QUANTIFYING UNCERTAINTY IN PRODUCTION FORECASTING USING MACHINE LEARNING

A Thesis

by

THOMAS MATTHEW BUTTON

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Duane McVay
Committee Members,	W. John Lee
	Siddharth Misra
	Shuang Zhang
Head of Department,	Akhil Datta-Gupta

December 2022

Major Subject: Petroleum Engineering

ABSTRACT

Estimating reserves—economically recoverable volumes of hydrocarbons in a company's portfolio—requires forecasting hydrocarbon production, which is prone to significant uncertainty and bias. Accurately quantifying this uncertainty is paramount to estimators understanding risk and projects meeting expectations.

Typically, production forecasts are made deterministically using Decline Curve Analysis (DCA). However, production forecasts can also be created probabilistically using Probabilistic Decline Curve Analysis (PDCA). In recent years, some reserves evaluators have turned to multivariate Machine Learning (ML) models to perform deterministic production forecasts, due to ML models' ability to handle large datasets and include properties other than production in the forecast. However, these models are deterministic and, to the best of my knowledge, there has been no standalone probabilistic adaptation published in the petroleum literature as of yet.

The aims of this research were to determine if a ML method was probabilistically reliable in forecasting production and to determine if the accuracy, probabilistic reliability, predicted uncertainty, and computational cost of this method was superior to an existing PDCA method.

A Gradient Boosting Regressor (GBR) was adapted to generate cumulative production predictions by training three separate models for each of the 10%, 50% and 90% quantiles. Predictions were made with this Gradient Boosting Regressor with Quantiles (GBRQ) method for future months based on the first 12 months of cumulative production history for the training wells, the target cumulative production at the forecasted month for the training wells, and the first 12 months of cumulative production history for the test wells.

Prediction accuracy was measured using the root mean square error (RMSE) between the predicted median (P50) and true values as well as between the predicted mean and true values. Probabilistic reliability was assessed using calibration plots in which the frequency with which actual production values were less than predicted production values at each quantile was plotted against the assigned probability. Predicted uncertainty was assessed using an average normalized uncertainty window and cost was compared on the basis of computational time.

The GBRQ method was more accurate at late times, was more probabilistically reliable, predicted less uncertainty, and was less computationally intensive than a published Probabilistic Decline-Curve-Analysis (PDCA) method for a dataset consisting of 438 conventional wells in the Midland Basin.

The GBRQ methodology can be useful to three groups: (1) reserves estimators, who can make point estimates and full forecasts of probabilistic production comparatively fast and with probabilistic reliability for large datasets; (2) reserves auditors, who can quickly use this method to compare with an auditee’s probabilistic production forecast; and (3) investors and banks, who can evaluate asset acquisitions and divestitures with well-calibrated probabilistic production forecasts.

DEDICATION

To my mother, Rene, and my father, Alvin who have supported throughout my graduate studies and career development.

ACKNOWLEDGEMENTS

I would like to thank Dr. Duane A. McVay, my committee chair for his fastidious guidance and support during my thesis research. I would also like to thank Dr. John Lee, Dr. Siddharth Misra, and Dr. Shuang Zhang for serving on my advisory committee.

I would also thank the Harold Vance Department of Petroleum Engineering at Texas A&M University for providing me a world class education and the tools needed to succeed in my future career.

Lastly, I would like to thank the lifelong friends I have made at Texas A&M University who inspired me to succeed and made learning such an enjoyable experience.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

The student's work was supervised by an advisory committee consisting of Dr. Duane A. McVay, Dr. W. John Lee, and Dr. Siddharth Misra of the Department of Petroleum Engineering and Dr. Shuang Zhang of the Department of Oceanography. All work was completed independently by the student.

Funding Sources

No outside funding was received for the research and compilation of this thesis.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1 Statement and Significance of the Problem

Reserves estimation is used to quantify the economically recoverable volumes of hydrocarbons in a company's portfolio. This involves forecasting hydrocarbon production, which is a process prone to significant uncertainty and biases. Accurately assessing this uncertainty is paramount to projects meeting expectations and estimators understanding risks. However, the hydrocarbon exploration industry does a poor job of assessing uncertainty. In fact, poor assessment of uncertainty has long plagued the U.S. oil and gas industry with overconfidence and optimism being especially prevalent (McVay and Dossary 2014).

One of the older and more well-known studies of overconfidence and optimism in the petroleum industry was in Capen's experiment as part of a Society of Petroleum Engineers (SPE) Distinguished Lecturer tour (1976). Capen asked 10 general-knowledge trivia questions to over 1200 petroleum professionals at local section meetings requiring estimates of a 90% confidence interval around the correct answer. Rather than produce a 90% confidence interval, the attendees produced, on average, a 32% confidence interval. This interval was narrower than expected and showed that the professionals were vastly overconfident. Capen also noted that the predictions attendees made tended to be optimistic. Despite observing their estimates to be overconfident, attendees did little to adjust their estimates in subsequent answers. Just as many professionals were biased in their answers to the general trivia questions, so are they biased in making project decisions. Capen pointed to investment underperformance in the petroleum industry and other institutions resulting from biased estimations.

A quarter century later, Brashear et al. (2001) compared the return on net assets for the largest U.S.-based E&P companies to their hurdle rates from 1990 to 2000. The authors found that the return on net assets was only 7% for projects with hurdle rates of 15%. The ranking of projects using deterministic methods overstated value, understated risk, and misallocated capital, according to the authors. Overstated values and understated risks imply overconfidence and optimism.

Fast forward 18 more years and overconfidence and optimism are still consistently present. The Wall Street Journal (2019) collected production forecasts and actual production data from 2014 to 2017 for 16,000 wells in Texas and North Dakota. Companies produced 10% less oil on average than their forecasts indicated. One company produced 25% less oil than it forecasted, three years in a row. The Wall Street Journal (WSJ) contended that many companies made extrapolations based on small clusters of prolific initial wells" while excluding the worst performing wells from the forecasts. This could explain why forecasts were so optimistic.

As a result, banks got tough on shale loans (WSJ 2019). Loan growth in the Permian Basin shrunk to 4.8% in the third quarter of 2019, far below the 7.5% average for Texas. Smaller operators were not able to handle restricted access to capital. As a result, 274 oil and gas producers filed bankruptcy alongside 330 oilfield services and midstream companies (Haynes and Boone 2022).

1.2 Status of the Question

The best way to measure overconfidence and optimism is to conduct lookbacks and compare probabilistic estimates to actual values (Alarfaj and McVay 2020). Since the predictions are continuous probabilistic assessments, they are expressed in terms of a cumulative distribution function (CDF). This research uses a conventional CDF, in which the P10 is a low value and the P90 is a high value. Using a conventional CDF, the proportion correct at a quantile is the proportion of the predicted values that are greater than the actual values (McVay and Dossary 2014). Thus, the P10 prediction is expected to be greater than the actual value for cumulative production only 10% of the time. Following this, a P50 prediction is expected to be greater than the actual value 50% of the time and a P90 prediction is expected to be greater than the actual value 90% of the time.

Deviation from perfect reliability results in measurable biases including overconfidence and optimism, as mentioned before (McVay and Dossary 2014). Overconfidence means that only a subset of the predicted distribution has been sampled and optimism means the distribution has shifted towards more desirable outcomes. An example of an estimated distribution that is both overconfident and optimistic is shown in **Fig. 1.1**:

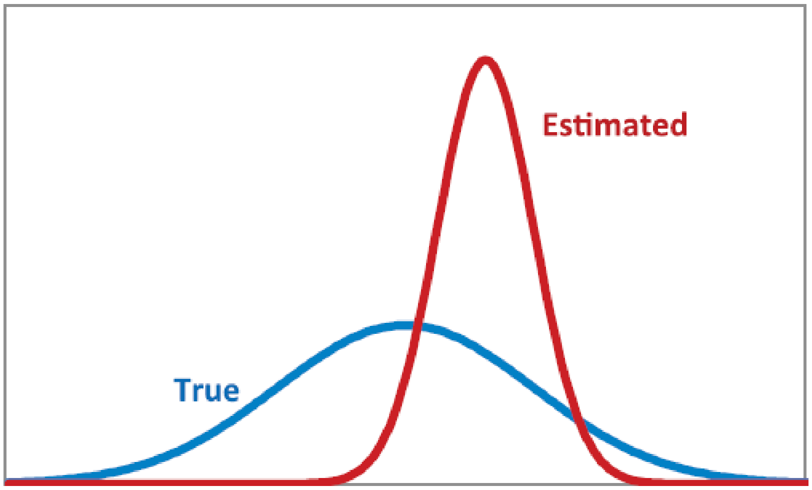


Fig. 1.1—Distribution changes due to overconfidence and optimism (McVay 2015).

To represent the fraction of the true distribution represented in the estimated distribution, the confidence bias parameter was first introduced by McVay and Dossary in 2014. A positive confidence bias parameter represents overconfidence, which is the underestimation of uncertainty while a negative confidence bias parameter represents underconfidence, which is the overestimation of uncertainty. Since underconfidence is not common, it will not be elaborated on in this thesis. Overconfidence bias ranges from zero to one. A value of zero means the entire true distribution was sampled while a value of one means a single point estimate was sampled. **Fig. 1.1** shows an estimated distribution in red that is narrower than the true distribution and is thus overconfident.

McVay and Dossary (2014) also introduced the directional bias parameter, which represents the shift of the estimated distribution relative to the true distribution. Directional bias ranges from negative one to positive one. A value of negative one signifies complete pessimism and means only the lowest possible outcomes were considered. A value of positive one signifies complete optimism and means only the highest possible outcomes were considered. A value of zero means no shift in the outcomes was considered. For value-based assessments, a value of one would be a rightward shift and for cost-based assessments, a value of one would be a leftward shift. In this case, cumulative production is a value-based assessment, so an optimistic bias would be a rightward shift. The rightward shift of the red curve in **Fig. 1.1** demonstrates the effect of optimism on a distribution.

These biases can be visualized and measured with a calibration plot in which the proportion of correct outcomes is plotted against the probability assigned. As a conventional cumulative distribution function is used, this would mean that for a given quantile, the proportion correct is the proportion of predictions greater than the actual values. In **Fig. 1.2**, overconfidence bias limits the size of the orange predicted distribution in the associated subplot, and this changes the slope of the calibration plot. The orange area in the below subplot represents the area of the full estimated distribution represented in the predicted distribution. As the overconfidence bias grows in magnitude, fewer of the values from the original distribution are sampled and this causes an associated decrease in slope on the calibration plot.

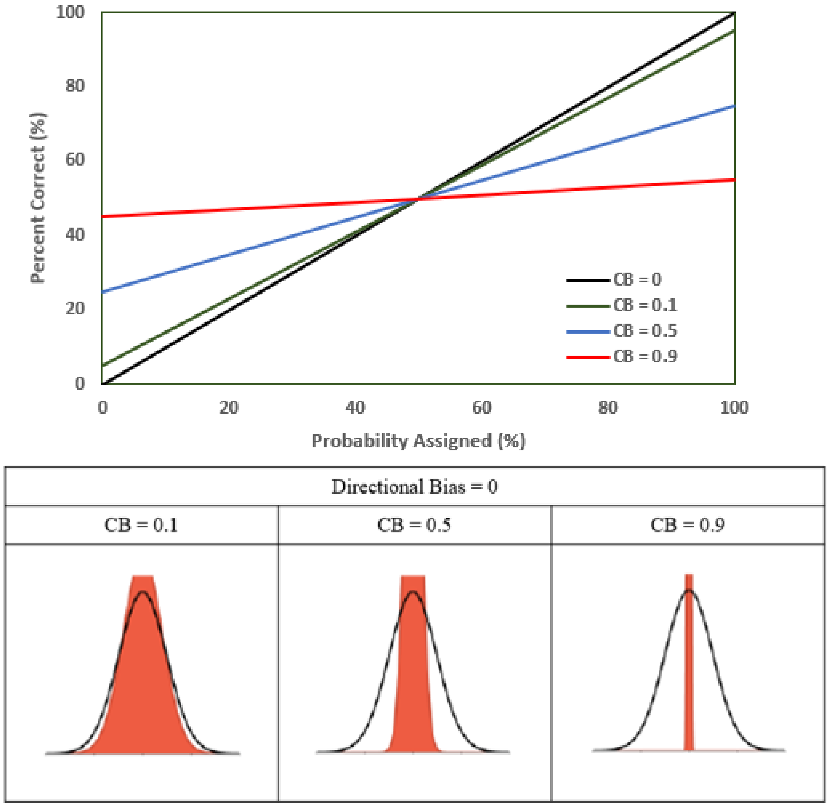


Fig. 1.2—Varying CB effect on full distributions (Modified from Alarfaj and McVay 2020).

In **Fig. 1.3**, directional bias shifts the distribution in the below subplot, and this results in a vertical translation on the calibration plot. Assuming an overconfidence bias of 0.5 in which the original distribution shown in orange is truncated, an increase in directional bias results in an upward translation of the line on the calibration plot while a decrease in directional bias results in a downward translation.

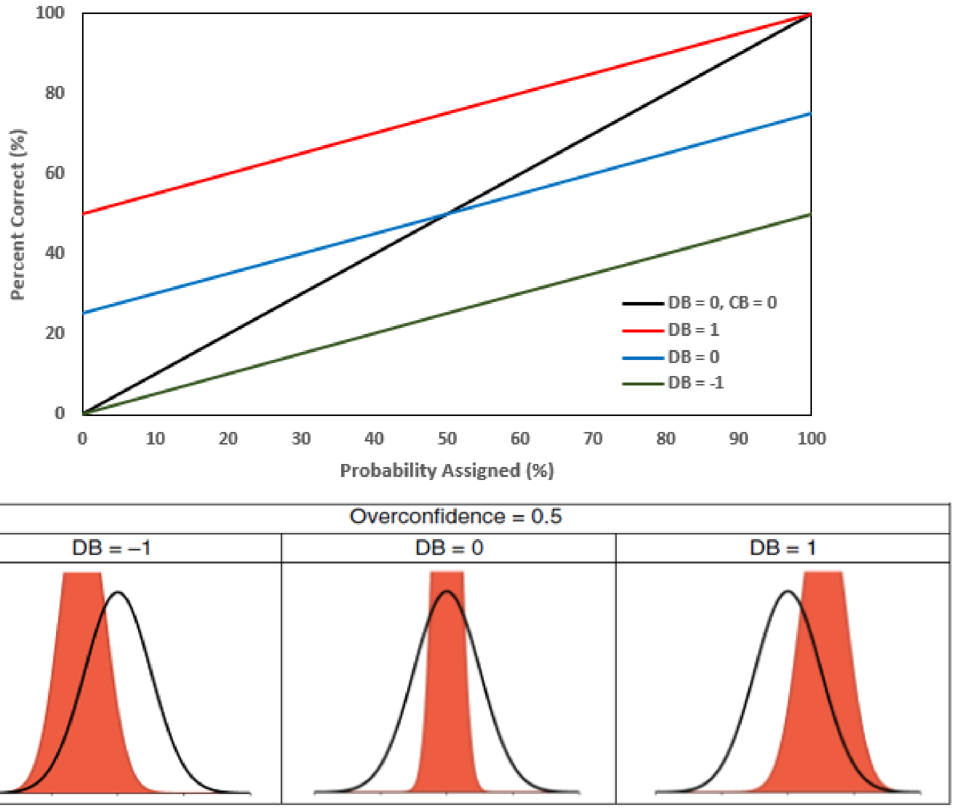


Fig. 1.3—Varying DB effect on full distributions (Modified from Alarfaj and McVay 2020).

McVay and Dossary (2014) demonstrated with simulation that even moderate levels of overconfidence and optimism could result in as much as 30% to 35% average reduction from estimated to realized portfolio values (Alarfaj and McVay 2020). Therefore, reliably measuring these biases is very important in measuring and understanding associated impacts on portfolios.

