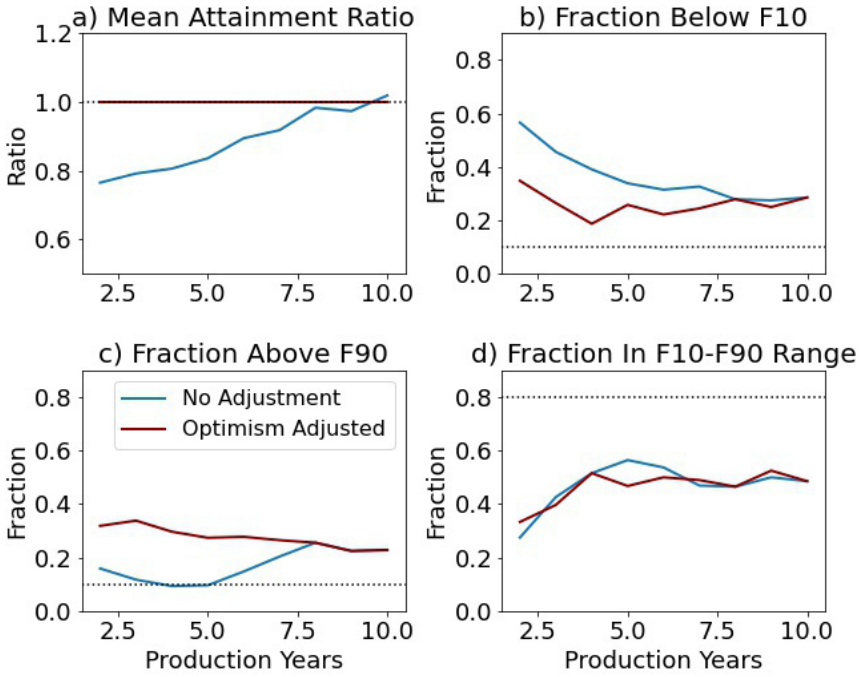

Fig. 6—Forecast performance statistics for all fields when the production start year is aligned with the first forecast year (PY1 = FY1), with and without adjustment for optimism bias. Subplot (a) shows the attainment ratio of production to the mean forecast, whereas figures (b) to (d) show the respective fractions of outcomes.

after 10 years of production. However, as argued above, projects with lifetimes over 10 years have usually received large additional investments and are not the same projects as the ones which were sanctioned. The discounted value streams also suffer greatly from the early shortfalls, which we show below.

In **Fig. 6**, mean production attainment is seen to be between 0.75 and 0.85 in PY2 to PY5. When optimism bias is adjusted for, the fraction of outcomes above F_{90} increases by approximately the same amount as the decrease of outcomes below F_{10}. This is also shown by the fraction of outcomes within $[F_{10}, F_{90}]$ being roughly the same with or without adjustment. Thus, the overprecision bias seems to be independent of optimism bias.

Fig. 7 shows the same data points as in **Fig. 6** but also includes delays in production start. Year 1 on the horizontal axis is now FY1 at project sanction. From **Fig. 7a**, average production attainment is seen to be at about 60% of the mean forecast after FY2. Furthermore, more than 70% of the fields have produced less than the F_{10} estimate after FY2 **(Fig. 7b)**. Forecasters are highly overconfident, shown by the much too narrow probability range—only 50% of the outcomes are within the F_{10}–F_{90} interval after FY4. Similarly to the time-aligned data above, the fraction of outcomes above F_{90} increases by the same amount as the reduction in outcomes below F_{10} when we adjust for optimism bias. Again, because of continuous additional investment after the FID, attention should mainly be paid to the first few years of production (FY2–FY5), so the picture is really bleaker than the trends in FY6 to FY10 suggest in **Fig. 7**.

Bratvold et al. (2020) and Nandurdikar and Wallace (2011) have previously documented a similar-looking picture but the data set is now twice as large as that in Bratvold et al.'s study, and Nandurdikar and Wallace's findings are only based on base-case forecasts. Thus, the presence of systematic bias in greenfield production forecasts is further corroborated by these results.

Production Shortfalls Entail Value Erosion. To further illustrate the negative impact of production shortfalls on valuation, it is of interest to compute the discounted revenue stream to the time of the investment decision. For valuation of new oil and gas projects, realistic uncertainty quantification is essential. However, such estimates encompass many different disciplines. The net present value of a potential investment is sensitive to factors such as the discount rate and the assumed future oil price. Greenfield investments may also be considered as a series of real options (Smith and McCardle 1999). The data set in this study does not include forecasts of capital expenditure at the time of the FID, and we do not know what assumptions were made about future oil prices in each case. Thus, instead of making a number of assumptions and attempting to compute the net present value for each field, we discount only the volume stream to the time of the FID. A discount rate of 8.5% is used. We consider cumulative production relative to the forecast at the end of PY4 (forecasts and production aligned) and FY4 (forecasts and production nonaligned).

Fig. 8 shows the ratio of discounted cumulative production to the discounted forecasts when forecasts and production are aligned. The average ratio with respect to the mean forecast after FY4 is then 76.3 % (i.e., an average shortfall in value of 23.7%). **Fig. 9** shows the same data as **Fig. 8**, but includes time delays. The average ratio is then 69.4% (i.e., with constant oil prices, an average shortfall in value of 30.6%). Thus, the relative contributions of shortfalls due to delays in production start and shortfalls after production start are about $(30.6 - 23.7)/30.6 \approx 23$ % and $23.7/30.6 \approx 77$ %, respectively **(Fig. 10).** Low production attainment after production start is not just due to poor subsurface understanding. Possible causes include low production efficiency, deviations from well drilling schedules, poor understanding of the reservoir, optimistic forecasts, and topside issues. These data clearly indicate that optimism bias with respect to production rates after production start have a larger effect than delays in production start on value erosion. In the remainder of the article, it is the nominal (nondiscounted) values which are used.

More Advanced Forecast Quality Metrics. The literature on forecasting methods and quality metrics is extensive (Gneiting et al. 2007; Hyndman and Athanasopoulos 2018), so it is sensible to apply some of the tried and tested methods from other research areas to the

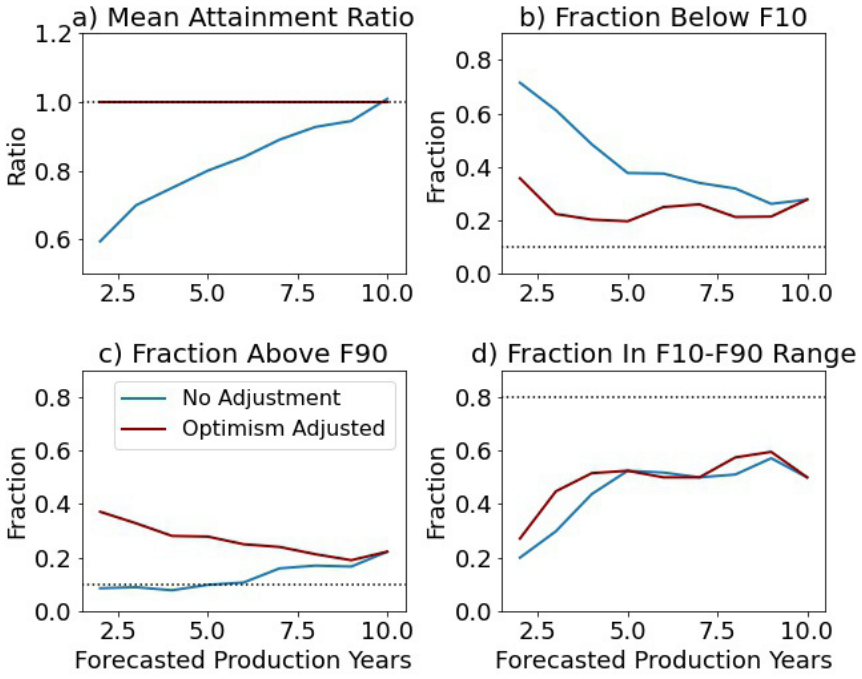

Fig. 7—Forecast performance statistics for all fields when delays in production start are included, with and without adjustment for optimism bias. The production years on the *x*-axes are the FYs at the time of the FID.