Accurate production forecasting is necessary so economic expectations are met throughout a well's life. Early in the life of a well when there is an absence of data to create reservoir simulations or geologic models, production forecasting is typically performed using an empirical decline curve model in a process known as Decline Curve Analysis (DCA). The most frequently used method in the oil and gas industry is the Modified Arps model due to its simplicity in application and general industry-wide understanding, according to Li, Billiter, and Tokar (2021):

$$q = \begin{cases} q_i(1+bD_it)^{-1/b}, & t \leq t_{lim} \\ q_{lim}exp[-D_{lim}(t-t_{lim})], & t > t_{lim} \end{cases} \quad \text{.........(1)}$$

$$t_{lim} = \frac{\frac{D_i}{D_{lim}}-1}{bD_i} \quad \text{.........(2)}$$

$$q_{lim} = q_i\left(\frac{D_{lim}}{D_i}\right)^{1/b} \quad \text{.........(3)}$$

Decline curve models, however, are typically applied deterministically; that is, uncertainty is not typically quantified. To quantify uncertainty, Probabilistic Decline Curve Analysis (PDCA) has been pursued as a topic of research. Jochen and Spivey (1996) introduced a bootstrap method and Cheng et al. (2010) developed a Modified Bootstrap Method (MBM) for probabilistic production forecasting. Gong et al. (2014) expanded on this work by creating a Bayesian probabilistic methodology using Markov-chain Monte Carlo (MCMC) coupled with Arps' DCA to quantify uncertainty in production forecasting. Kuzma et al. (2014) also created a generative model (GM) with the aim of simulating real noise and artifacts in the production history. While these methods have been shown to generate probabilistically well-calibrated forecasts in certain applications, computational times are slow: 30 seconds per well for the Gong et al. MCMC PDCA and three to five minutes per well for the Cheng et al. MBM PDCA method. The Kuzma et al. paper does not provide computational detail nor enough evidence of probabilistic calibration. These methods also face the same problems as with decline curves. That is, parameters for a model are fit to a well's production history and generate a smooth prediction curve. However, this does not consider underlying reservoir signals, such as boundaries not seen in the production data, production from other wells, and interactions with other wells.

As technology progressed to handle more data, forecasters began using multivariate Machine Learning (ML) methods to forecast or assist in forecasting production. ML methods are divided into supervised and unsupervised learning with the main difference being the use of labels for data to provide context in supervised learning algorithms. The labels split the data into features and targets, which provides supervised learning algorithms a channel in which to learn the relationship between the two entities. Supervised learning algorithms can further be divided into classification and regression in which classification algorithms classify testing data into various categories and regression algorithms identify a relationship between dependent and independent variables.

Clustering, association, and dimensionality reduction are examples of unsupervised learning algorithms with at least one energy exploration application of identifying spatial importance in well location for production forecasting (Harris 2014). Linear regression, logistic regression, support vector machines, random forests, artificial neural networks, and gradient boosting regression are examples of supervised learning algorithms. One application in production forecasting has been to use a neural network to investigate the pattern between selected reservoir and hydraulic fracture parameters and decline parameters for a logistic growth model (Li and Han 2017). This method also employed principal component analysis, a form of unsupervised learning, to quantify variance in the principal component space and the key factors that influence production rate.

Another application has been to learn the posterior distribution of model parameters for a transient hyperbolic DCA model given production data and a specification of prior beliefs (Fulford et al. 2016). Another unique application has been to forecast cumulative production using one year of cumulative production and geographic, wellbore, spacing, and completions properties as well as a dynamic production rescaling method (Li, Billiter, and Tokar 2021).

In these ML applications, the accuracy of deterministic production forecasts improved when compared to conventional decline curve models. Uncertainty was quantified in only one paper (Fulford et al. 2016), which used a ML algorithm to identify parameters for the transient hyperbolic model. The paper used the same PDCA method as Gong et al. (2014) with the simple addition of a regression algorithm to learn the prior distribution of decline curve model parameters. This means the same pitfalls of PDCA, that is, fitting smooth models that do not capture underlying reservoir signals or include production from other wells, is present even in the only application of machine learning to assist in probabilistic production forecasts.

In conclusion, PDCA methods reliably quantify uncertainty but are computationally intensive and do not capture underlying reservoir signals or production from other wells in the forecast. ML methods have been developed to forecast production, but none function as standalone probabilistic methods. The Fulford et al. (2016) application of ML assisted PDCA in finding a prior distribution of decline curve parameters, so the same limitations that apply to PDCA also apply to those authors' analysis. The ability for ML methods to capture underlying reservoir signals and production from other wells is also a handicap. Production forecasts cannot be made past times for which there is no production in other wells. There is a need for a standalone probabilistic ML method that can reliably assess uncertainty in production forecasts.

1.3 Research Objectives

The objectives of this thesis were to determine if a Gradient-Boosting-Regressor ML regression method was probabilistically reliable in forecasting production and to determine if the accuracy, probabilistic reliability, predicted uncertainty, and computational cost of this ML method was superior to an existing PDCA production forecasting method.

2. METHODOLOGY

2.1 General Steps

The steps for conducting my research were as follows:

1. Gathered production data. In this case with limited availability of data, only publicly available data were used. The wells needed to be in the same reservoir, produce primarily oil or gas, and either consist of all vertical or all horizontal wells. In my work, only vertical oil producing wells were chosen. The dataset also must have at least 40 wells with more than 20 months of production history to have been considered. The first dataset contained 40 wells with at least 21 months of production history and the second dataset contained 438 wells with at least 21 months of production history

2. Preprocessed the data. Ensured that just production data were included, then removed ramp-up and downtime production data, as is common in production forecasting. Production rate data are converted to cumulative production data and arranged into features and targets for the GBRQ method. Time data were converted to elapsed time with the removal of ramp-up and downtime production.

3. Generated predictions for a specific month using the GBRQ and PDCA methods.

4. Generated a calibration plot for each method using the predictions generated from Step 3 to compare accuracy, probabilistic reliability, predicted uncertainty, and computational cost.

5. Repeated steps three and four to make predictions for each month.

2.2 Metrics for Comparison

As mentioned before, accuracy, probabilistic reliability, predicted uncertainty, and computational cost made up the basis for comparison.

Accuracy was measured using the root mean square error (RMSE) between the predicted median (P50) and true values, as well as between the predicted mean and true values where the mean was estimated using Swanson's rule. Production data are typically distributed lognormally with a right skew. A distribution of cumulative production for the Midland dataset at Month 440 shows exactly this behavior below (**Fig. 2.1**). The mean was 30,704 STB while the median was 23,206 STB, verifying the right skew of the distribution.

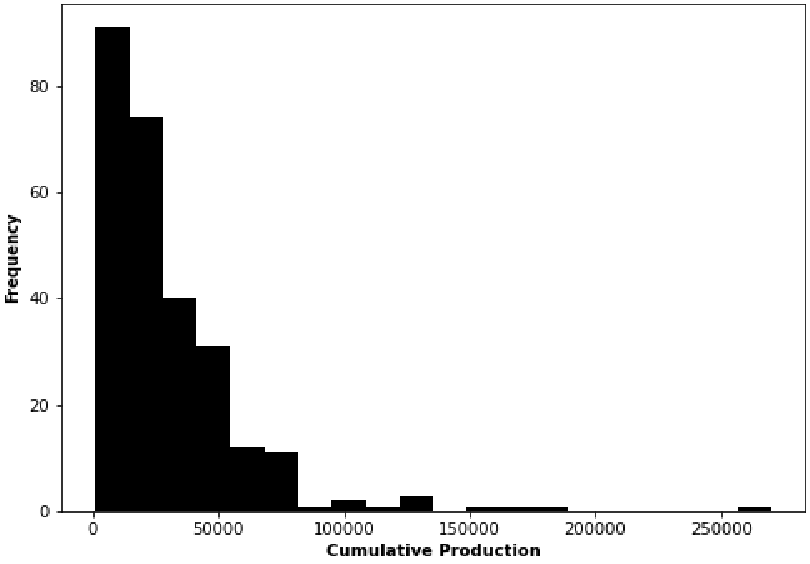


Fig. 2.1—Histogram of Month 440 cumulative production (Midland dataset).

Probabilistic reliability was assessed using calibration plots in which the frequency of actual production values less than the predicted production values at each quantile was plotted against the estimated probability.

From calibration plots, the calibration score, biases, and coverage ratio can be calculated. The calibration score was given by Lichtenstein and Fischhoff (1977):

$$Calibration\ Score\ (CS) = \frac{1}{N}\sum_{t=1}^{T} n_t(r_t - c_t)^2 \qquad (4)$$

where N is the total number of responses, n_t is the number of times the response was used, r_t is the probability assigned, c_t is the proportion of items greater than the actual for the probability assigned, and T is the total number of response categories used. A perfectly calibrated model has a calibration score = $((0.9\text{-}0.9)^2 + (0.5\text{-}0.5)^2 + (0.1\text{-}0.1)^2)/3 = 0$. The calibration score includes the effects of both confidence bias and directional bias. Thus, it was the primary metric for comparing probabilistic reliability.

The confidence and directional biases can be observed from calibration plots as changes in slope and vertical translations respectively. However, they can also be directly calculated. Alarfaj and McVay (2020) developed equations to relate slope m and intercept a of a calibration plot to confidence bias (CB) and directional bias (DB) as follows:

$$CB_{OC} = 1 - m \qquad (5)$$

$$DB_{OC} = \frac{2a}{1-m} - 1 \qquad (6)$$

$$CB_{UC} = \frac{1}{m} - 1 \qquad (7)$$

$$DB_{UC} = 1 - \frac{2a}{1-m} \qquad (8)$$

The subscript OC represents overconfidence while the subscript UC represents underconfidence.

Coverage ratio represents the portion of the distribution of actual values sampled in the predicted distribution and is the inverse of confidence bias. This can be calculated as $CR = \frac{c_{90}-c_{10}}{90-10}$, where c_{90} is the proportion of items greater than the actual for an assigned probability of 90% and likewise for c_{10}. A perfect coverage ratio would be 1.

The following is an example for measuring the biases. **Fig 2.2** below is a sample calibration plot of synthetic data.

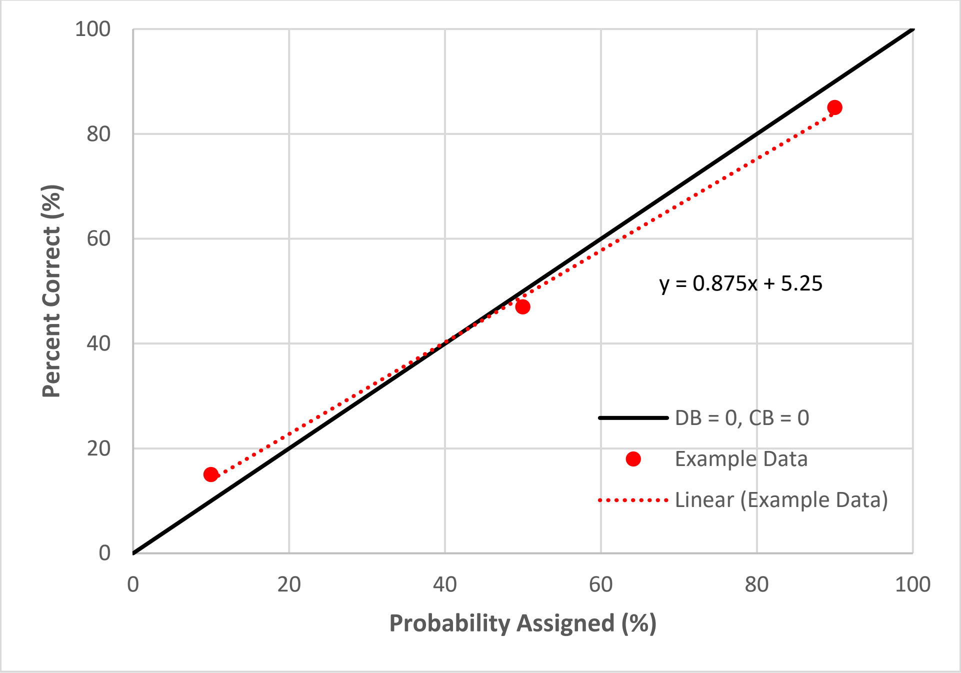


Fig. 2.2—Example calibration plot.

The above data points represent P10, P50, and P90 estimates. More specifically, the x coordinates represent the probability assigned while the y coordinates represent proportion of predictions greater than the actual values at the assigned probability. The predictions for the 10th percentile in this case turned out to be greater than the actual value 15% of the time (10,15), the predictions for the 50th percentile were greater than the actual 47% of the time (50,47), and the predictions for the 90th percentile were correct 85% of the time (90,85). The anticipated probability range for the P90-P10 interval was 90 – 10 = 80%, but the actual range was 85 – 15 = 70%. The narrower actual range means underestimation of uncertainty and thus overconfidence. This is verified by using **Eq. 5** to calculate CB_{OC} as 1 – 0.875 = 0.125. This small positive value verifies the presence of subtle overconfidence. Directional bias DB_{OC} can be calculated with **Eq. 6** as $\frac{2(.0525)}{0.125} - 1 = -0.16$. This means the predictions were slightly pessimistic.

Predicted uncertainty was assessed using an average normalized uncertainty window. This was found by calculating the difference between the P90 and P10 values divided by the P50 value and averaging for all the wells at a given month. This quantifies how much uncertainty is forecasted. If two methods have equal accuracy and probabilistic reliability, the method with less predicted uncertainty is the superior probabilistic method.

Cost was compared on the basis of computational time. The method that takes less computational time is preferable.

3. COMPARISON OF THE RELIABILITY OF GBRQ AND PDCA

3.1 Data Acquisition and Cleaning

There were two datasets analyzed: a 130 well set from the DJ Basin and a 448 well set predominantly from the San Andres Formation in the Permian Basin. The 130 well dataset will hereafter be referred to as the "DJ dataset" and the 448 well dataset will hereafter be referred to as the "Midland dataset." Both datasets consisted of publicly available data gathered from Enverus. As downtime and ramp-up time were removed, production history for available wells was compressed. This resulted in fewer wells with at least 21 months of production history. Given that the vast majority of the 130 wells in the DJ dataset had histories shorter than 21 months due to them being unconventional, the final well count was 40 out of 130 with at least 21 months of production history. There were even fewer wells with at least 24 months (or two years) of production history, so the forecast on the DJ dataset was limited to 21 months. The Midland dataset had significantly more wells with longer production histories. There were 438 wells that had at least 24 months of production history and 270 wells that had at least 440 months. The DJ dataset was useful to see how the GBRQ and PDCA methods performed on small dataset sizes while the Midland dataset was useful for comparing the two methods at longer forecast lengths. The concentration of wells for the Midland dataset is shown by orange dots in **Fig. 3.1**.

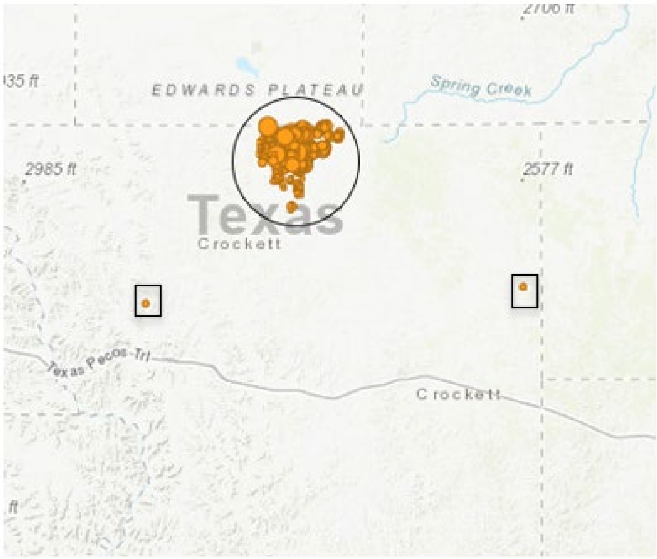


Fig. 3.1—Midland dataset before area of interest filter.

It is obvious that two wells are not geographically close to the other wells. This may indicate they are in a different reservoir. In fact, the well to the southwest of the main cluster is 17.6 miles away from the southernmost well in the cluster while the southeastern well is 24.8 miles from the southernmost well in the cluster. The other wells are less than two miles from each other, so it is obvious that these two wells should be removed from the analysis. The area of focus then becomes the 448 closely grouped wells (**Fig. 3.2**). Of these 448 wells, 438 have at least 21 months of production history, which is necessary for comparison with the DJ dataset.