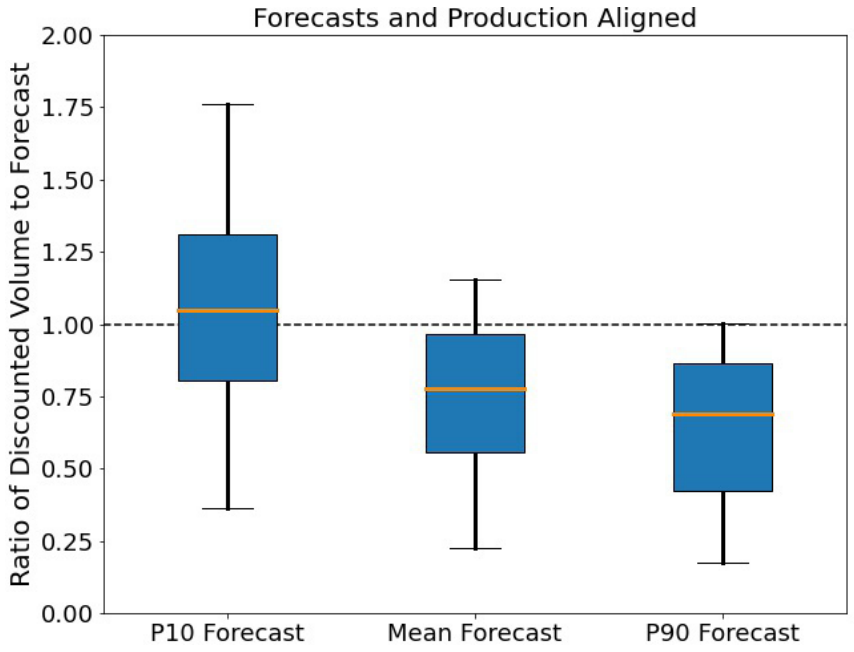

Fig. 8—Ratio of the discounted volume stream for actual production relative to the forecasts until FY4, when forecasts and production are aligned by start year. The box shows the p_{25}–p_{75} range and the median in orange, and the whiskers show the p_{05}–p_{95} range.

production forecasts. *Density forecasts* are common in finance (Amisano and Giacomini 2007) and meteorology (Murphy and Winkler 1992) but less so in other fields. Using a full probability density allows for more advanced analyses than just using the forecast mean or two percentiles. In oil and gas, stochastic methods for reservoir evaluation have been used in industry at least since the 1980s (Haldorsen and Damsleth 1990). However, we do not know the specifics of the workflows that have been used by different operators to generate the forecasts in this data set. Thus, it is necessary to make some assumptions about the underlying distribution.

Fitting a Continuous Probability Distribution. As the forecasts in this data set only include a mean and two percentiles, the amount of information in the forecasts is limited. By making some reasonable assumptions about the underlying statistical distribution, it is possible to fit a probability density function. Clearly, the lower limit is zero, and there is an upper limit on production capacity due to equipment and materials on platforms and in subsea installations. Hence, the forecast probability density functions are assumed by the authors to be unimodal and not very long-tailed. In addition, F_M is usually closer to F_{90} than to F_{10}, so the distribution needs to allow for skewness.

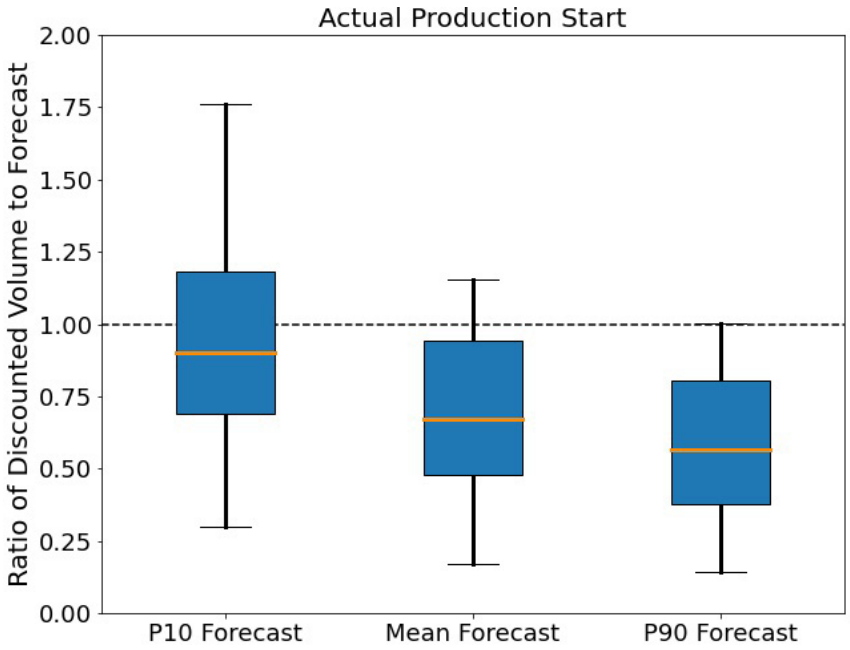

Fig. 9—Ratio of the discounted volume stream for actual production relative to the forecasts until FY4, with actual production start. The box shows the p_{25}– p_{75} range and the median in orange, and the whiskers show the p_{05}– p_{95} range.

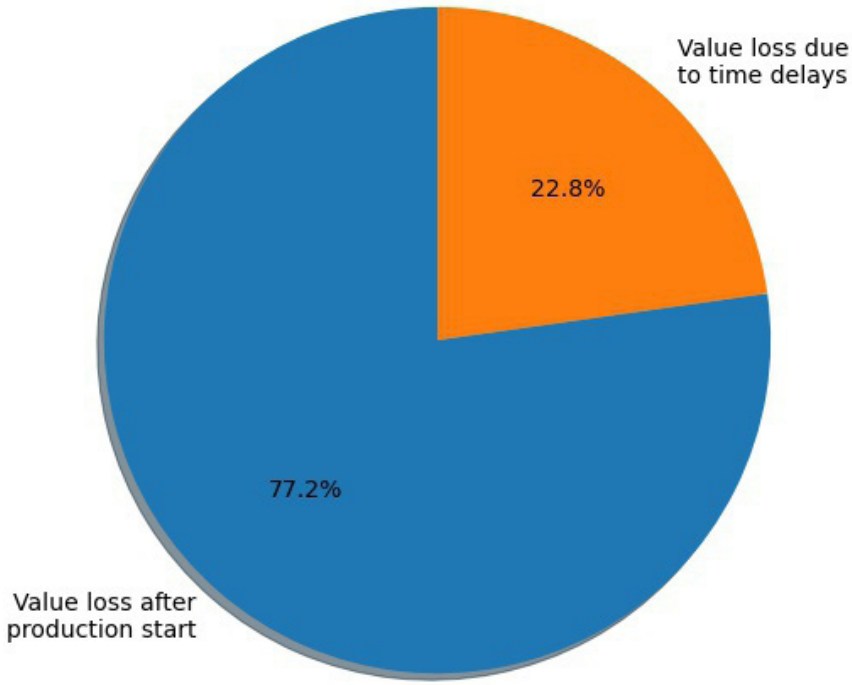

Fig. 10—Relative importance of effects after production start and delays in production start on value loss with respect to the mean forecast.

There are several theoretical distributions which are relevant, but one needs to fit the function to the percentiles and to the mean. Here, the so-called symmetric percentile triplet metalog equations (Keelin 2016) are used to fit a cumulative density function (CDF). Two advantages of using a metalog distribution are (1) that percentiles can be used directly as input to the fitting procedure and (2) that boundedness below is easy to implement. One must however find F_{50} by optimization, so that the mean of the distribution is as close as possible to the forecast mean F_M. The resulting probability density function is unimodal and has relatively short tails. Given a $\{F_{10}, F_{50}, F_{90}\}$ triplet, the semibounded symmetric percentile triplet metalog density function is unique and can be computed directly from a set of analytical equations.

Production Attainment Percentiles from the CDF. After fitting the CDFs, it is straightforward to compute the production attainment percentiles, instead of just the attainment ratio with respect to the mean forecast, for each field for each year. **Fig. 11** shows a subset of these results for nonaligned data after FY4, grouped into quartiles. The diagonal plots show the marginal distributions of the quartiles

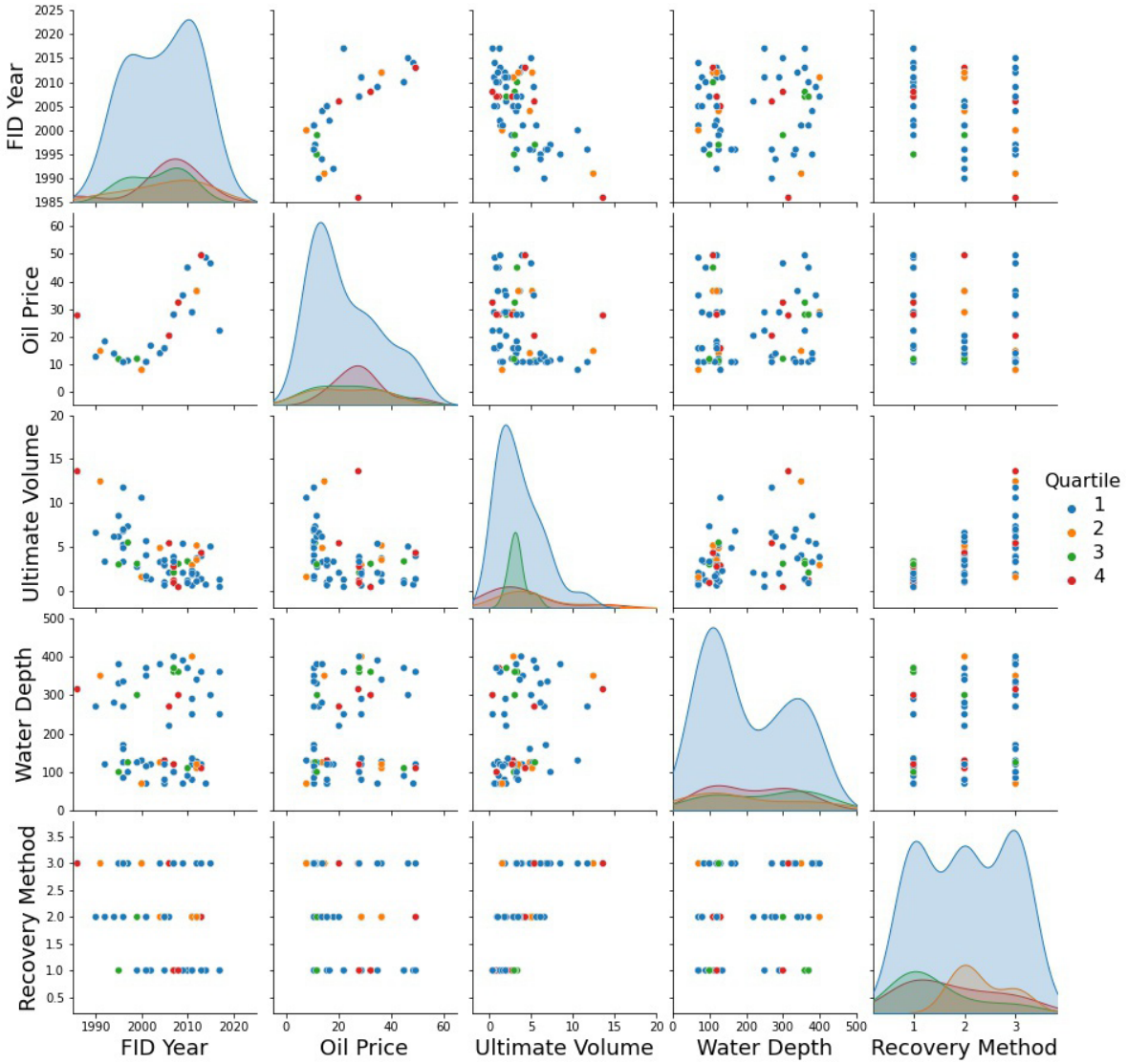

Fig. 11—Production attainment quartiles after FY4 against FID year (x_1), oil price 2 years before the FID (x_2), ultimate forecasted volume (x_4), water depth (x_5), and recovery method (x_8). The area under the curve is proportional to the number of outcomes within the respective quartile.

(Q1–Q4) against four of the parameters in **Table 1**, and the area under the curve is proportional to the number of outcomes within each quartile. The production outcomes are not evenly distributed: 68% are in the first quartile (i.e., below the 25th percentile). If there is a predictive relationship between any of the features and production attainment, the four quartile densities should have different shapes and means. However, it is difficult to discern any clear patterns. For example, the mode is located approximately at the same field volume for all four densities, and the densities vs. recovery method look almost flat.

Off the diagonal, **Fig. 11** shows the pairwise relations between features. There is obviously not independence between all these feature pairs. For instance, oil price and FID year are tightly coupled as oil price changes gradually over time. Recovery method is also seen to be related to field volume—gas injection (x_8: 3) is used at all fields above a certain size, whereas fields which produce with only pressure depletion (x_8: 1) are smaller on average than the rest of the data set.

Forecast Performance Score. The continuous ranked probability score (CRPS) is a commonly used metric for density forecasts and is a so-called strictly proper scoring rule[ii] (Gneiting et al. 2007). It penalizes both the width of the forecast and the distance to the outcome, so a narrow forecast around the actual outcome is optimal. The minimum (and best) score is $S(F, y) = 0$, which is when the density forecast $F(x)$ has zero uncertainty and is a step function at the outcome value y. However, a problem with the CRPS integral (see Appendix B) in the context of the oilfield data is that it yields higher scores for large fields than for small fields, as the absolute numbers vary significantly. A scale-independent version of this scoring rule is proposed instead. We refer to the appendix for details. **Fig. 12** shows an example forecast CDF and the resulting forecast score as a function of the outcome. The score is symmetric about the mean and the minimum coincides with the forecast mean.

[ii] A scoring rule is strictly proper if it is uniquely optimized (in expectation) by only the true probabilities.

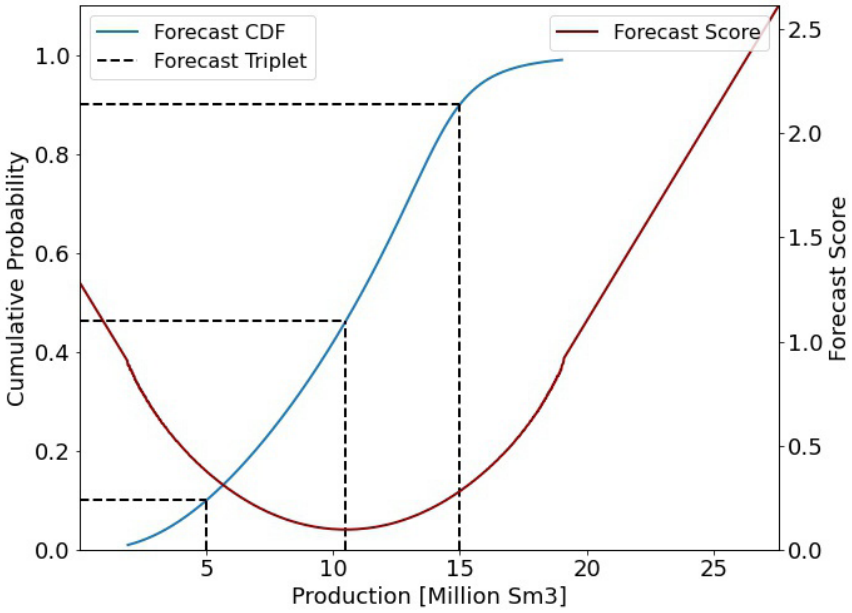

Fig. 12—Example forecast triplet, metalog density $F(x)$, and score $S(F, y)$ as a function of the production outcome y on the horizontal axis. The example forecast triplet is $\{F_{10}, F_M, F_{90}\} = \{5, 10.5, 15\}$ and the minimum score is 0.098 for $y = 10.5$.

Do Field Features Predict Bias?

The key question which we address in this study is whether field features can be used as predictors of forecasting bias. As the number of features in **Table 1** is relatively high compared to the size of the data set (71 in FY2 and 66 in FY4), it is important to avoid statistical overfitting (Hastie et al. 2009): The risk of finding spurious correlations is proportional to the number of statistical relationships and methods which are tested. Fortunately, statistical learning is possible even when the number of features is much higher than the number of observations (Hastie et al. 2009), but it is essential to use cross-validation and regularization. Furthermore, only a few predefined hypotheses are tested, and only methods which are not susceptible to overfitting are used. As a first step, we consider univariate tests of independence between the predictors and production attainment.

Univariate Tests of Independence. Before considering more advanced supervised learning methods, it is also of interest to test the features and the attainment ratios for statistical independence. **Fig. 13** shows the attainment ratio vs. the 16 features, as well as the numerical scores of two tests:

1. *p*-values from an *F*-test (Box 1953) of a hypothetical linear relationship between the attainment ratio after FY4 and the field features. The null hypothesis is that an intercept-only model between each predictor and the response variable is as good as a linear model. Thus, low *p*-values indicate that the linear model has explanatory power.
2. The mutual information (Kraskov et al. 2004) is a measure of overall dependency, not necessarily linear, between attainment ratio and features. Values closer to 1 than to 0 indicate a strong relationship.

The variables with the lowest *F*-test *p*-values and the highest mutual information scores are highlighted in red. Minimum and maximum reservoir quality are the only predictors which have *p*-values close to 0.01. However, there are only two fields which have a maximum reservoir quality score outside the central range (2–4), and the majority (73%) have a score of 3 ("good"). The two fields which have "excellent" reservoir properties have slightly exceeded production expectations. Likewise, almost all fields have a minimum reservoir quality score in the 1–3 range. A drawback with reservoir quality description as a predictor is that it is continuously updated and is unlikely to be completely representative of the beliefs at the time of the FID.