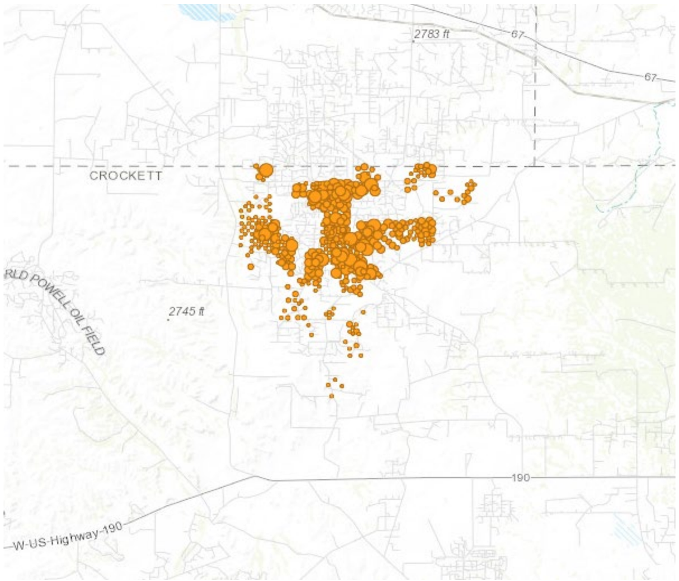


Fig. 3.2—Midland dataset after area of interest filter.

Once the area of interest filter is applied, production data characteristics were examined. The Midland dataset essentially consisted of six different types of behavior (**Fig. 3.3**). Out of 438 wells, the vast majority of wells in the dataset behaved similar to wells 127188628 in the top right and 127663805 in the bottom right, the only difference being a short steep initial decline in the top right well. Fewer than five wells behaved like 127240426 in the middle left with a spike in noise while declining from middle to late times. Fewer than five wells also behaved similar to 127312421 in the middle right with significant noise at middle times only. About 15 wells experienced a spike in production at 260 months similar to well 127178493 in the top left, which could indicate the beginning of some effort to enhance production at the field level, such as artificial lift. No significant jumps in production were observed at other times besides Month 260. Lastly, well 127363981 in the bottom left is a standalone case of a significant step change at Month 110. This does not occur in other wells.

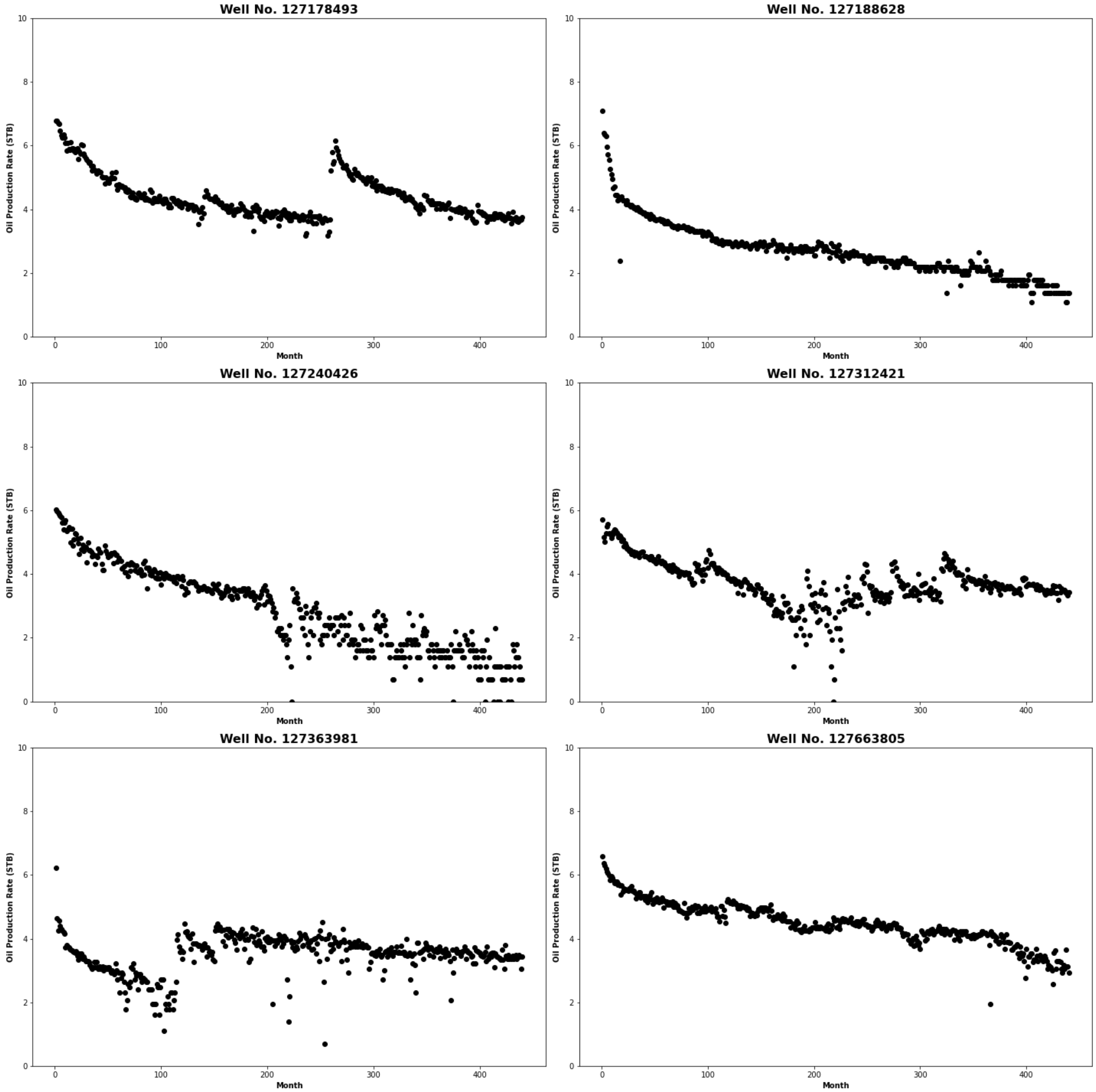


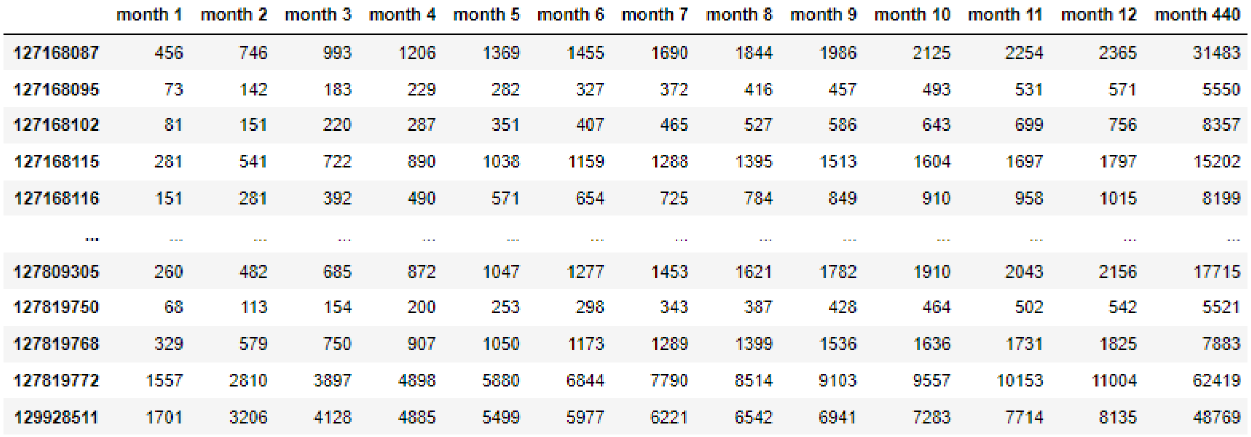
Fig. 3.3—Representative production cases for Midland dataset.

These representative cases demonstrate that while most wells experienced long-term exponential decline without significant noise or interruption, the dataset contains outlier cases that add uncertainty to production forecasts.

3.2 Data Preprocessing

The DJ dataset consists of monthly oil production rate vs time data while the Midland dataset consists of monthly oil, gas, and water production rate vs time data as well as activity status. For this research, only monthly oil production rate vs time data were needed. Since Python was used to do the analysis, production data were filtered using the Pandas library.

Next, as is commonly performed with traditional DCA, ramp-up production (q before q_{max}) and downtime ($q = 0$) were removed. The DJ dataset consisted of much higher production, so downtime was removed using a minimum of 200 STB/M instead of zero flow rate. Next, feature and target selection were performed so that the GBRQ method could identify the relationship between the features (predictor variables) and targets (response variables). The 12 features in this case were the first 12 months of cumulative oil production and the target was the month forecasted to. In other words, one year of production history was used to forecast to various point estimates of cumulative production in the future. **Fig. 3.4** shows the structure of the preprocessed data. The first unlabeled column is well number, the next 12 columns are the 12 features, and the last column is an example target of Month 440.



	month 1	month 2	month 3	month 4	month 5	month 6	month 7	month 8	month 9	month 10	month 11	month 12	month 440
127168087	456	746	993	1206	1369	1455	1690	1844	1986	2125	2254	2365	31483
127168095	73	142	183	229	282	327	372	416	457	493	531	571	5550
127168102	81	151	220	287	351	407	465	527	586	643	699	756	8357
127168115	281	541	722	890	1038	1159	1288	1395	1513	1604	1697	1797	15202
127168116	151	281	392	490	571	654	725	784	849	910	958	1015	8199
...	...	...	...	...	...	...	...	...	...	...	...	...	...
127809305	260	482	685	872	1047	1277	1453	1621	1782	1910	2043	2156	17715
127819750	68	113	154	200	253	298	343	387	428	464	502	542	5521
127819768	329	579	750	907	1050	1173	1289	1399	1536	1636	1731	1825	7883
127819772	1557	2810	3897	4898	5880	6844	7790	8514	9103	9557	10153	11004	62419
129928511	1701	3206	4128	4885	5499	5977	6221	6542	6941	7283	7714	8135	48769

Fig. 3.4—Dataframe ready for GBRQ analysis.

3.3 Generate Predictions

3.3.1 Gradient Boosting Regressor with Quantiles (GBRQ)

A Gradient Boosting Regressor (GBR) is a supervised machine learning regression algorithm that constructs an additive model in forward stages, which can be trained with various differentiable loss functions. Each stage involves fitting a decision tree to the negative gradient of the loss function. In layman's terms, a weak learner, which is usually a decision tree, is trained and predicted on the training data and the residuals between the prediction and training data are used to train the next decision tree. The weight coefficients of the next decision tree are fit to the residuals of the previous tree and new predictions are made with accompanying residuals. The process continues with the new residuals being used to train the weight coefficients of the next tree. **Fig. 3.5** shows this process.

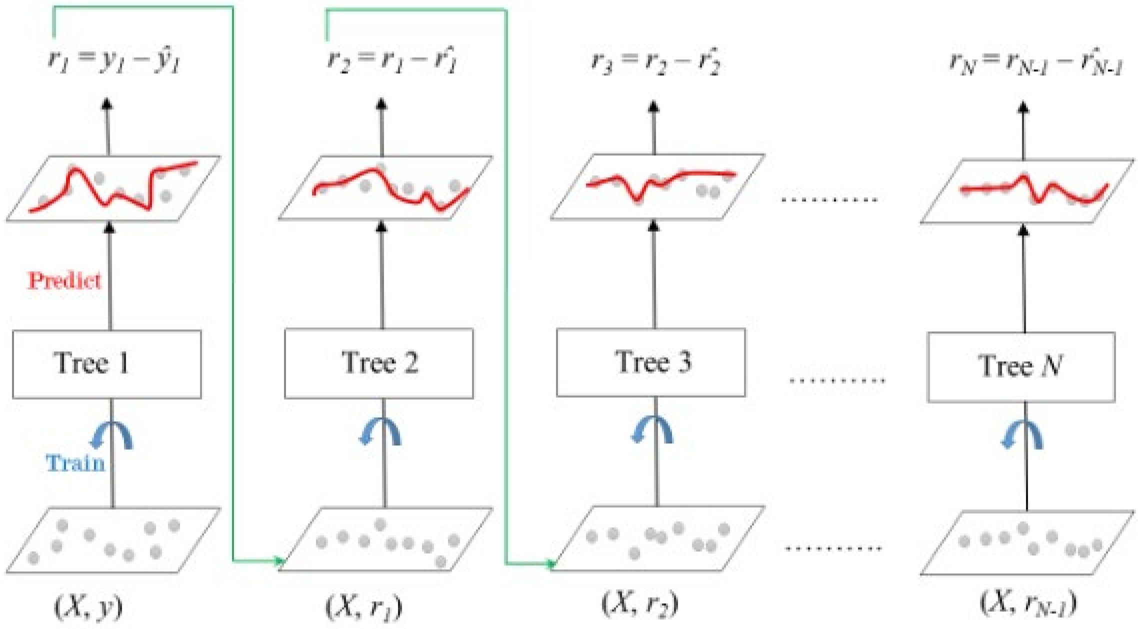


Fig. 3.5—Diagram of a Gradient Boosting Regressor (Siakorn, Wikimedia Commons).

Boosting is designed to create strong learners from weak learners. In the context of machine learning, weak learners are marginally better than random guessing. While it may seem logical to begin with a strong learner, this limits the learning process and introduces significant bias, which is not ideal. Boosting contrasts with bootstrap aggregation, or "bagging," which is common for algorithms such as random forest. Bagging involves the creation of many decision trees that sample with replacement from the original distribution of data and the results are aggregated into a final model (**Fig. 3.6**). Bagging limits learning because new trees do not learn from older trees. In boosting, successive trees are built from previous trees, which all learn in conjunction as opposed to in isolation. This results in computational and accuracy improvements.

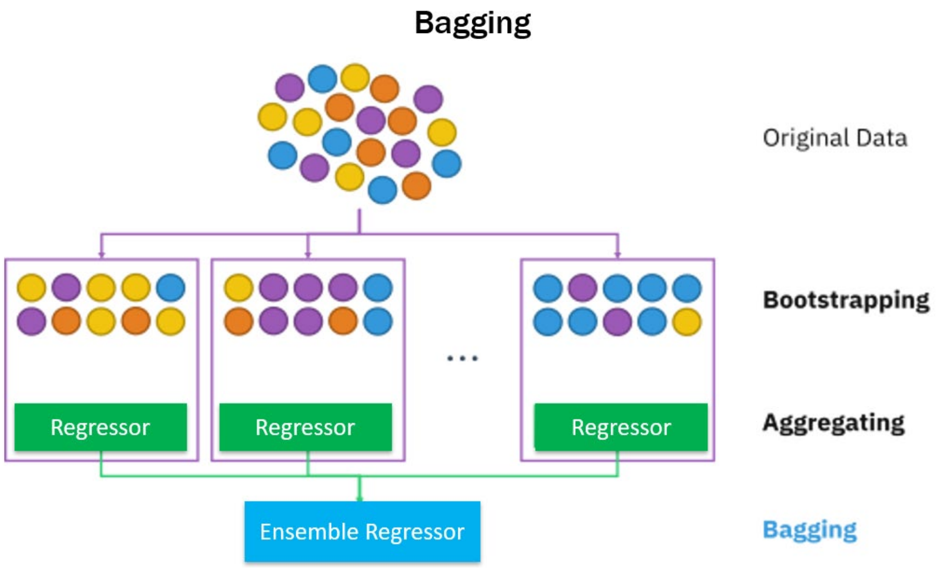


Fig. 3.6—Diagram of bagging (Siakorn, Wikimedia Commons).

The optimization of ML algorithms involves minimizing the loss function. A loss function measures how far a prediction is from the true value. The GBR model can be used with different loss functions: squared error, absolute error, huber, and quantile. The quantile loss function was used as it can quantify uncertainty by making predictions fit to different quantiles. The quantile loss function is defined as follows:

$$L(y_i{}^p, y_i) = max[Q(y_i - y_i{}^p), (Q-1)(y_i - y_i{}^p)] \qquad (9)$$

where $L(y_i^p,y_i)$ is the loss function, y_i is the actual value at the i^{th} data point, y_i^p is the predicted value at the i^{th} data point, Q is the quantile, and *max* refers to the max value within the brackets. Since Q is between zero and one, the first term is positive and dominates when underpredicting while the second term is positive and dominates when overpredicting. When trying to predict the median ($Q = 0.5$), both terms are penalized equally. When $0.5 < Q < 1$, underpredictions are penalized more heavily, and when $0 < Q < 0.5$, overpredictions are penalized more heavily.

Gradient boosting uses the residuals of the previous iteration to train the decision tree in the next iteration, but these residuals will be fit to quantiles instead of the exact value. This allows the GBR method to generate predictions for a quantile.

In probabilistic estimates of production, the analyst performing forecasts is commonly interested in the P10, P50, and P90 values. These correspond to the 10th percentile, 50th percentile, and 90th percentile likelihood of production on a cumulative distribution function. The GBR model by itself can only fit to one quantile, so three GBR models must be created. These three models made up the GBRQ approach.