The two features with the highest mutual information scores are x_7, investments per unit reserves, and x_8, the recovery method. If the test of a linear relationship gives an insignificant *p*-value but the mutual information score is high, it usually means that there is a pattern which the linear model does not capture. But here, the scores are not very high (between 0.1 and 0.2), and a visual inspection of the plots (e.g., the dip in production attainment against x_7), shows that the patterns are unlikely to represent something meaningful.

Machine Learning Methods for Classification and Regression. It is essential that what a prediction model learns about a training data set generalizes to other data points. We use the following classes of supervised learning methods, which are all known to be robust against overfitting:

- Random forest (RF) regression and classification is a tree-based ensemble method (Hastie et al. 2009). Each tree is built by bootstrap sampling from the training set and by adding randomness to the splits of the decision nodes within the tree. Both contribute to reduction of variance. While individual trees are typically subject to high variance and overfitting, this effect cancels out when the average of a forest of trees is used.
- Partial least squares (PLS) regression (Geladi and Kowalski 1986), such as principal component regression, is used to compute linear combinations of the predictor variables. However, instead of finding the direction of maximum variance in the predictors, PLS aims to maximize the variance explained in the response variable. It is a regression method which is particularly suited when there are many predictor variables with presence of multicollinearity. Use of PLS regression is common in fields such as bioinformatics and chemometrics, where this is often the case. Much like Lasso and ridge regression (Hastie et al. 2009), overfitting is controlled by the number of components that are used.

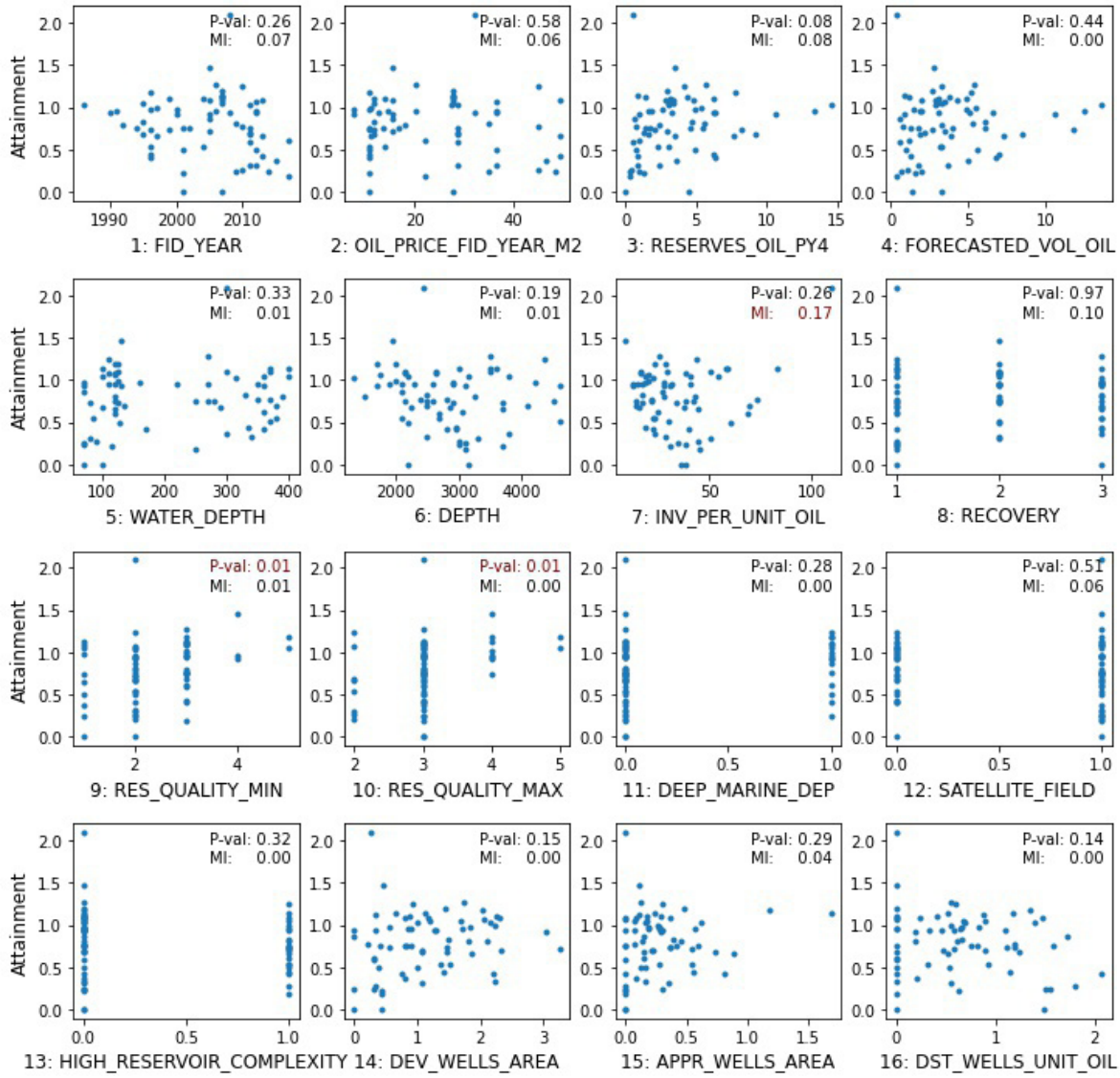

Fig. 13—Univariate tests.

- Support vector machines (Cortes and Vapnik 1995) were originally invented for the purpose of binary classification. The underlying idea is to construct a hyperplane which optimally separates labeled data points. By using a so-called *kernel function* (Schölkopf 2000), the data can also be mapped into another lower-dimensional space. The advantage is that support vector classification, which is a linear method, then yields a computationally inexpensive and nonlinear decision boundary. We use a so-called radial basis function kernel (Musavi et al. 1992), where the regularization parameter (C) and the kernel coefficient (γ) are optimized. Thus, it is a type of prediction model which accounts for the most important nonlinear effects. A disadvantage of the kernel support vector machine is that it is in practice a black-box model—the feature importance weights are not available. Support vector regression (SVR) is based on the same ideas as the support vector classification. Compared with ordinary linear regression, the advantage of SVR is that outliers are given no weight—only data points within a certain margin from the optimal hyperplane are used in the model.
- Naive Bayes classification is based on the "naive" assumption of independence between features (Hastie et al. 2009), so that the class probability is proportional to the product of 16 conditional probabilities and the class prior. Despite the model simplicity, it can achieve high accuracy rates. Here, we use a Gaussian naive Bayes classification.

The methods above have several parameters which must be set, such as the number of trees in an RF model and the strength of the regularization term in an support vector classification. For some of the parameters, the default values in the Python Scikit-learn library are used while others are optimized by cross-validation (Hastie et al. 2009). **Fig. 14** shows the contour plot of the coefficient of determination for RF regression vs. the number of trees and the maximum number of features. In this case, the optimal number of features is one, and the score is negative for all combinations of these parameter values. In other words, the regression model has no explanatory power.

For each classifier or regression model, a pipeline with preprocessing and feature selection is created. Tree-based models do not require standardization of the data because each tree is a hierarchy of decision rules. However, standardization is necessary for support vector machines and PLS. Before fitting a prediction model to each fold of training data, half of the predictors are removed by feature selection (Chandrashekar and Sahin 2014). Instead of univariate testing of correlations, this is done by fitting a simpler machine learning model in a first step, which is then tested for feature importance.

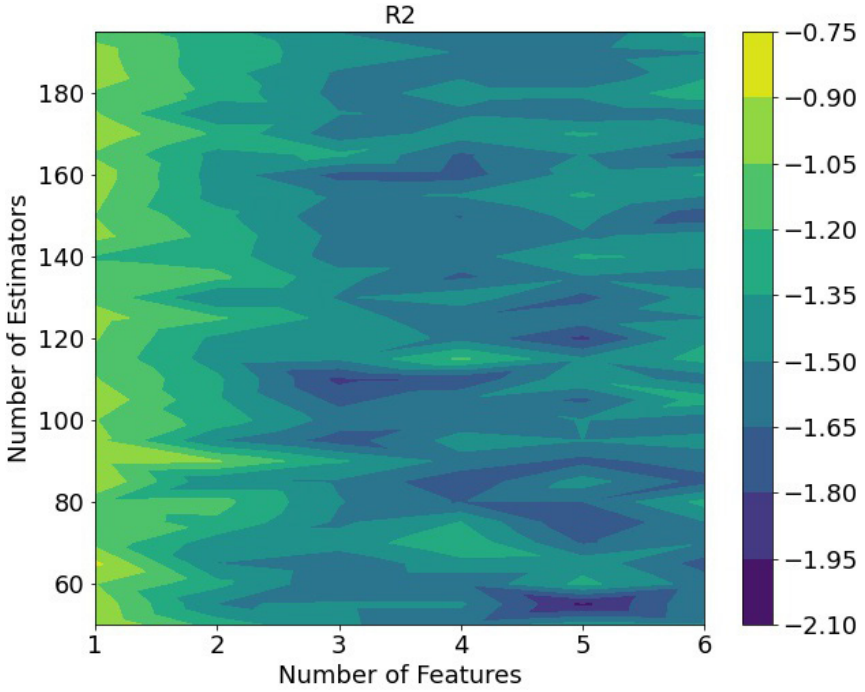

Fig. 14—Optimal parameter search for RF regression of the forecast quality score in Eq. B-2 using all parameters in Table 1 as predictors. The contour plot shows the R^2 score (higher scores are better). The optimal choice is one feature, and the score is not sensitive to the number of estimators (number of trees).