Besides the features and targets, which are used to train the model parameters, hyperparameters are used to control the learning process. Normally in ML model training, the dataset is split into one training and one testing fold (as well as one validation fold in some applications). The model is trained on the training fold and tested on the testing fold and hyperparameters are tuned to achieve the most accurate predictions. However, the goal in this case is to achieve reliable probabilistic calibration and not necessarily highest accuracy. Furthermore, the effects of hyperparameter tuning on probabilistic calibration have not been properly established and a link may not exist. In initial testing, the three different GBR models at different quantiles had different sets of hyperparameters optimized for accuracy, which changed depending on how the data were split. Intuitively, it does not make sense to have different hyperparameters for different quantiles of the same dataset. After tuning model hyperparameters for accuracy, an increase in calibration score was observed. Thus, given these observations and concerns, the default hyperparameters of a 0.1 learning rate, 100 decision trees, subsample fraction equal to one, and three nodes were used and no hyperparameter tuning was done.

Rather than use a conventional approach with one training and one testing split to generate predictions, a process known as Cross Validation Prediction was used. The data were randomly and equally split into 10 folds. The model was trained on nine folds and made predictions on the outlying testing fold. The same splits were maintained, but a different fold was then used as the testing fold. Ultimately, the process was repeated until each fold was the outlying testing fold once and thus predictions were made for every single well in the dataset (**Fig. 3.7**). This process was applied to each of the three GBR models so that there were P10, P50, and P90 predictions for each well. The Cross Validation Prediction process makes sure that the splitting of the data does not control the outcomes of training and testing, which is important for skewed datasets like this one (**Fig. 2.1**). The entire process from importing the data to generating predictions is detailed in **Fig. 3.8**.

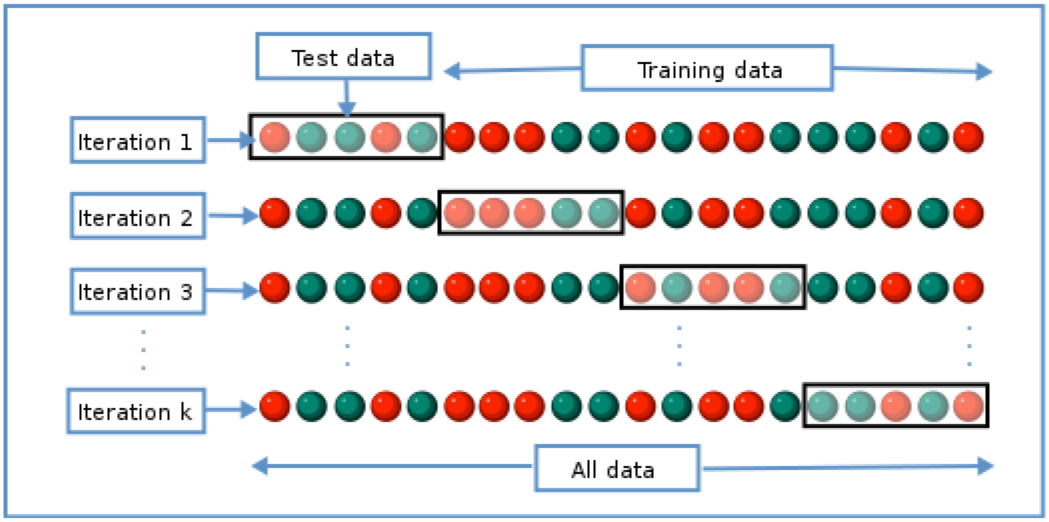
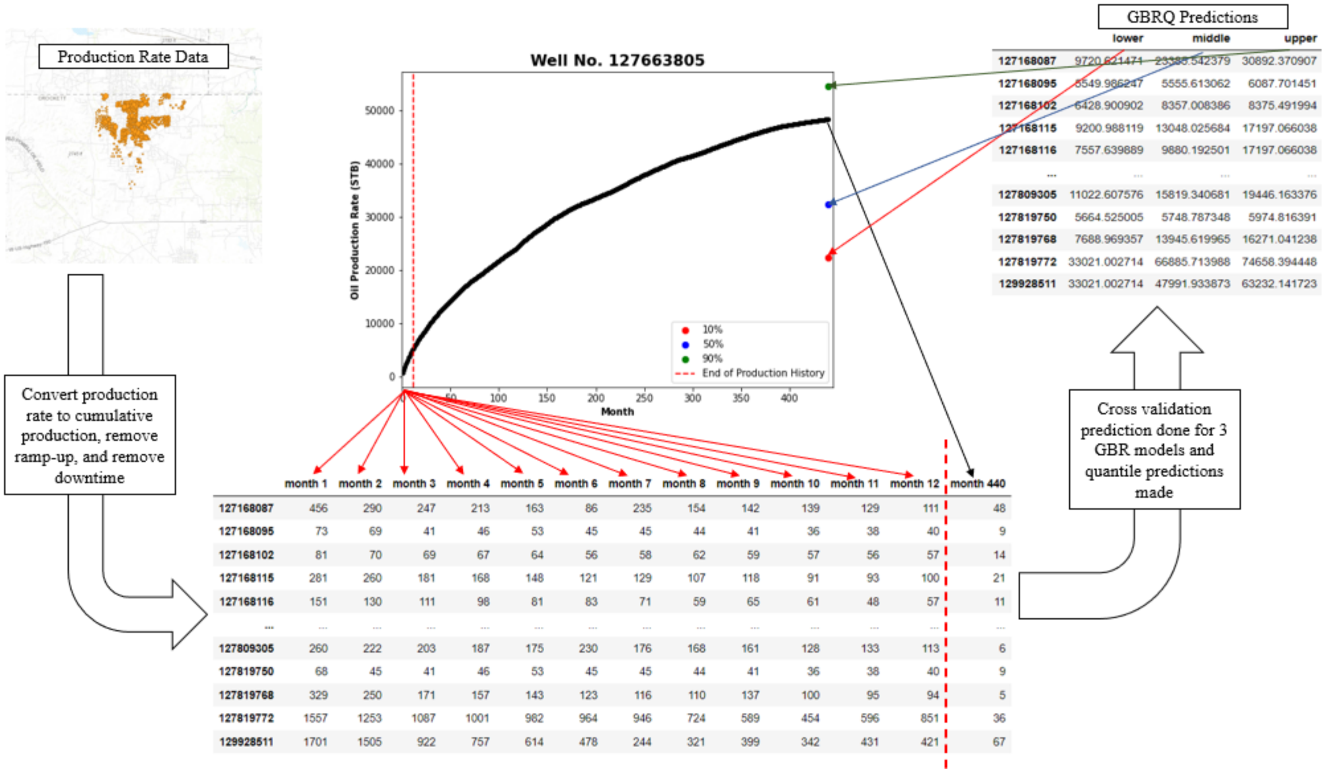


Fig. 3.7—Diagram of cross validation (Siakorn, Wikimedia Commons).


Production Rate Data
Convert production rate to cumulative production, remove ramp-up, and remove downtime
Well No. 127663805
Oil Production Rate (STB)
0
10000
20000
30000
40000
50000
0
50
100
150
200
250
300
350
400
Month
10%
50%
90%
End of Production History
Cross validation prediction done for 3 GBR models and quantile predictions made

GBRQ Predictions

	lower	middle	upper
127168087	9720.621471	23385.542379	30892.370907
127168095	5549.986247	5555.613062	6087.701451
127168102	6428.900902	8357.008386	8375.491994
127168115	9200.988119	13048.025684	17197.066038
127168116	7557.639889	9880.192501	17197.066038
...	...	...	...
127809305	11022.607576	15819.340681	19446.163376
127819750	5664.525005	5748.787348	5974.816391
127819768	7688.969357	13945.619965	16271.041238
127819772	33021.002714	66885.713988	74658.394448
129928511	33021.002714	47991.933873	63232.141723

	month 1	month 2	month 3	month 4	month 5	month 6	month 7	month 8	month 9	month 10	month 11	month 12	month 440
127168087	456	290	247	213	163	86	235	154	142	139	129	111	48
127168095	73	69	41	46	53	45	45	44	41	36	38	40	9
127168102	81	70	69	67	64	56	58	62	59	57	56	57	14
127168115	281	260	181	168	148	121	129	107	118	91	93	100	21
127168116	151	130	111	98	81	83	71	59	65	61	48	57	11
...	...	...	...	...	...	...	...	...	...	...	...	...	...
127809305	260	222	203	187	175	230	176	168	161	128	133	113	6
127819750	68	45	41	46	53	45	45	44	41	36	38	40	9
127819768	329	250	171	157	143	123	116	110	137	100	95	94	5
127819772	1557	1253	1087	1001	982	964	946	724	589	454	596	851	36
129928511	1701	1505	922	757	614	478	244	321	399	342	431	421	67

Fig. 3.8—Summary of GBRQ process.

Fig. 3.9 and **Fig. 3.10** show the P10, P50, and P90 predictions for cumulative production plotted against the actual values for all wells in the DJ and Midland datasets, respectively. The x axis represents the actual cumulative production for Month 21 while the y axis represents the predicted cumulative production for Month 21. The x and y axes range from 30,000 STB/D to 190,000 STB/D for the DJ dataset and from zero STB/D to 45,000 STB/D for the Midland dataset. The red datapoints represent predictions for the P10 model, the blue datapoints represent predictions for the P50 model, and the green datapoints represent predictions for the P90 model. The dotted black unit-slope line helps gauge accuracy. The closer to the line a P50 prediction is, the closer that prediction is to the actual value and thus the more accurate that prediction is (assuming the P50 is quantity used to measure accuracy). While the usual goal of training a deterministic machine learning model is to generate predictions as close to the actual values as possible, the goal of creating a probabilistically reliable model differs. In the case of **Fig. 3.9** and **Fig. 3.10**, a perfect probabilistically calibrated approach would mean the green dots are above the dotted black unit-slope line 90% of the time, the blue dots are above the line 50% of the time, and the red dots are above the line 10% of the time. The predictions for the DJ dataset (**Fig. 3.9**) appear to be less well behaved than the predictions for the Midland dataset (**Fig. 3.10**). The poor behavior in the DJ dataset predictions can be attributed to the fewer number of wells in the DJ dataset as compared with the Midland dataset, which is less information for the GBRQ model to use during training. Since all other wells were taken into account when generating predictions, the Midland dataset has significantly more information for fitting quantile predictions than the DJ dataset.

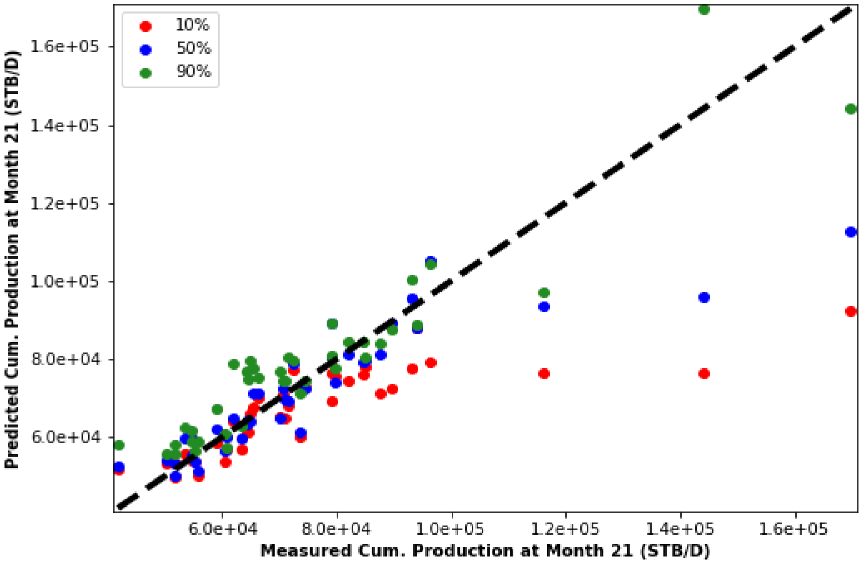


Fig. 3.9—GBRQ predictions for DJ dataset.

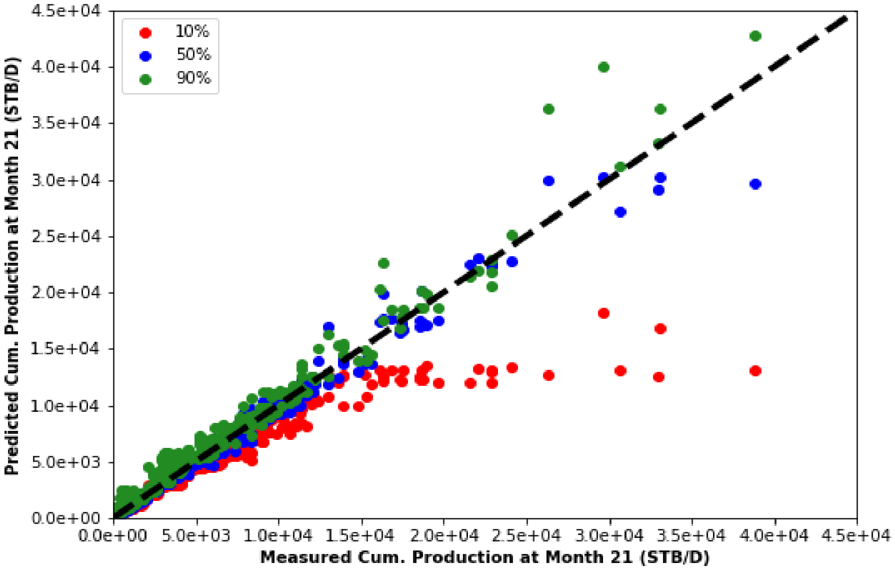


Fig. 3.10—GBRQ predictions for Midland dataset.

The reliability of these predictions was checked in Section 3.4 with calibration plots by plotting the proportion of predictions greater than actual vs the probability assigned.

3.3.2 Probabilistic Decline Curve Analysis (PDCA)

As mentioned before, analytical models that are typically applied for decline curve analysis are deterministic. Jochen and Spivey (1996) and Cheng et al. (2010) developed bootstrap methods to generate probabilistic production forecasts in single wells based on DCA models of existing production. When tested on a sample dataset of 100 oil and gas wells, it was found that the Cheng et al. method covered 80% of the true incremental production over the P90-P10 range for incremental production while the Jochen and Spivey method only covered 40% of the true incremental production over the same range of incremental production. In an ideal scenario, 80% of the true incremental production is covered over the P90-P10 range and thus the Cheng et al. Modified Bootstrap Method (MBM) was much closer to ideal coverage. However, the MBM developed by Cheng et al. requires a least-squares fit for each well at each month with each fit requiring multiple Newton iterations. This means each well requires three to five minutes to calculate probabilistic production forecasts. Gong et al. (2014) have since created a Bayesian method, which can quantify reserves uncertainty as reliably as the MBM by combining Markov-chain Monte Carlo (MCMC) simulation with Arps' DCA. Since this method is faster and just as reliable as the MBM, this was used as a benchmark for comparison to the GBRQ method rather than the MBM.

The MCMC method requires an iterative process to calculate a Markov chain that contains the desired posterior distribution. Since the posterior distribution of parameters is unknown, using the Metropolis-Hastings algorithm is needed to directly sample from a proposal distribution. This proposal distribution consists of parameters for Arps decline curves. Although other decline curve methods are readily available, the Arps method is the mostly widely used and understood method in the petroleum industry and has been in use for over 70 years, thus the Arps method was used. Specifically, a single segment hyperbolic decline model without a terminal exponential decline was used. This might be a problem for Midland dataset at very late times when wells have the possibility of interfering with one another. The bounds for the parameters in the prior distribution were chosen to be $0.01 < q_i < 1{,}000{,}000$, $0.0001 < D_i < 50$, and $0 < b < 2$, where q_i is in STB/D, D_i is in 1/years and b is dimensionless. These ranges were chosen to be wide enough so that any reasonable initial parameter values were included.