Two standard metrics are used to measure the performance of the supervised learning models. For classification, we resort to the Brier score:

$$B(\mathbf{y}, \hat{\mathbf{p}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2, \tag{2}$$

where n is the number of observations, y_i is a binary outcome variable, and p_i is the estimated probability that $y_i = 1$. For example, if $p_i = 0.9$ and $y_i = 1$, then $B = 0.01$, whereas if $p_i = 0.1$ and $y_i = 1$, then $B = 0.81$. Thus, the better the forecast calibration is, the lower the Brier score will be. For regression, the coefficient of determination is used:

$$R^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}, \tag{3}$$

where $\hat{y}_i$ is the prediction, y_i is the continuous outcome variable, the denominator is the residual sum of squares, and the numerator is the total sum of squares around the mean $\bar{y}$. Thus, $R^2 = 1$ implies that the model explains all the variance in the response variable. On the other hand, the score goes negative if the predictions $\hat{\mathbf{y}}$ perform worse than always predicting the mean. In the next sections, all reported scores are the averages over the test sets with 10-fold cross-validation.

Classification of Fields with Low Production Attainment. For exploration and production companies, project risk assessment is of utmost importance. Most of the investments in this data set are also very capital and resource intensive. However, as seen for the forecasts in **Fig. 6**, more than 40% of the fields on the NCS have cumulative production levels below the F_{10} forecast in PY4. Being able to identify combinations of project risk factors would be highly advantageous to avoid future value erosion. Thus, a first question is whether the features in **Table 1** can be used to classify fields with subsurface properties where forecasts are typically much too optimistic. **Fig. 15** shows the classification performance of the RF classification, support vector classification, and naive Bayes classification when forecasts are aligned with production start. In all three cases, predicting the mean of the training sample outperforms these classifiers. Hence, there seems to be no predictive power in the field features with respect to this response variable.

Classification of Low Forecast Quality Scores. To expand on the question above, one may ask whether it is possible to relate the field features to poor forecasts in general (i.e., forecasts which have either low precision or which are too optimistic or too pessimistic). All these outcomes result in a degradation of the forecast quality score (Appendix B). In contrast with the section above, delays in production start are now included. With a threshold of $S_t = 0.5$, 27% of the data points have higher (worse) scores. As seen in **Fig. 16,** also in this case, all three classifiers have higher Brier scores than the zero-skill rate.

While these findings may be disappointing to investment analysts in oil and gas, it is consistent with long-running research on cost overruns in megaprojects. According to Flyvbjerg et al. (2018), the root causes of cost overruns are neither complexity, business cycles, nor geology but always psychological, motivational, and political biases. With this in mind, it should not be surprising that underestimates of upside and downside risk are not statistically related to any of the features in **Table 1**.

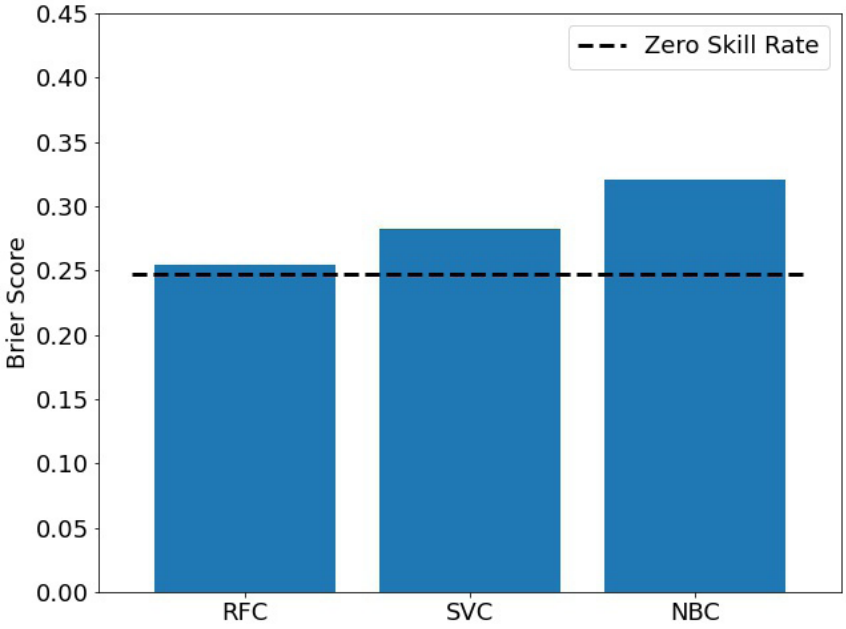

Fig. 15—Performance of classifiers of fields with cumulative production below the F_{10} forecast after PY4. Lower scores are better, and the zero-skill rate is the average Brier score achieved when one simply predicts the mean of the training data.

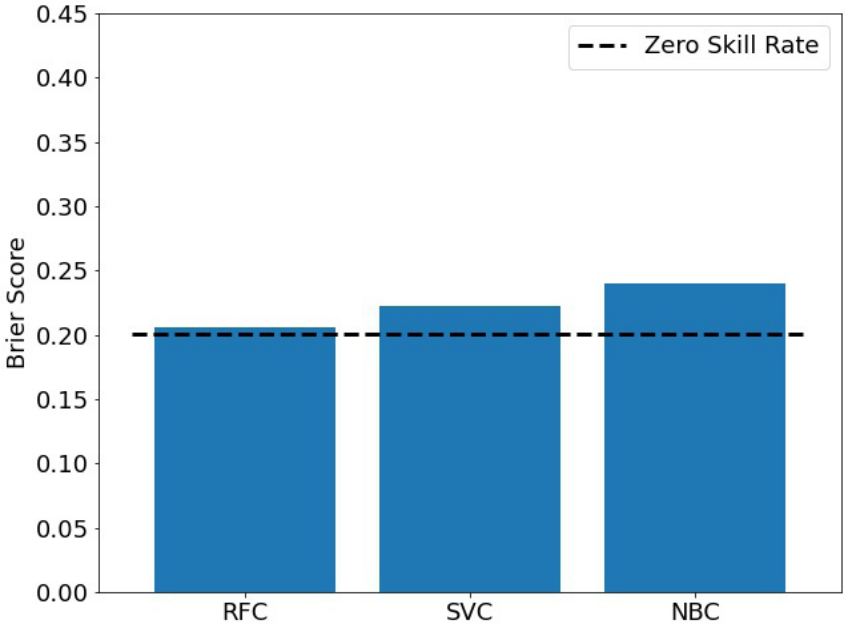

Fig. 16—Performance of classifiers of fields with a forecast score (Eq. B-2) above 0.5 after FY4. Lower scores are better, and the zero-skill rate is the average Brier score achieved when one simply predicts the mean of the training data.

Regression Analysis of Optimism Bias. As a third possibility, we consider three linear and nonlinear regression models (RF, PLS, and SVR) between production attainment (including delays in production start) and the field features after FY4. SVR is the best-performing model: **Fig. 17** shows the sorted response values, which range from 0 to 2.1. A successful model should capture at least some of the slope. Instead, the predictions $\hat{\mathbf{y}}$ behave more or less like noise around the green line, which is the zero-skill mean response $\bar{y}$. **Fig. 18** confirms this impression: The coefficient of determination is negative for all three models, so none of the regression models explain any of the variability in the response variable. Thus, the mean of the training data is a better guess for an out-of-sample data point than any of the model predictions.