At each step s in the Markov Chain, a candidate $\theta_{proposal}$ is drawn from the proposal distribution. The probability that this candidate is accepted ($\theta_s = \theta_{proposal}$) is α and the probability of rejection ($\theta_s = \theta_{s-1}$) is $\alpha - 1$ where:

$$\alpha = \min[1, \frac{\pi(\theta_{proposal}|y)q(\theta_{s-1}|\theta_{proposal})}{\pi(\theta_{s-1}|y)q(\theta_{proposal}|\theta_{s-1})}] \quad \text{...}(10)$$

The MCMC method consists of the following steps:

0. Set $s = 1$ and $\ln(q_i)$, $\ln(D_i)$, and b equal to the least-squares best fit.
1. Generate a sample $\theta_{proposal}$ [$\ln(q_i)$, $\ln(D_i)$, and b] from the proposal distribution.
2. Calculate acceptance ratio by use of **Eq. 10**.
3. Generate a random number between zero and one.
4. If the random number is less than the acceptance ratio, accept $\theta_{proposal}$ (i.e., $\theta_s = \theta_{proposal}$). Otherwise, $\theta_s = \theta_{s-1}$.
5. $s = s+1$. If s is less than maximum chain length, go to step 1.

In terms of synthetic realizations, 1000 were used. Relative error decreased with increasing number of MCMC iterations and it was noticed that an acceptable level of error was reached with 1000 iterations. Further increases in iterations resulted in very small incremental gains with large increases in computational time. Other model inputs include 12 months of production history, logarithmic regression, a triangular distribution with noninformative priors for the q_i and D_i parameters and a triangular distribution of informative priors for b estimated from 197 Barnett Shale Gas wells. **Fig. 3.11** and **Fig. 3.12** show predicted cumulative production values at month 21 plotted against the actual cumulative production at Month 21. The red dots represent P10 predictions, the blue dots represent P50 predictions, and the green dots represent P90 predictions. The dotted black line represents a scenario in which the predicted values are the exact same as the actual values. The predictions seem to follow a unit-slope more closely than the GBRQ predictions. To quantify the difference in bias and calibration, calibration plots are shown and discussed in Section 3.4.

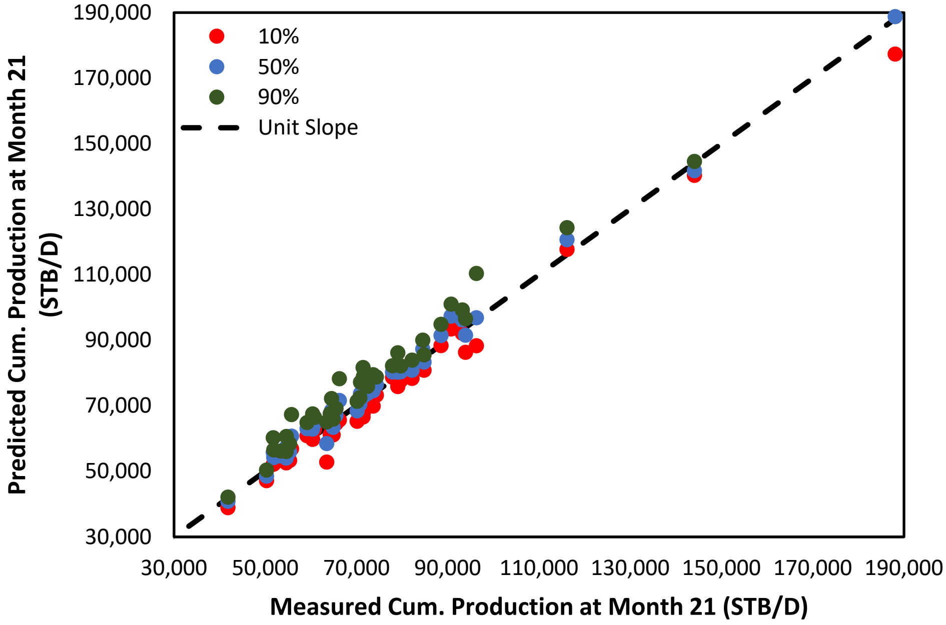


Fig. 3.11—PDCA predictions on DJ dataset.

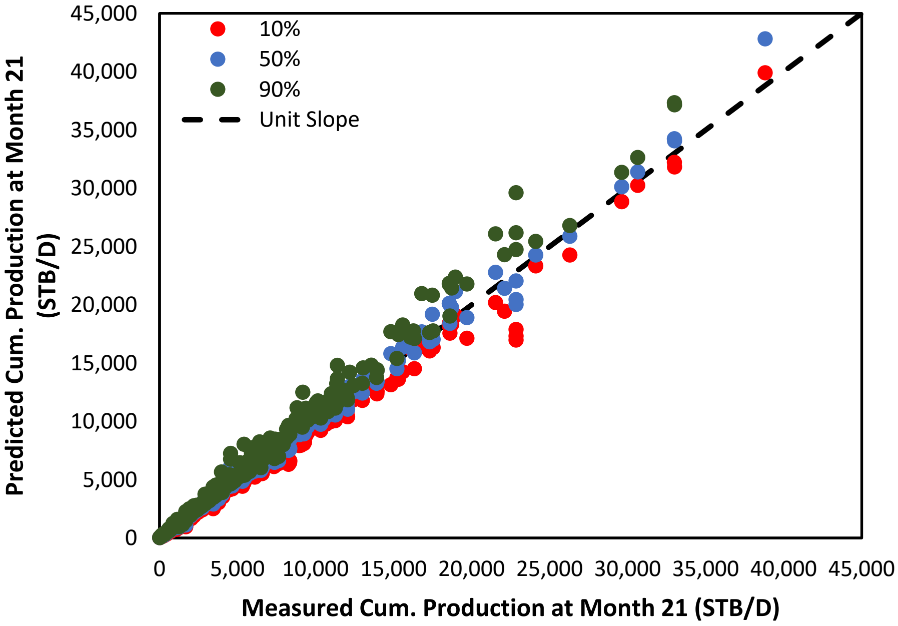


Fig. 3.12—PDCA predictions on Midland dataset.

3.4 Generate Calibration Plots

3.4.1 DJ dataset

The predictions for cumulative production at Month 21 for the DJ dataset in Section 3.3 were converted to calibration plots. To create these plots, the proportion of predictions greater than actual values is plotted against the probability assigned for 10%, 50%, and 90% probabilities. **Fig. 3.13** shows the calibration plot for the GBRQ predictions on the DJ dataset at Month 21. Upon inspection, the calibration plot is not well calibrated due to the obvious difference in slope and translation between the actual values and perfect calibration line. Since the slope was less than one, the predictions were overconfident. The overconfidence bias can be calculated as $CB_{OC} = 1 - 0.59 = 0.41$ from **Eq. 5**. There is a vertical translation downwards and the directional bias can be calculated as $DB_{OC} = 2*(.1698)/0.41 - 1 = -0.17$ using **Eq. 6**. This calculation quantifies the pessimism. Thus, the model is both moderately overconfident and slightly pessimistic for this dataset at 21 months.

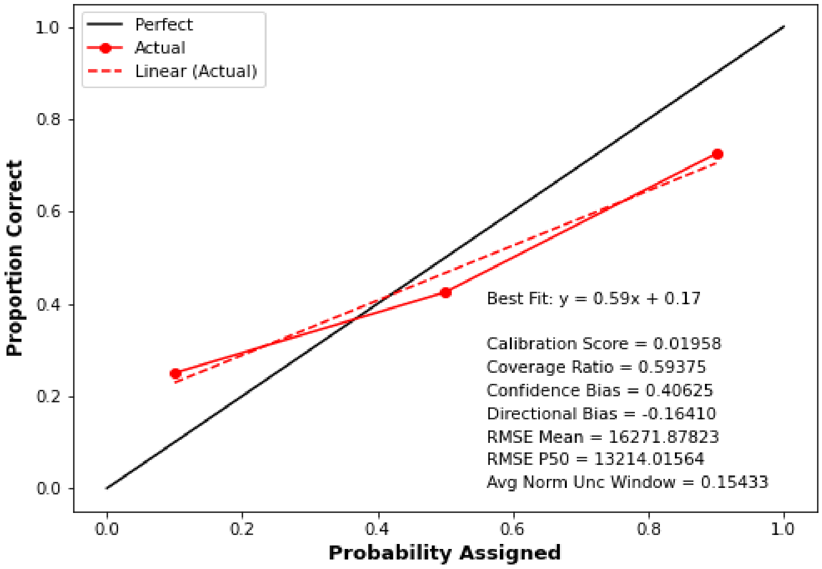


Fig. 3.13—GBRQ calibration plot (DJ dataset).

Fig. 3.14 shows the calibration plot for the PDCA predictions on the DJ dataset at Month 21. The PDCA CS score of 0.02354 was higher and thus worse than the GBRQ CS score of 0.01958. The calibration curve is translated downwards, and the slope of the best fit line is exactly one. The calibration score is higher than that of the GBRQ method. The downwards vertical translation indicates pessimism while the slope of one indicates no confidence bias, since $CB_{OC} = 1 - 1 = 0$. However, this would mean the directional bias is undefined since using **Eq. 6** would yield a divide-by-zero error. This would violate the assumptions of the confidence and directional bias equations laid out by Alarfaj and McVay (2020) and would thus potentially yield infinite pessimism. Thus, the PDCA model has no confidence bias but yields significant pessimism for this dataset.

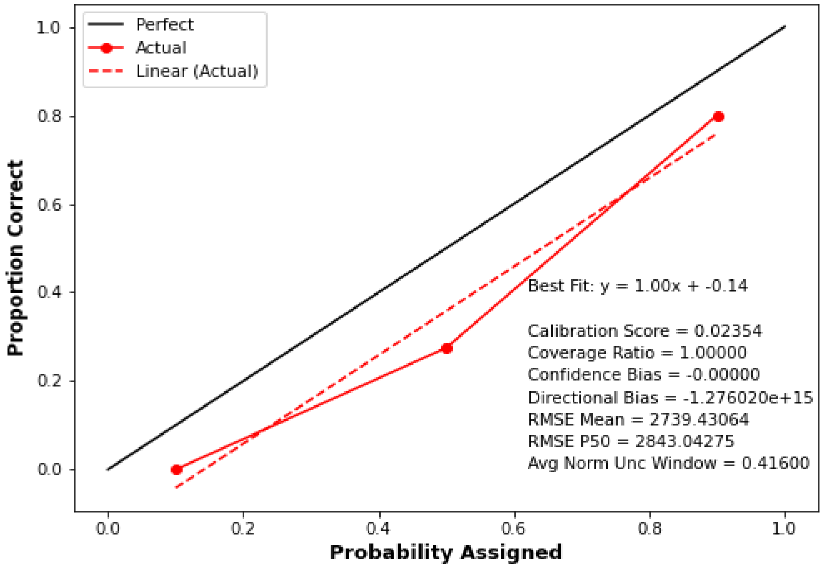


Fig. 3.14—PDCA calibration plot (DJ dataset).

It is obvious that both calibration plots are not well-calibrated both from initial inspection and from calculations of confidence and directional biases. Since Gong et al.'s PDCA implementation with 197 wells achieved very good probabilistic calibration and since machine learning models benefit from larger datasets, the same experiment was run on the larger dataset, the 438 well Midland dataset.

3.4.2 Midland dataset

Fig. 3.15 shows the calibration plot for the GBRQ predictions on the Midland dataset at Month 21. The calibration score of 0.00102 for the Midland dataset was a significant improvement over the calibration score of 0.01958 for the DJ dataset. It seems there is very slight overconfidence with almost non-existent vertical translation or directional bias. The overconfidence bias can be calculated as CB_{OC} as 0.10 from **Eq. 5** and $DB_{OC} =$ 0.128 from **Eq. 6**. Negligible optimism is visible in the calibration plot and confirmed by **Eq. 6**. Thus, the GBRQ model is slightly overconfident and slightly optimistic for this dataset at 21 months.

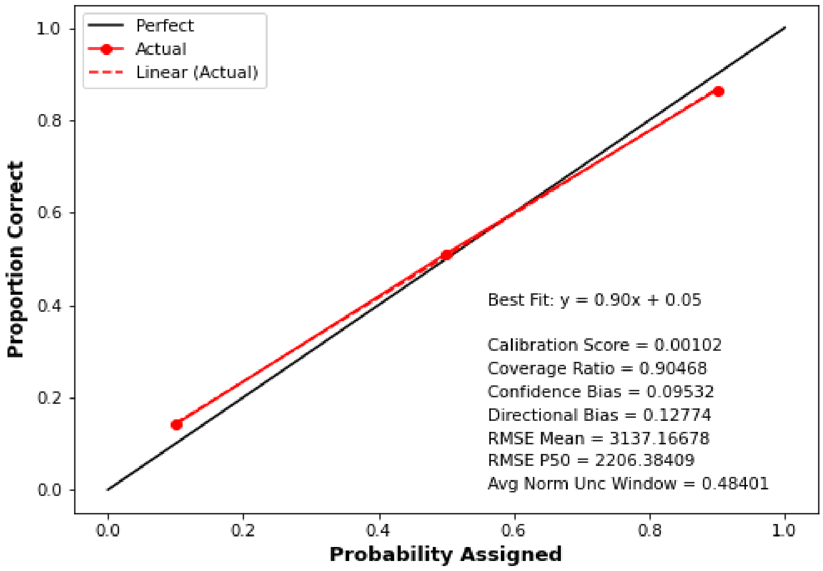


Fig. 3.15—GBRQ calibration plot (Midland dataset).

Fig. 3.16 shows the calibration plot for the PDCA predictions on the Midland dataset at Month 21. Upon inspection, this calibration plot is also much better calibrated than the predictions for the DJ dataset counterpart. The calibration score of 0.00255 is a significant improvement over the calibration score of 0.02354 for the DJ dataset. Although not clear what the confidence bias is by looking at the graph, slight optimism is indicated by a vertical shift upwards in the actual values. The overconfidence bias can be calculated as CB_{OC} = -0.02 from **Eq. 5**. The negative value indicates very little underconfidence and thus **Eq. 8** must be used to calculate DB_{UC} = 1 - 2*(.03)/-0.02 = 4. At extremely low values of confidence bias, assumptions were again violated for the equations in Alarfaj and McVay (2020); the directional bias value is outside the bounds of complete pessimism and complete optimism. From this analysis, the PDCA model is both slightly underconfident and optimistic for this dataset at 21 months.

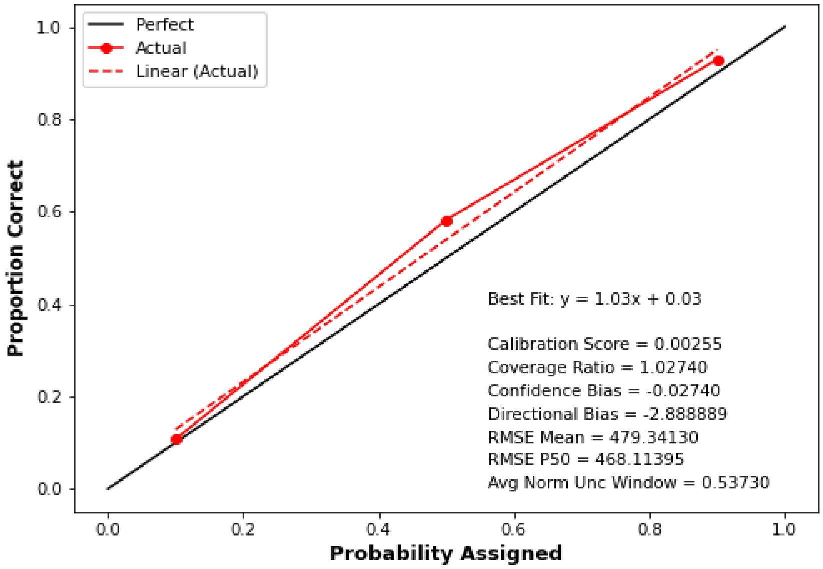


Fig. 3.16—PDCA calibration plot (Midland dataset).

3.4.3 Model Comparisons

The results of the model comparisons for the DJ dataset are summarized in **Table 1**.

	Metric	GBRQ	PDCA
Accuracy	RMSE Mean (STB)	16271.8782	2739.4306
	RMSE P50 (STB)	13214.0156	2843.0427
Probabilistic Reliability	Calibration Score	0.0195	0.0235
	Coverage Ratio	0.5937	1
	Confidence Bias	0.4062	0
	Directional Bias	-0.1641	-1.276E15
Uncertainty	Average Normalized Uncertainty Window	0.1543	0.416
Cost	Computational Time (min)	<1	55

Table 1—GBRQ vs PDCA statistics for DJ dataset at Month 21 (40 wells).

Recalling that the GBRQ calibration plot (**Fig. 3.13**) was moderately overconfident and slightly pessimistic while the PDCA calibration plot (**Fig. 3.14**) contained no confidence bias and was significantly pessimistic, there is now a quantitative basis for comparison. The calibration score for the GBRQ method is lower than that of the PDCA method. This means the GBRQ model is overall better probabilistically calibrated for the DJ dataset. The moderate overconfidence and slight pessimism of the GBRQ model was thus less biased overall when compared to the lack of confidence bias and significant pessimism from the PDCA model. Recall that the coverage ratio measures the fraction of the true distribution sampled in the predicted distribution. As a general rule for probabilistic models, a coverage ratio of one means 100% of the true distribution was sampled in the predicted distribution, which is ideal for a probabilistic model. This is the same as saying the model had no confidence bias. The coverage ratio was much less than one for the GBRQ method and exactly one for the PDCA method, meaning the PDCA method had superior coverage. The RMSE scores for the mean and median of the GBRQ were much higher than those of PDCA, indicating more accurate predictions of the true value by the mean and median of PDCA. The GBRQ method also predicted far less uncertainty than the PDCA method shown as a smaller average normalized uncertainty window. This is likely due to the inclusion of more information in the predictions as the GBRQ method considers production from other wells. If reliably assessing uncertainty, the uncertainty should decrease as more relevant information is included in the forecast. Lastly, the computational time was significantly less for the GBRQ than the PDCA. For the DJ dataset of 40 wells, the GBRQ method was overall better calibrated, less uncertain and faster while the PDCA method had better coverage and was more accurate.

The results of the calibration plot analysis for the Midland dataset are summarized in **Table 2**.

	Metric	**GBRQ**	**PDCA**
Accuracy	RMSE Mean (STB)	3137.1667	479.3413
	RMSE P50 (STB)	2206.3841	468.1139
Probabilistic Reliability	Calibration Score	0.001	0.0025
	Coverage Ratio	0.9046	1.0274
	Confidence Bias	0.0953	-0.0274
	Directional Bias	0.1277	2.8889
Uncertainty	Average Normalized Uncertainty Window	0.484	0.5373
Cost	Computational Time (min)	<1	324

Table 2—GBRQ vs PDCA statistics for Midland dataset at Month 21 (438 wells).

For the 438-well Midland dataset, calibration score for the GBRQ model was less than the score for the PDCA model. This means the GBRQ was overall better calibrated for this dataset as well. The GBRQ model was slightly overconfident and slightly optimistic while the PDCA model was slightly underconfident and significantly optimistic. The PDCA model again had a coverage ratio closer to one than the GBRQ model. The RMSE for the mean and median of the GBRQ model were significantly greater than the RMSE values for the PDCA model, meaning that PDCA again had more accurate predictions. The GBRQ also had a lower average normalized uncertainty window meaning less uncertainty was predicted. Lastly, the computational time was again significantly lower for GBRQ while PDCA increased significantly due to the large addition of data.

For the Midland dataset, the GBRQ method was better calibrated and faster while the PDCA method had more accurate predictions of the actual values. Since the two methods were close in calibration, the predicted uncertainty could be compared directly. The GBRQ method predicted less uncertainty than the PDCA method. Thus, the GBRQ and PDCA methods outperformed each other in the same metrics regardless dataset size when forecasting to 21 months of cumulative production given 12 months of forecast history.

3.5 Longer Forecast Times

The analysis was extended to other forecasted months. This was not possible for the DJ dataset because there was a lack of wells with adequate production history to extend the forecast beyond 24 months. Thus, analysis for other forecast target months was done using the Midland dataset.

The GBRQ method was used to generate probabilistic predictions for each month for each well in the Midland dataset (**Fig. 3.17**) for the same representative wells in **Fig. 3.3**. The PDCA method was also used to generate rate-time production profiles, which were converted to cumulative production profiles (**Fig. 3.17**) simply by summing the rate values of successive months. The thick black profile represents actual cumulative production, the green profile represents P90 predictions, the blue profile represents P50 predictions, and the red profile represents P10 predictions. Since the GBRQ predictions incorporated cumulative production for other wells at each month, those profiles were noisy. The PDCA predictions were based on decline curve models, so those profiles were smooth. To enhance readability, the GBRQ predictions were smoothed using a five-month-rolling average.

The same cases were visualized in **Fig. 3.18** on semi-log production-rate-vs-time plots. Since the GBRQ method predicted cumulative production, the production-rate profile was calculated by differentiating the cumulative-production profile; however, the initial rate calculations were far too noisy. Instead, the production rate was calculated using the five-month-rolling average applied first to the cumulative data and then a 15-month-rolling average applied to the calculated rate data. This mixture of smoothing was determined through trial and error and resulted in production rate profiles that were much more readable. Since the PDCA models predicted smooth decline curves, they were plotted directly. The legend for each sub plot was the same as in **Fig. 3.17** except for the inclusion of an additional actual production profile that was calculated and smoothed in the same way as the GBRQ profiles.

Well 127178493 in the top left of **Fig. 3.17** and **Fig. 3.18** was representative of wells with enhancement to production rate around the 260-month mark (**Fig. 3.18**). The GBRQ P90 profile overpredicted, the P50 profile fit closely, and the P10 profile underpredicted the actual production profile (**Fig. 3.17**). The PDCA profiles had similar behavior. The P50 profile of the GBRQ prediction did a better job of picking up the enhancement to production rate because the GBRQ method included cumulative production data from other wells with a similar enhancement to production.

Well 127188628 in the top right of **Fig. 3.17** and **Fig. 3.18** was representative of wells that experienced steep initial decline in production rate followed by an exponential decline, as shown by the linear decline of the actual production profile (**Fig. 3.18**). The PDCA P10 profile underpredicted, the P50 profile fit closely and subsequently underpredicted at later times, and the P90 profile overpredicted the actual production profile (**Fig. 3.17**). The GBRQ P10 profile stayed close to the actual profile while the P50 and P90 profiles overpredicted the actual production profile.

Well 127240426 in the middle left of **Fig. 3.17** and **Fig. 3.18** was representative of wells with significant oscillations in decline after Month 260 (**Fig. 3.18**). The GBRQ P10 profile underpredicted, the P50 profile fit closely, and the P90 profile overpredicted the actual production profile (**Fig. 3.17**). The GBRQ P10 and P50 profiles began to trend upward at later times. The PDCA P10 and P90 profiles behaved similarly, but the P50 profile overpredicted the actual production profile.

Well 127312421 in the middle right of **Fig. 3.17** and **Fig. 3.18** was representative of wells with heavy oscillations in production rate and very shallow decline (**Fig. 3.18**). The GBRQ P10 profile underpredicted, the P50 profile fit closely, and the P90 profile slightly overpredicted the actual production profile (**Fig. 3.17**). All three PDCA profiles vastly overpredicted by assuming an exponential decline.

Well 127363981 in the bottom left of **Fig. 3.17** and **Fig. 3.18** was representative of a single unique well with some form of production enhancement early in the well's life (**Fig. 3.18**). All three GBRQ profiles overpredicted before the enhancement (**Fig. 3.17**). After the enhancement, the P10 underpredicted, the P50 closely fit, and the P90 overpredicted the actual production profile. The PDCA model acted in the opposite manner with the P10, P50, and P90 profiles underpredicting, closely fitting, and overpredicting the actual production profile, respectively, before the enhancement. After the enhancement, all three PDCA profiles underpredicted. This is because the PDCA model was only fit to the first 12 months of production before the enhancement was performed. On the other hand, GBRQ incorporated late-time production from other wells and was able to pick up on the enhancement.

Finally, well 127663805 in the bottom right of **Fig. 3.17** and **Fig. 3.18** was representative of wells with no significant noise or initial steep decline (**Fig. 3.18**). These wells also had much larger cumulative production over well life (**Fig. 3.17**). The GBRQ P10 and P50 profiles underpredicted while the P90 profile slightly overpredicted when compared to actual production. The PDCA P10 and P50 profiles behaved similarly. However, the PDCA P90 profile slightly underpredicted when compared to actual production. Since these wells had larger production values, the predictions for the GBRQ method included cumulative production data for wells with less production and caused the P10 and P50 profiles to vastly underpredict the actual values. The informative prior distribution for *b* values in the PDCA method seemed to work well for the other representative cases but, it seems to have caused underpredictions in this case.

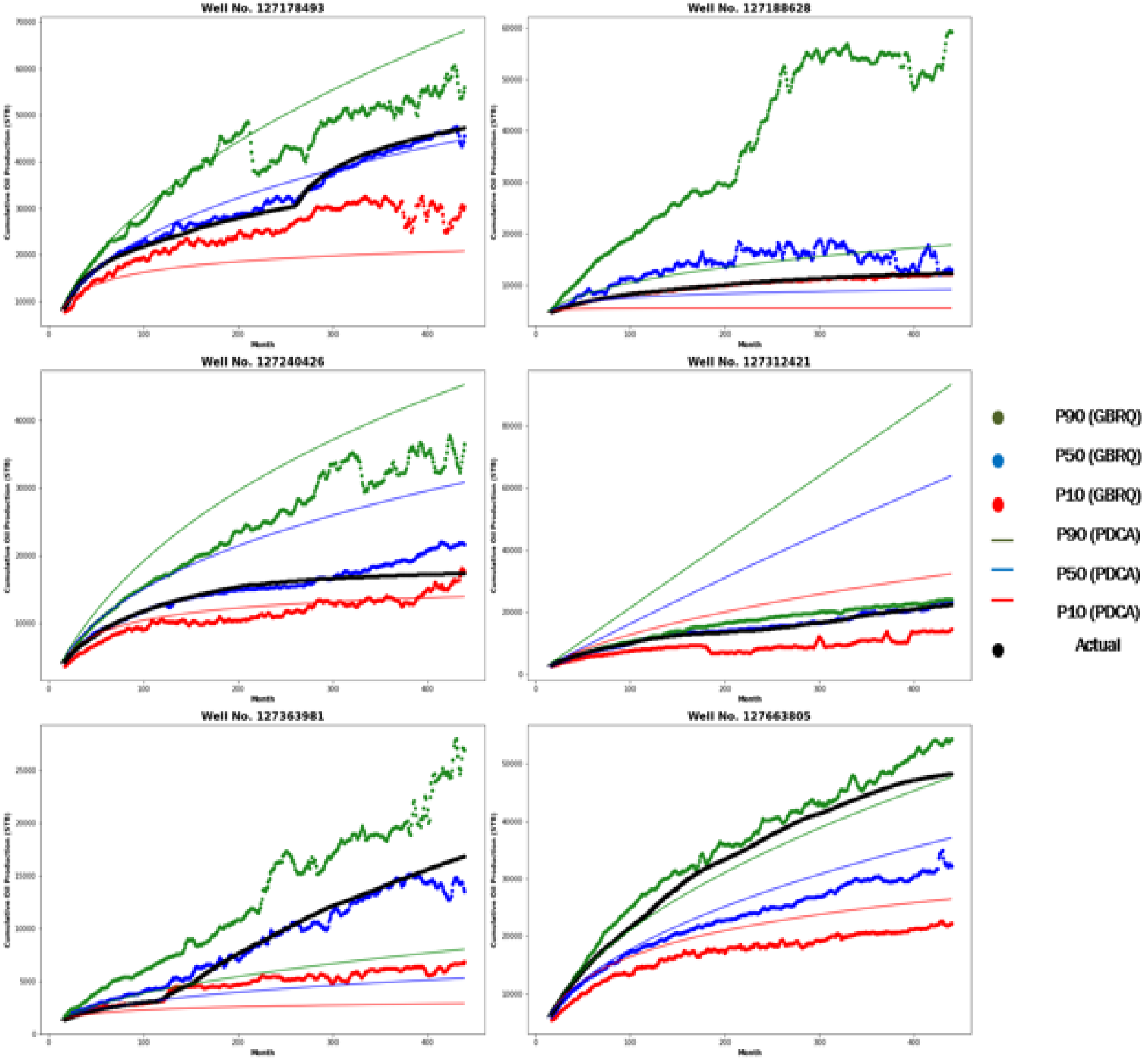


Fig. 3.17—Cumulative production predictions for representative cases of Midland dataset.

The semi log production rate profiles in **Fig. 3.18** show the noise in actual production rate better as well as the enhancements in production more clearly. The PDCA method did a solid job forecasting smoothed profiles when enhancements to production were not present. In cases with enhancements to production, the PDCA method underpredicted with all three profiles. The GBRQ method was able to pick up the enhancements to production due to the inclusion of cumulative production from other wells, but was significantly noisier than the PDCA method. While these figures quantify uncertainty, they do not quantify bias. Furthermore, the models were forecast to 440 months of production, which was a little more than half the 790 months of production history some of the other wells in the Midland dataset contained. The reason for not plotting all of the history is discussed in the next section.

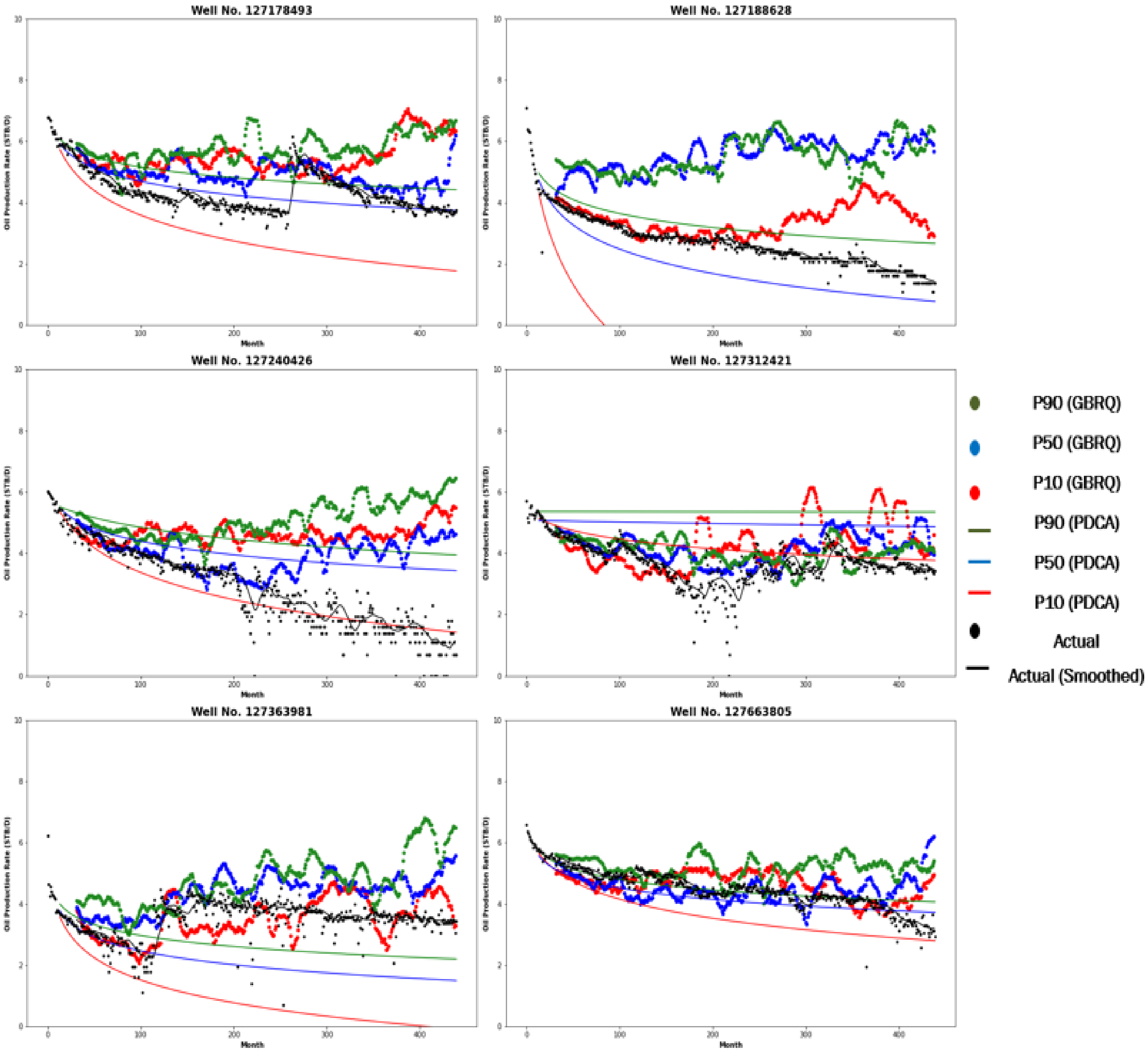


Fig. 3.18—Semi-log production rate predictions for representative cases of Midland dataset.