It is natural to compare these results with the conclusions of Nandurdikar and Wallace (2011). They claim that appraisal strategy, broadly grouped into three categories (conservative, moderate, and aggressive), and API gravity are two project parameters which are strongly related to the attainment ratio. However, these conclusions are based on linear regression with no cross-validation, and no individual data points are shown. The *p*-values they report for the linear models for these features (0.019 and 0.042) are not "very statistically significant," as claimed by the authors. Furthermore, they present data on self-reported "root causes" by the project teams as explanations for what has gone wrong in each case, which are grouped into reservoir, wells, facilities, and other issues. Their data are clearly

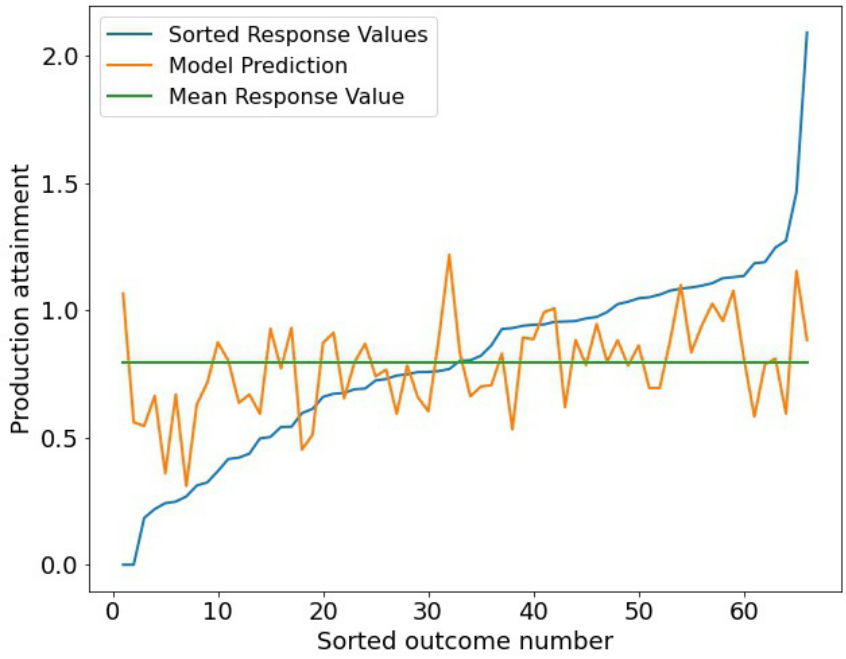

Fig. 17—Sorted production attainment outcomes after PY4, model predictions, and the response mean.

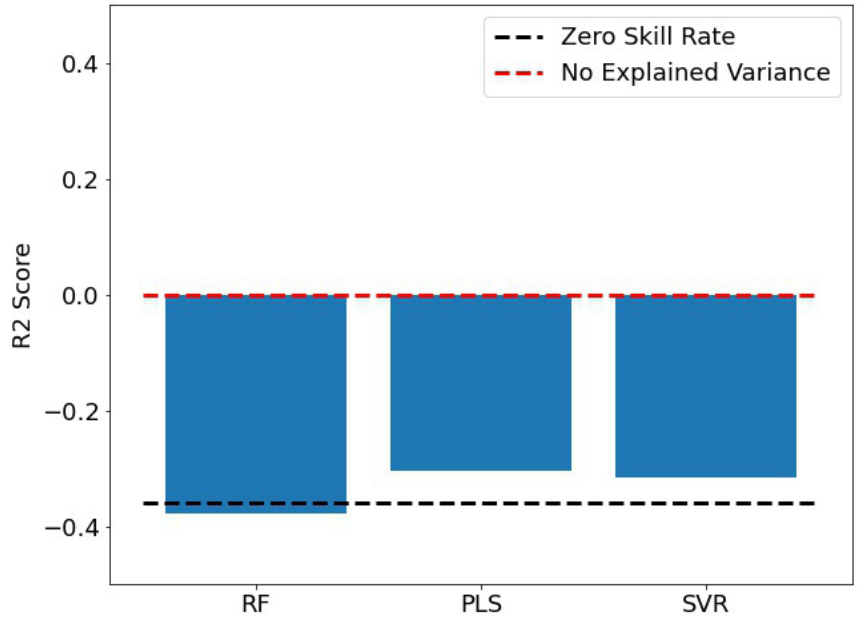

Fig. 18—Performance of the three regression models for prediction of attainment ratios. Positive R^2 scores imply that the model explains some of the variance, negative scores mean that performance is worse than an intercept-only model of the average of the entire data set. The zero-skill rate is obtained by predicting the average of the training data for each cross-validation fold.

interesting, but it is an open question whether they are more reliable than the forecasts themselves, as other types of bias, such as political bias and hindsight bias (Flyvbjerg 2021), may be at play. The data set in this study does not include the same type of soft information, but we find no evidence that there is a causal relationship between hard field characteristics and optimism bias in forecasts.

Zero-Skill Reference Forecasts Outperform the Operator Forecasts. Exploration and production companies spend enormous resources on uncertainty quantification models and forecasting, so one might expect the quality to be proportional to the effort. In the self-assessments by project teams in the data set of Nandurdikar and Wallace (2011), the overall production forecast quality was rated as "good" or "excellent" in 75% of the cases. Only 23% were rated as "fair" or "poor." A striking result in the same study is that production attainment looks like it is independent of these assessments: The distributions of outcomes relative to the forecasts look exactly similar for the two categories. In light of the additional findings in this study, it is therefore tempting to ask whether less expensive and simpler forecasts are viable. There is now at least half a century of solid research which shows that conventional methods of forecasting and estimation, both on an individual level and on an organizational level, are subject to various types of bias (Flyvbjerg et al. 2018). The so-

called reference class forecasting is an intuitively appealing and well-established approach used in many other fields (Kahneman 2011). This means that forecasts are based on past results instead of just project-specific information.

Reference Rates Based on Past Data. What could be a feasible reference forecast for a greenfield in oil and gas? Admittedly, it is a different case than, for example, weather forecasting, where forecasts are issued on a daily basis, so the number of empirical data points is very high, which again makes statistical learning and bias correction more straightforward. For instance, a reference temperature forecast could be the smooth seasonal mean for a set of geographical coordinates.

For oil production forecasts for a specific field, an obvious possibility is to compute the reference curve based on empirical production data from analog fields where production has ceased. The approach we propose here is described in further detail by Nesvold and Bratvold (2022), where the number of abandoned fields on the NCS is $n_e = 19$. First, the empirical production rates are normalized as follows: The time interval is first set to [0, 1] and the area under the curve is normalized so it integrates to one. A set of weights $w_1, ..w_{n_e}$, such that $\sum_{i=1}^{n_e} w_i = 1$, is assigned to the empirical curves based on similarity in size with an individual new field (in this case, Gaussian weighting). The output is then a reference shape for the production rate for the new field **(Fig. 19).** The mean forecast F_M is then found by scaling the cumulative version of this curve to match the estimated ultimate production volume and the expected number of production years (based on linear regression of past field lifetimes against field volumes) at the time of the FID. For the F_{10} and F_{90} forecasts, we multiply this reference curve by 0.35 and 1.6, respectively. These factors are rough estimates based on past data.

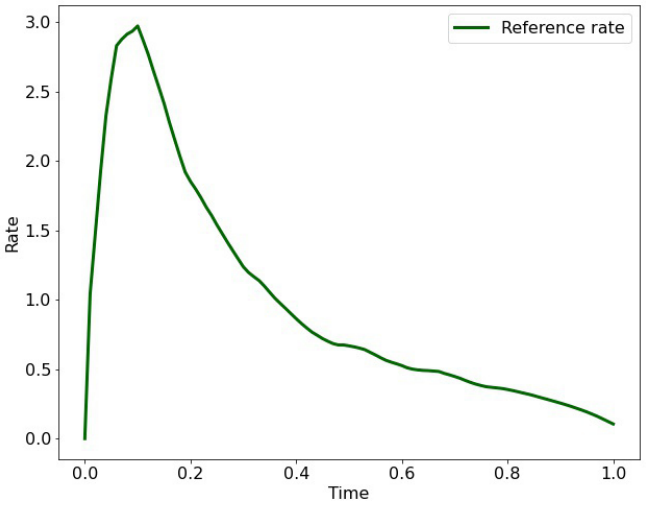

Fig. 19—Example reference production rate.

Performance. **Fig. 20** shows the performance statistics of this simple model compared to the original forecast data. It is clear that both optimism bias and overprecision bias have been reduced significantly. The attainment ratio is seen to be slightly above 1—this may be due to technological and organizational improvement. After all, the reference forecast is based on production rates for fields with production start in the 1970s and 1980s. Importantly, the uncertainty range in production is seen to be much more realistic. If the original forecasts also had been unbiased, it is likely that different decisions would have been made—such as whether to develop a field and what type of development concept to use.

Discussion

Good decision-making is in itself not a prerequisite for successful oilfield development, as there are many factors of uncertainty which can contribute both positively and negatively to profitability. For example, fluctuating oil and gas prices can cause delays in production start to have a positive impact on payback time on investments. Early cost overruns may be seen as insignificant when the field lifetime is extended for decades. However, as this study shows, in the long run, straightforward statistics reveal the degree of optimism and overconfidence which is present in investment decisions for these megaprojects. Only the production shortfalls (not cost overruns) are considered here, but it is clear that such forecast bias contributes to significant value erosion. The volumetric attainment ratio is on average only 60% of the expectation after FY2, and 70% of the outcomes are below F_{10} estimate **(Fig. 7)**. Furthermore, overprecision bias, measured by the fraction of outcomes outside the F_{10}–F_{90} range after adjustment for optimism bias **(Fig. 6)**, is independent of the latter. Overprecision bias also contributes to significant value erosion, both because downside risk is not accounted for and because production equipment may not be dimensioned for potential upside. As seen in **Fig. 10**, almost 80% of the value reduction in the first 4 years of the field lifetime is due to biased forecasts of factors which play out after production start. But this is only value reduction due to optimism bias. The actual loss is certainly higher. Getting the probabilities right is an essential component of high-quality decision-making (Howard 1968). Clearly, if the forecasts had been unbiased at the time of the FID, other investment decisions would have been made. According to Taleb (2007), bad forecasts are comparable to criminal offenses because of the damage they do to society.

Production shortfalls and optimistic forecasts are often explained away in anecdotal fashion. High oil prices, reservoir complexity, and poor field development strategies are examples of explanations which are given for optimism and overconfidence bias. While Nandurdikar and Wallace (2011) attempt to make connections between field features and poor forecasts through self-reported causes by the project teams, such insights are subject to significant hindsight bias (Flyvbjerg 2021), rendering them largely irrelevant.

The NPD has done an invaluable job in collecting numerical and qualitative data on hydrocarbon exploration and production on the NCS over several decades, which allows for transparency in a global industry which is characterized by secrecy and competitiveness. The data do not include soft information such as the level of competency, incentives for realism in forecasts, and choice of modeling techniques. But soft data are likely to be related to hard features, such as field size, project sanction year, and reservoir type: Large projects often have more experienced management teams; uncertainty modeling has evolved enormously over the past four decades; and different types of geology require different workflows and expertise. However, the analyses in this study show that the features have practically no

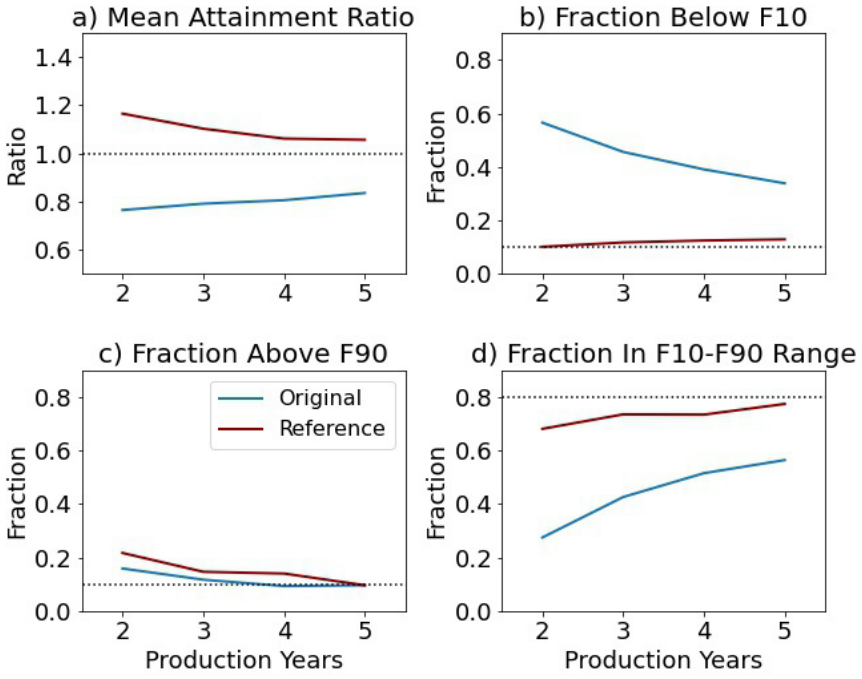

Fig. 20—Reference forecast performance compared with the original operator forecasts.

predictive power of forecast quality, whether they are used individually or simultaneously and whether a linear or an advanced, nonlinear model is used for prediction. The only possible exception seems to be reservoir quality, which is a qualitative description updated by the NPD over time, along with production history. Thus, it seems unlikely that projects with risks of bias or poor forecasts can be identified beforehand based on field features.

Why are forecasts biased? Flyvbjerg et al. (2018) acknowledges that scope changes and project complexity can cause cost increases and benefit decreases, but they are not the actual *root causes* of overruns and shortfalls with respect to estimates. The causal chain starts with planners who systematically underestimate known and unknown uncertainties and overestimate expected performance. We also agree with Nandurdikar and Wallace (2011) about the importance of having the right incentives in place: If nobody is accountable for the quality of the forecasts and if no time is devoted to learning from past mistakes, it is unlikely that performance will improve. For calibration, both Flyvbjerg et al. (2018) and Kahneman et al. (2011) commend reference class forecasting as an inexpensive and robust approach to reduce strategic misrepresentation, optimism bias, the planning fallacy, and other classic organizational biases. As described in the previous section, past production data for offshore fields on the NCS yields a reference shape for the production curve. If this curve is scaled to match the expected production volume and multiplied by empirically estimated scalar factors for the low and high estimates, the result is that the mean estimate after PY4 is adjusted down, F_{90} is adjusted up, and F_{10} is adjusted down. As seen in **Fig. 20**, the results are dramatically better than with the original forecasts. It is certainly food for thought that the enormous amounts that are poured into uncertainty quantification workflows do not lead to output which outperforms a zero-skill, free-of-charge reference forecast. We certainly do not suggest that probabilistic forecasting is futile work; on the contrary, it is essential for rational decision-making about investments. But current approaches in the hydrocarbon industry are not value-adding. It seems like a change of perspective from the inside to the outside view (Kahneman 2011) is necessary to address the bias problem.

Acknowledgments

We are very grateful to the following two organizations:

- Equinor ASA for the interest, time, and the financial support for the research project on forecasting methods.
- The NPD for access to the forecast data for research purposes and for structuring other high-quality public data about the NCS.

References

Amisano, G. and Giacomini, R. 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *J Bus Econ Stat* **25** (2): 177–190. https://doi.org/10.1198/073500106000000332.

Bergli, S. and Falk, K. 2017. Cause and Impact Analysis of Cost and Schedule Overruns in Subsea Oil and Gas Projects–a Supplier's Perspective. In *INCOSE International Symposium*, Vol. 27, 309–321. Hoboken, New Jersey, USA: Wiley Online Library.

Bieker, H. P., Slupphaug, O., and Johansen, T. A. 2007. Real-Time Production Optimization of Oil and Gas Production Systems: A Technology Survey. *SPE Prod & Oper* **22** (4): 382–391. SPE-99446-PA. https://doi.org/10.2118/99446-PA.

Box, G. E. P. 1953. Non-Normality and Tests on Variances. *Biometrika* **40** (3–4): 318–335. https://doi.org/10.1093/biomet/40.3-4.318.

Bratvold, R. B., Mohus, E., Petutschnig, D. et al. 2020. Production Forecasting: Optimistic and Overconfident—Over and Over Again. *SPE Res Eval & Eng* **23** (3): 799–810. SPE-195914-PA. https://doi.org/10.2118/195914-PA.

Chandrashekar, G. and Sahin, F. 2014. A Survey on Feature Selection Methods. *Comput Electr Eng* **40** (1): 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Cortes, C. and Vapnik, V. 1995. Support-Vector Networks. *Mach Learn* **20** (3): 273–297. https://doi.org/10.1007/BF00994018.

Dahl, R. E., Lorentzen, S., Oglend, A. et al. 2017. Pro-Cyclical Petroleum Investments and Cost Overruns in Norway. *Energy Policy* **100**: 68–78. https://doi.org/10.1016/j.enpol.2016.10.004.