3.6 Forecast Length Limits of Midland dataset

As mentioned before, the well counts of the datasets were limited by available production history for a hindcast. The DJ dataset consisted of 90 out of 130 wells with fewer than 21 months of production history, so the dataset was limited to just 40 wells. The Midland dataset consisted of 10 out of 448 wells with fewer than 21 months of production history and was thus limited to 438 wells. As more production history was required to perform hindcasts at later months, there were fewer wells in the dataset with adequate production history. **Fig. 3.19** is a plot of well count vs months of production history available for the Midland dataset.

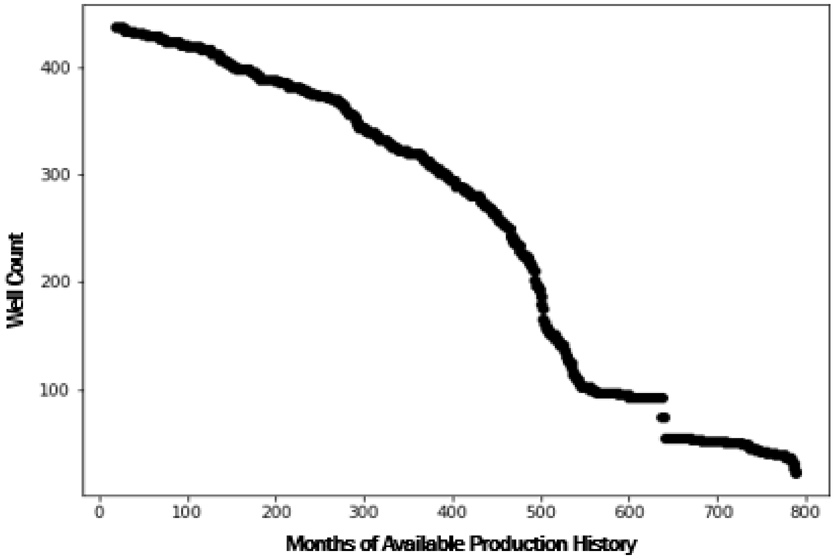


Fig. 3.19—Well count vs months of available production history (Midland dataset).

The dataset begins with 438 wells available at 21 months forecasted. As the number of months required for a hindcast increased, the number of wells with available production decreased more and more rapidly until eventually reaching a steep decline between 450 to 550 months and a sudden drop at 640 months. Note also that the number of wells decreased from 297 to 97 between 450 and 550 months. It was noted before that both the GBRQ and PDCA models performed much better for the Midland dataset than the DJ dataset, primarily due to the increase in available data in the Midland dataset. To further examine the relationship between more wells and better performance, predictions were generated with the GBRQ model at each month between 21 and 800, calibration plots were generated from those predictions, and a calibration score was calculated at each month (**Fig. 3.20**). The model was well calibrated as represented by low calibration scores until the rapid decline in well count beginning after Month 450. The calibration scores then increased rapidly and indicated the negative effect of a drop in well count on model calibration.

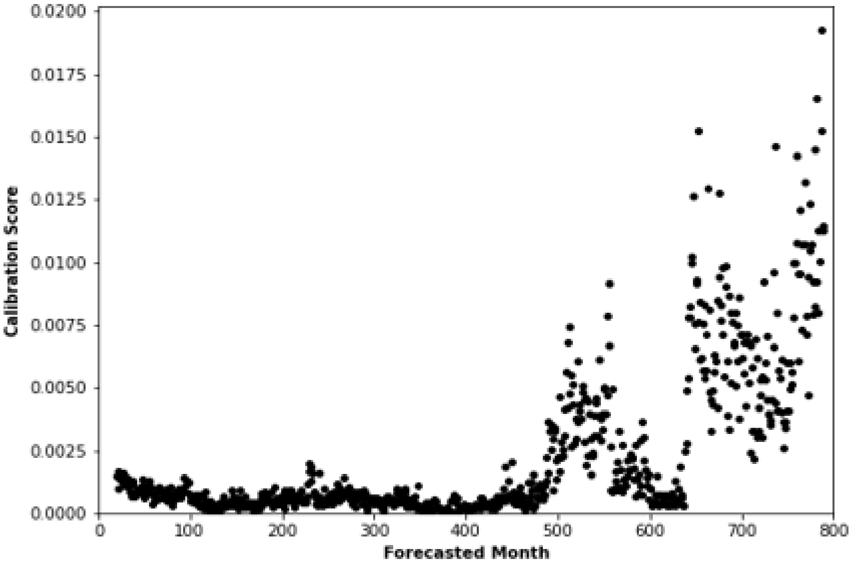


Fig. 3.20—Calibration score vs forecasted month (GBRQ Midland dataset).

The same calibration plots at each month were used to calculate coverage ratio (**Fig. 3.21**). The dotted blue line represents perfect coverage. Coverage ratio of the GBRQ model increased towards perfect coverage until dropping severely at the same time of 450 months.

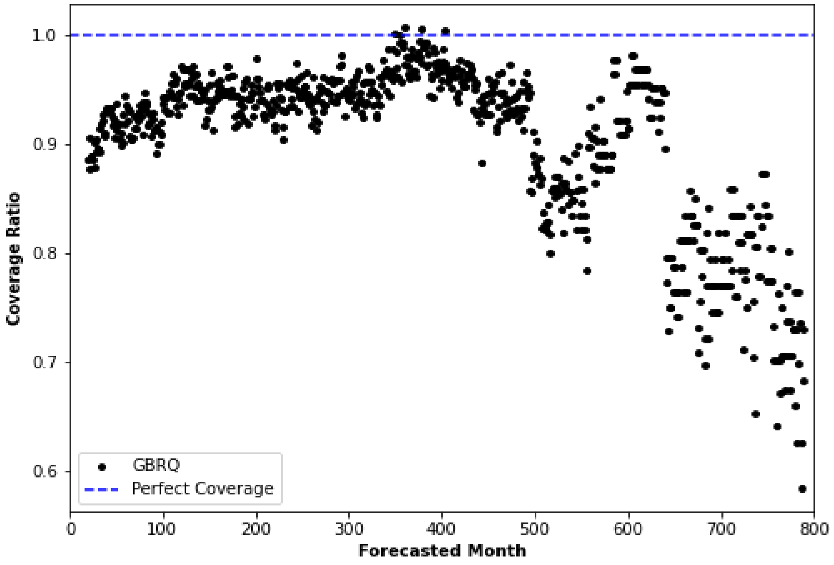


Fig. 3.21—Coverage ratio vs forecasted month (GBRQ Midland dataset).

The same calibration plots were used to calculate confidence bias (**Fig. 3.22**). Confidence bias, which has an inverse relationship with coverage ratio, did not change significantly with greater forecasted months until experiencing significant noise as well count plummeted.

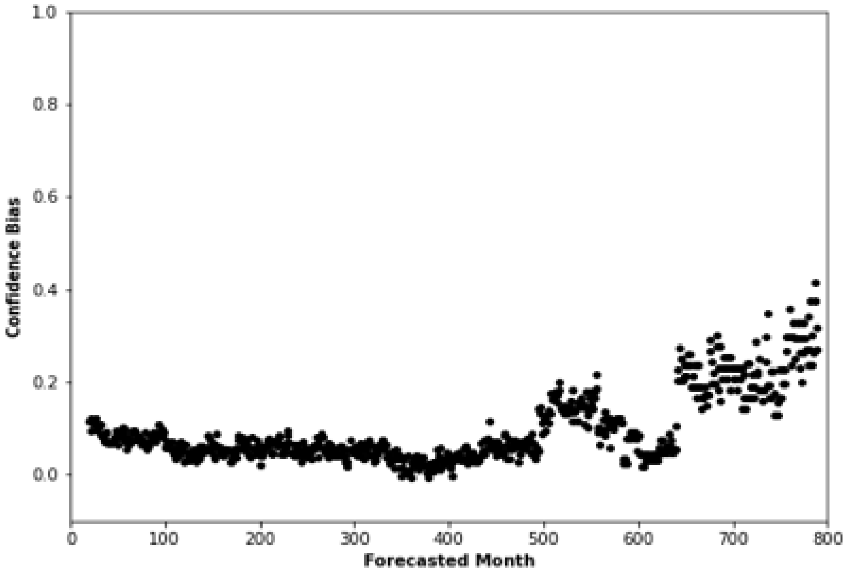


Fig. 3.22—Confidence bias vs forecasted month (GBRQ Midland dataset).

Fig. 3.23 shows the directional bias variation with forecasted month and was created in the same way as previous plots. The blue line represents complete optimism while the red line represents complete pessimism. The black dots represent the directional bias values. Other than unusual noise at 370 and 600 months, there was no trend in directional bias. In other words, the drop in well count at later times did not seem to influence directional bias.

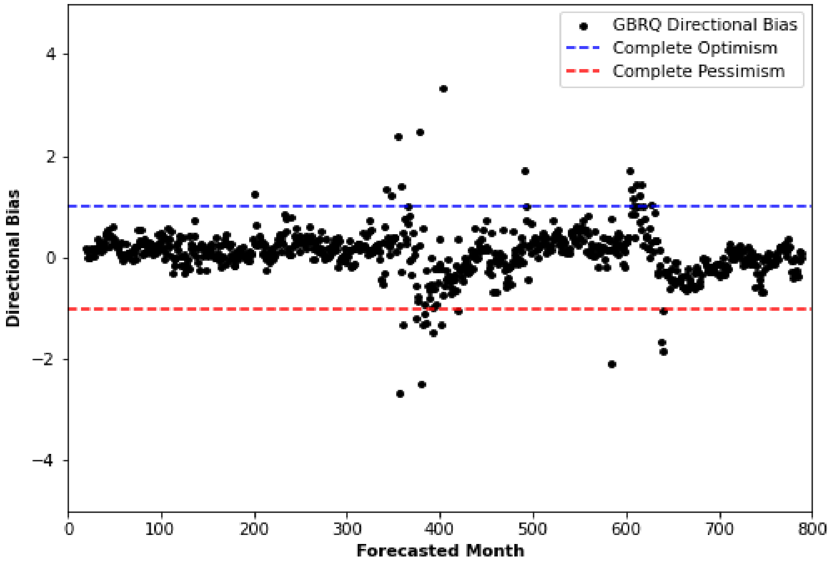


Fig. 3.23—Directional bias vs forecasted month (GBRQ Midland dataset).

Fig. 3.24 shows the RMSE between the mean of the predictions and actual values in black, as well as the RMSE between the median of the predictions and actual values in red. The mean of the predictions was calculated using Swanson's rule as previously mentioned. The plot shows that the median was a more accurate predictor of production than the mean and that both RMSE values increased steadily with longer forecast period until experiencing a sudden drop followed by noise when reaching the well count drop at 450-500 months.

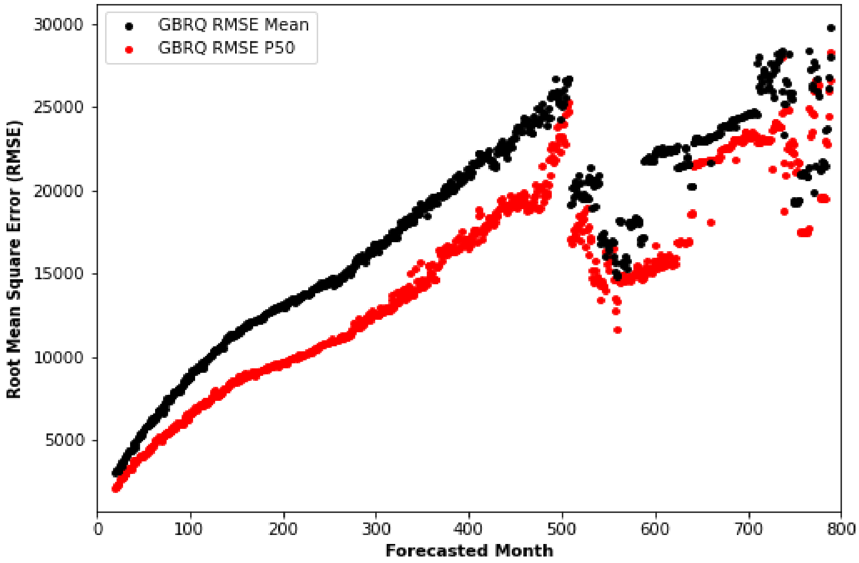


Fig. 3.24—RMSE vs forecasted month (GBRQ Midland dataset).

Fig. 3.25 shows the average normalized uncertainty window as a function of the month forecasted to. This window is a measure of predicted uncertainty, so the smaller the window the less uncertainty the model predicts, which is preferred. Values increased steeply then stayed around 0.8 up to 450 months before dropping quickly. The values then declined rapidly starting at 550 months.

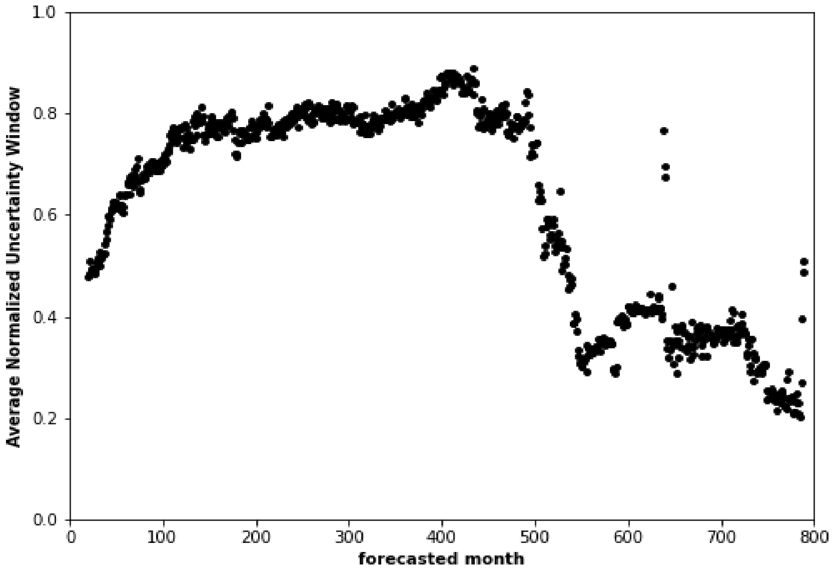


Fig. 3.25—Uncertainty window vs forecasted month (GBRQ Midland dataset).

The significant drop in well count after Month 440 caused irregularities in nearly all the measures of GBRQ predictions. Thus, comparison between the GBRQ and PDCA methods was limited to the interval from zero to 440 months.

3.7 GBRQ vs PDCA (Midland dataset)

The GBRQ and PDCA models were compared over a forecast time interval of 21 to 440 months. The PDCA model generated predictions for Months 21, 24, 36, 60, 120, 240, 360, and 440 using 12 months of production history. This is because the PDCA method took hours as opposed to seconds to run each case and this way important trends in statistics for the PDCA method could still be observed without computational time being an inhibitor. The GBRQ method generated predictions for each month between Month 21 (438 wells) and Month 440 (297 wells). **Fig. 3.26** to **Fig. 3.30** show metric comparisons vs forecast month, tables of statistics at each key time are presented in **APPENDIX A**, and calibration plots for the PDCA and GBRQ runs at each key time are presented in **APPENDIX B** and **APPENDIX C,** respectively.

Fig. 3.26 shows calibration score for both methods. The GBRQ method is shown in black and the PDCA method is shown in red. The PDCA method has a higher calibration score over the entire forecast period and is thus less well calibrated overall compared to the GBRQ method.

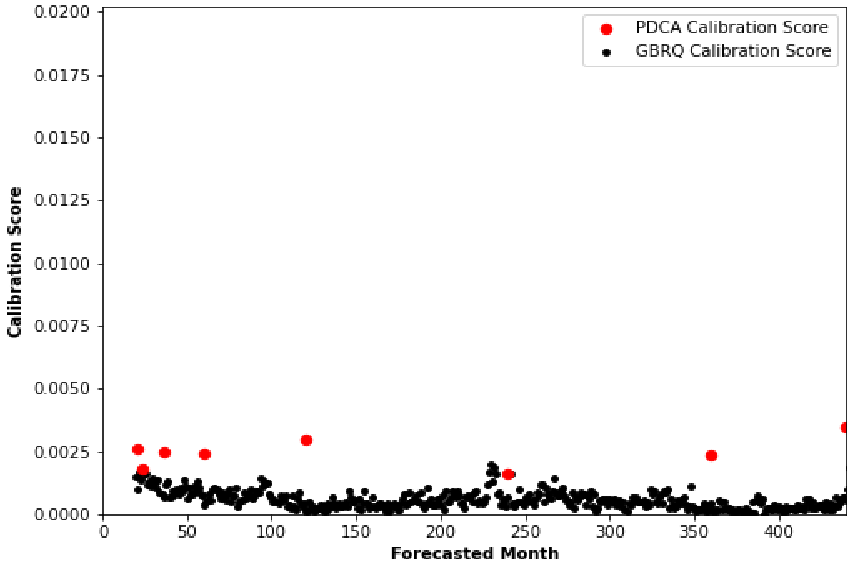


Fig. 3.26—Calibration score vs forecasted month (GBRQ vs PDCA).