De Reyck, B., Grushka-Cockayne, Y., Fragkos, I. et al. 2017. Optimism Bias Study: Recommended Adjustments to Optimism Bias Uplifts. UK Department for Transport, UK, Horseferry Road, London.
Esmaeili, I. and Kashani, H. 2022. Managing Cost Risks in Oil and Gas Construction Projects: Root Causes of Cost Overruns. *ASCE-ASME J Risk Uncertain Eng Syst Part A: Civ Eng* **8** (1): 04021072. https://doi.org/10.1061/AJRUA6.0001193.
Flyvbjerg, B. 2014. What You Should Know about Megaprojects and Why: An Overview. *Proj Manag J* **45** (2): 6–19. https://doi.org/10.1002/pmj.21409.
Flyvbjerg, B. and Rasmussen, V. K. 2021. Heuristics for Masterbuilders: Fast and Frugal Ways to Become a Better Project Leader. *SSRN J*. https://doi.org/10.2139/ssrn.4159984.
Flyvbjerg, B. 2021. Top Ten Behavioral Biases in Project Management: An Overview. *Proj Manag J* **52** (6): 531–546. https://doi.org/10.1177/87569728211049046.
Flyvbjerg, B., Ansar, A., Budzier, A. et al. 2018. Five Things You Should Know about Cost Overrun. *Transp Res Part A Policy Pract* **118**: 174–190. https://doi.org/10.1016/j.tra.2018.07.013.
Geladi, P. and Kowalski, B. R. 1986. Partial Least-Squares Regression: A Tutorial. *Anal Chim Acta* **185**: 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.
Gneiting, T., Balabdaoui, F., and Raftery, A. E. 2007. Probabilistic Forecasts, Calibration and Sharpness. *J R Stat Soc B* **69** (2): 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x.
Griffin, D. and Tversky, A. 1992. The Weighing of Evidence and the Determinants of Confidence. *Cogn Psychol* **24** (3): 411–435. https://doi.org/10.1016/0010-0285(92)90013-R.
Haldorsen, H. H. and Damsleth, E. 1990. Stochastic Modeling. *J Pet Technol* **42** (4): 404–412. SPE-20321-PA. https://doi.org/10.2118/20321-PA.
Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning*, second edition. New York, USA: Springer Series in Statistics, Springer. https://doi.org/10.1007/978-0-387-84858-7.
Hoffrage, U. 2016. *Overconfidence*, second edition, Vol. 16, 291–314. Hove, UK, England: Psychology Press.
Howard, R. A. 1968. The Foundations of Decision Analysis. *IEEE Trans Syst Sci Cyber* **4** (3): 211–219. https://doi.org/10.1109/TSSC.1968.300115.
Hyndman, R. J. and Athanasopoulos, G. 2018. *Forecasting: Principles and Practice*, second edition. Melbourne, Australia: OTexts.
Jergeas, G. F. and Ruwanpura, J. 2010. Why Cost and Schedule Overruns on Mega Oil Sands Projects? *Pract Period Struct Des Constr* **15** (1): 40–43. https://doi.org/10.1061/(ASCE)SC.1943-5576.0000024.
Kahneman, D. 2011. *Thinking, Fast and Slow*, first edition. New York, USA: Farrar, Straus and Giroux.
Kahneman, D., Lovallo, D., and Sibony, O. 2011. *Before You Make That Big Decision*, 51–60. Brighton, Massachusetts, USA: Harvard Business Review.
Keelin, T. W. 2016. The Metalog Distributions. *Decis Anal* **13** (4): 243–277. https://doi.org/10.1287/deca.2016.0338.
Kraskov, A., Stögbauer, H., and Grassberger, P. 2004. Estimating Mutual Information. *Phys Rev E Stat Nonlin Soft Matter Phys* **69** (6 Pt 2). https://doi.org/10.1103/PhysRevE.69.066138.
Morris, S. 1990. Cost and Time Overruns in Public Sector Projects. *Econ Polit Wkly*: M154–M168.
Murphy, A. H. and Winkler, R. L. 1992. Diagnostic Verification of Probability Forecasts. *Int J Forecast* **7** (4): 435–455. https://doi.org/10.1016/0169-2070(92)90028-8.
Musavi, M. T., Ahmed, W., Chan, K. H. et al. 1992. On the Training of Radial Basis Function Classifiers. *Neural Netw* **5** (4): 595–603. https://doi.org/10.1016/S0893-6080(05)80038-3.
Nandurdikar, N. and Wallace, L. 2011. Failure to Produce: An Investigation of Deficiencies in Production Attainment. Paper presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 30 October–2 November. SPE-145437-MS. https://doi.org/10.2118/145437-MS.
Nesvold, E. and Bratvold, R. B. 2022. Debiasing Probabilistic Oil Production Forecasts. *Energy* **258**. https://doi.org/10.1016/j.energy.2022.124744.
NPD. 2022. Norwegian Petroleum Directorate Fact Pages. (accessed 30 January 2022).
Oglend, A. and Osmundsen, P. 2016. Cost Overruns in Norwegian Oil and Gas Projects: A Long-Tailed Tale. In *IAEE Energy Forum*. Bergen, Norway: Bergen Open Access Publishing.
Schölkopf, B. 2000. The Kernel Trick for Distances. In *Advances in Neural Information Processing Systems 13*. Cambridge, Massachusetts, USA: MIT Press Direct.
Smith, J. E. and McCardle, K. F. 1999. Options in the Real World: Lessons Learned in Evaluating Oil and Gas Investments. *Oper Res* **47** (1): 1–15. https://doi.org/10.1287/opre.47.1.1.
Taleb, N. 2007. *The Black Swan: The Impact of the Highly Improbable*. New York, New York, USA: Random House.

Appendix A—Transforms Applied to Field Features

Because of the natural distributions of field features related to volume and area, it is necessary to apply a transform to get something closer to a uniform or natural distribution. **Table A-1** describes the transforms applied to the features in **Table 1**. **Fig. A-1** shows an example for field area.

Feature Number	Transform
1	–
2	–
3	Oil reserves: square root
4	Mean volume forecast: square root
5	–
6	–
7	Investments per unit oil: square root of ratio
8	–
9	–
10	–
11	–

Table A-1—Transforms applied to field features.

Feature Number	Transform
12	–
13	–
14	Development wells per unit area: square root in denominator
15	Appraisal wells per unit area: square root in denominator
16	Number of DST logs per unit volume: square root in denominator

Table A-1 (continued)—Transforms applied to field features.

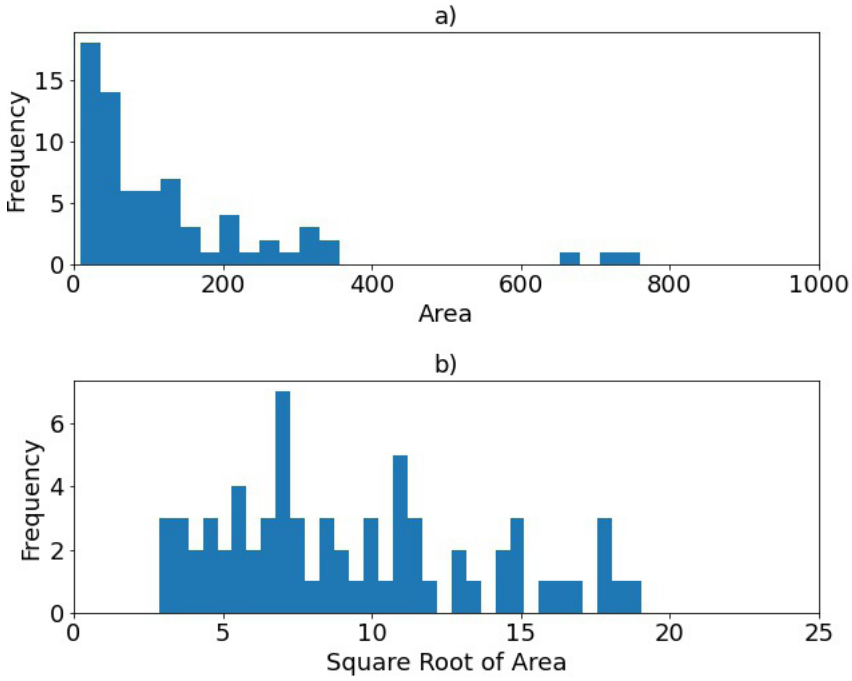

Fig. A-1—Square root transform.

Appendix B—Scale-Independent Forecast Score

The CRPS is a function of the production outcome y for a density forecast $F(x)$:

$$\mathrm{CRPS}(F,y)=\int (F(x)-\mathbb{1}[x\geq y])^2 dx. \tag{B-1}$$

To get a scale-independent score based on the CRPS, a log transform is applied to the percentiles of the CDF, and the CRPS in Eq. B-1 is computed. We refer to this term as $\mathrm{CRPS}_{log}(\cdot)$. However, this operation makes the score asymmetric about y—forecasts which are too high would be penalized less than low forecasts, so a symmetric score is computed by averaging the CRPS above and below the forecast mean M by the distance $|y-M|$. A Winkler term (Hyndman and Athanasopoulos 2018) is used outside the range $[F_{01}$–$F_{99}]$. In summary, the modified score used in this study is:

$$S(F,\mathrm{y})=\begin{cases}\mathrm{CRPS}_{log}(F,y^+)+\mathrm{CRPS}_{log}(F,y^-)+\dfrac{F_{01}-y}{F_{99}-F_{01}} & \text{if } y<F_{01}\\ \mathrm{CRPS}_{log}(F,y^+)+\mathrm{CRPS}_{log}(F,y^-) & \text{if } y\in[F_{01},F_{99}],\\ \mathrm{CRPS}_{log}(F,y^+)+\mathrm{CRPS}_{log}(F,y^-)+\dfrac{y-F_{99}}{F_{99}-F_{01}} & \text{if } y>F_{99}\end{cases} \tag{B-2}$$

where $y^+=y$ and $y^-=\max\{0,2M-y\}$. **Fig. 12** shows an example density forecast F as well as the resulting score $S(F,y)$.