Fig. 3.27 shows coverage ratio for both methods. The associated colors for GBRQ and PDCA are the same as in **Fig. 3.26** with the addition of a blue dotted line to show perfect coverage. The PDCA method had close to a perfect coverage ratio for very early forecast times before dropping and staying below the GBRQ method until Month 440. These initial perfect coverage ratios were consistent with the analysis on the DJ dataset, but the more complete analysis shows that the GBRQ method had better coverage than the PDCA method for forecast times after Month 24.

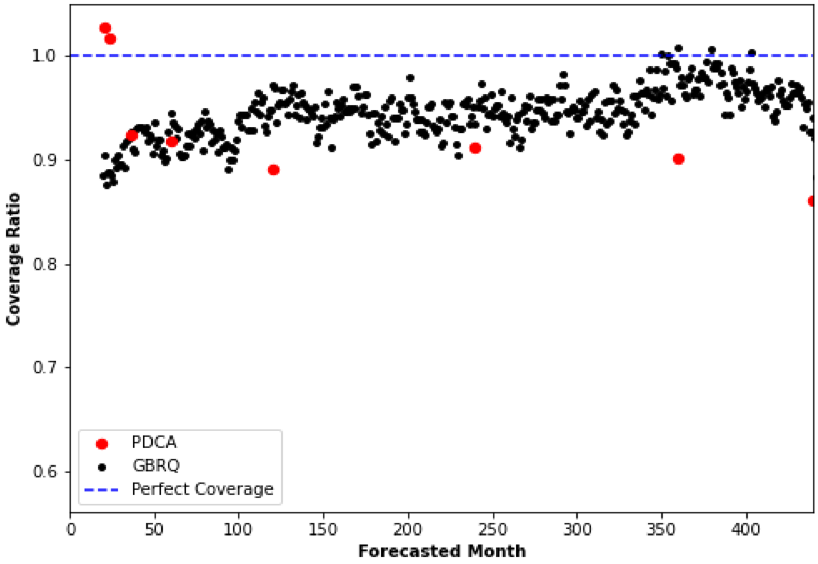


Fig. 3.27—Coverage ratio vs forecasted month (GBRQ vs PDCA).

Fig. 3.28 shows confidence bias for both methods. Confidence bias has an inverse relationship with coverage ratio. The PDCA method has a similar inverse trend of very low confidence bias at early times followed by a larger confidence bias than the GBRQ method until Month 440. This matched the behavior in **Fig. 3.27**.

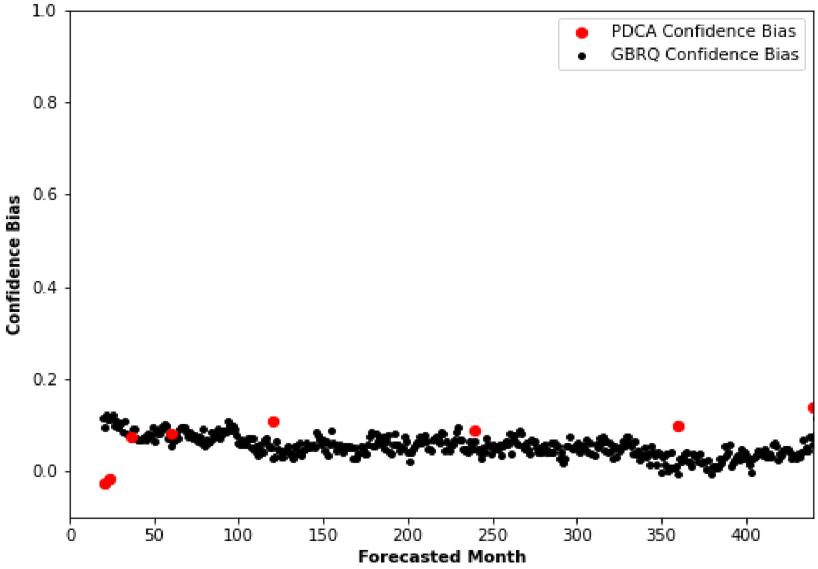


Fig. 3.28—Confidence bias vs forecasted month (GBRQ vs PDCA).

Fig. 3.29 shows directional bias for both methods. The colors for the GBRQ method are the same as in the figures above with the addition of a dotted blue line to represent complete optimism and a dotted red line to represent complete pessimism. The GBRQ method had no trend in directional bias and stayed close to zero until experiencing noise around Month 370. The PDCA method experienced very high optimism at early times (due to near-zero confidence bias) and trended downwards towards no directional bias at late times. The PDCA method had greater directional bias than the GBRQ method for almost the entire duration of the forecast period.

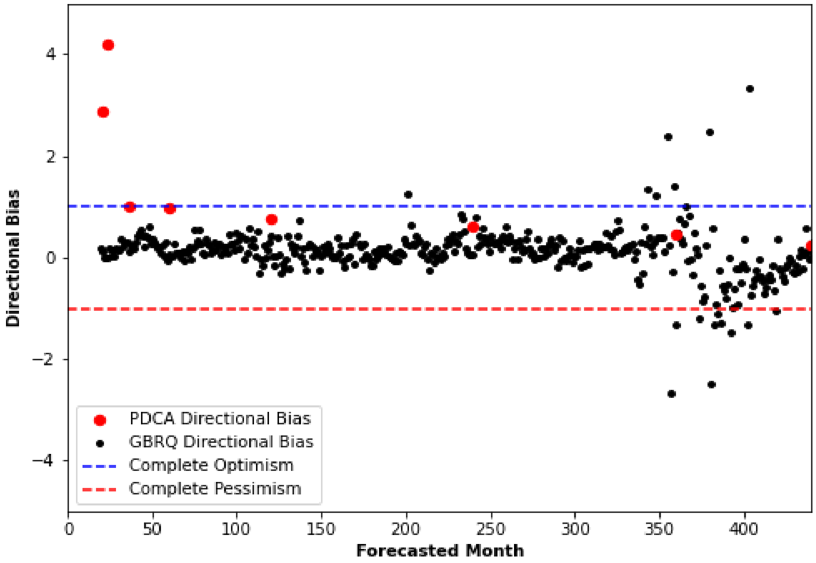


Fig. 3.29—Directional bias vs forecasted month (GBRQ vs PDCA).

Fig. 3.30 shows the RMSE for the mean and median for both methods. The RMSE for the mean is shown as small black circles for the GBRQ method and large red circles for the PDCA method while the RMSE for the median is shown as smaller black triangles for the GBRQ method and large red triangles for the PDCA method. The RMSE for the median was more accurate for the GBRQ method with both errors trending upwards steadily over the entire forecast period. The RMSE's for the PDCA method were indistinguishable until Month 240 when the median became more accurate. At Month 360 the mean becomes more accurate and at Month 440 the median again becomes more accurate. Thus, neither the mean nor median can be deemed superior to the other for the PDCA method. The PDCA method is more accurate than the GBRQ method prior to Month 120 while the GBRQ method is much more accurate after Month 120. The GBRQ method is more accurate after Month 120 likely due to the GBRQ method's ability to incorporate long-term production of other wells at later forecast times. The PDCA method does not have this ability.

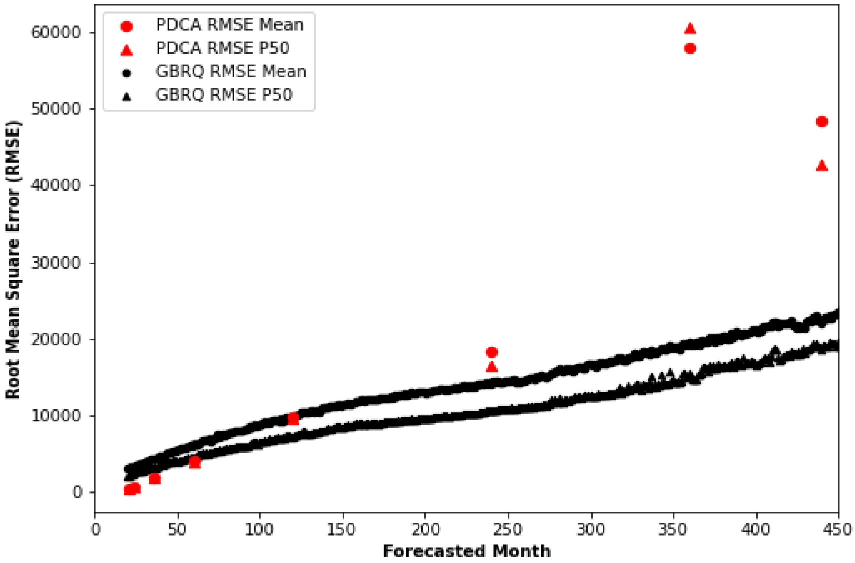


Fig. 3.30—RMSE (P50 and mean) vs forecasted month (GBRQ vs PDCA).

Finally, **Fig. 3.31** shows the average normalized uncertainty window for both methods as a function of forecast length. The PDCA uncertainty window increased rapidly at first then increased at a shallower slope at later times. The GBRQ method also increased rapidly at first, but increased only slightly after about 120 months. The GBRQ method predicted less uncertainty throughout the forecast period and significantly less uncertainty at large forecast times because of the inclusion of cumulative production information from other wells. The GBRQ method incorporates more information than the PDCA method. If you are reliably assessing uncertainty, then as you add more relevant information, the uncertainty should decrease.

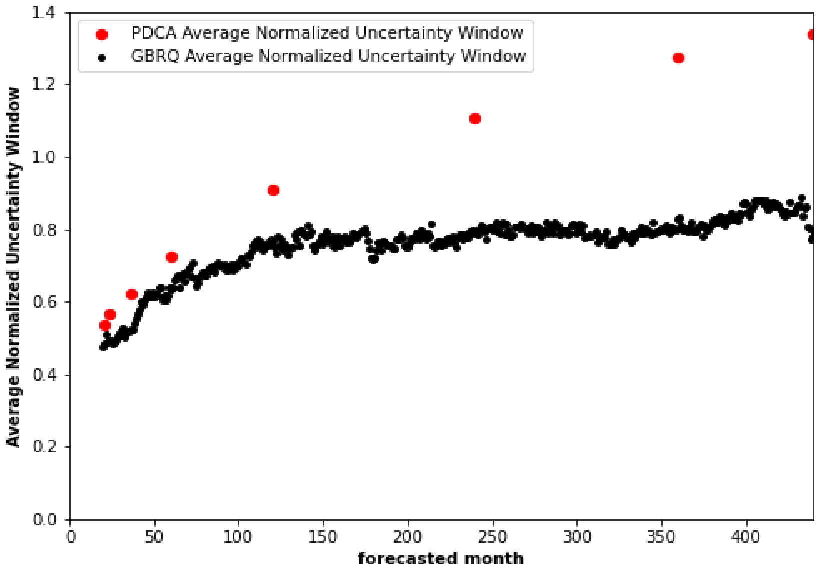


Fig. 3.31—Average normalized uncertainty window vs forecasted month (GBRQ vs PDCA).

In summary, predictions with the GBRQ method were overall better calibrated (with correspondingly better calibration scores and coverage ratios as well as lower confidence and directional biases) over the entire 21-440 month forecast period. The GBRQ method also had greater accuracy for both the mean and median after about Month 120 and had lower predicted uncertainty over the entire forecast period than the PDCA method. The GBRQ method generated the forecasts for all wells in the forecast period in 45 min while the PDCA method took over three and a half hours, plus additional post-processing time to calculate metrics. The GBRQ method was also limited by available production history. It was only able to forecast to a month in which other wells already had production history. The decline models in the PDCA method do not have this limitation. Lastly, the inclusion of wells with enhancements to production, initial steep declines, and oscillations in production made the GBRQ profiles much noisier than the PDCA profiles.

The GBRQ method could be useful to reserves estimators who would like to make fast and reliable probabilistic forecasts. Reserves auditors can also use this method to generate a probabilistic forecast of production to compare with an auditee's forecast. Lastly, investors and banks can create probabilistic production forecasts for asset acquisition and divestiture evaluation.

4. LIMITATIONS AND FUTURE WORK

4.1 Limitations

The inputs for both models were limited to 12 months of production history in the DJ and Midland datasets. The GBRQ method was also limited to default hyperparameters for the initial weak learner and the PDCA method was limited to using Arp's decline model. The earliest forecast was to Month 21 and the latest was Month 440 for the Midland dataset because of the well count drop. Computational time was also a constraint for the PDCA method, which prevented a full realization of analysis at each month like the GBRQ method. Lastly, the time used was elapsed time from initial production, but this limited the ability of the GBRQ model to identify possible field-wide effects at the same date on multiple wells.

4.2 Future Work

Future research should include using production histories other than 12 months for predicting future performance. This has already been performed for PDCA (Gong et al. 2014) in which improvements in accuracy and probabilistic reliability were noted with increasing length of history. Thus, this should be done with the GBRQ method.

The GBRQ method can also include data other than production data in its forecast of future cumulative production. The additional data could include completion data, as well as geologic and petrophysical properties. As mentioned before, actual time should also be included as an input.

The effect of changing hyperparameters on probabilistic reliability and other metrics would also be a useful future study.

5. CONCLUSIONS

For the 438-well, conventional-oil-well dataset in the Midland basin, the GBRQ ML method was clearly superior to the PDCA method. The GBRQ method was better calibrated than the PDCA method. Calibration scores were lower, coverage ratio was superior, overconfidence was lower, and optimism was lower. The PDCA method made more accurate predictions for the first half of forecast length (fewer than 120 months), but the GBRQ method made more accurate predictions for the second half (greater than 120 months). The GBRQ method also predicted less uncertainty than the PDCA method. The GBRQ method created full forecasts in significantly less time than the PDCA method. The GBRQ method was also able to perform a probabilistic production forecast for a single month in seconds, which was an option the PDCA method did not have.

The GBRQ method had very noisy forecasts when compared to the PDCA method. This was because the PDCA method utilizes decline curve models while the GBRQ method utilizes a separate ML model for each month of predicted future cumulative production. Because the GBRQ method incorporates cumulative production from other wells at the future predicted month of interest, it (1) is able to better predict long-term production, including changes in future production trends, than the DCA-based PDCA method, and (2) predicts lower uncertainty than the PDCA method because it is incorporating more information into the prediction than the PDCA method, which considers only the historical production for each well in forecasting production. However, this feature of the GBRQ—incorporating cumulative production from other wells at the future predicted month of interest—is also a disadvantage in that it cannot forecast past the date for which production data is available from analogous wells, while the DCA-based method can forecast as far into the future as desired.

NOMENCLATURE

Acronyms

CDF	Cumulative Distribution Function
DCA	Decline Curve Analysis
E&P	Exploration and Production
EN	Elastic Net
GBR	Gradient Boosting Regressor
GBRQ	Gradient Boosting Regressor with Quantiles
MAE	Mean Absolute Error
MBE	Mean Bias Error
MBM	Modified Bootstrap Method
MCMC	Markov Chain Monte Carlo
ML	Machine Learning
MM	Multi-Model
MSE	Mean Squared Error
PDCA	Probabilistic Decline Curve Analysis

Symbols - Units

b	Hyperbolic Exponent (Loss Ratio), Dimensionless
CS	Calibration Score
c_t	Proportion of items greater than the actual for the probability assigned
D_i	Nominal Decline Rate at time 0, 1/t
D_{lim}	Terminal Nominal Decline Rate when $t \geq t_{\text{lim}}$, 1/t
$L(y_i,y_i^p)$	Loss Function
n_t	Number of times the response was used
N	Total number of responses
Q	Quantile
q	Instantaneous Production Rate at time t, STB/D
q_i	Instantaneous Production Rate at time 0, STB/D
q_{lim}	Instantaneous Production Rate at time t_{lim}, STB/D
q_{max}	Maximum Instantaneous Production Rate, STB/D
r_t	Probability assigned in calibration score
T	Total number of response categories used
t	Time, months

t_{lim}	Time when Nominal Decline decreased from D_i to D_{lim}, months
x	Variable for Normalized Uncertainty Window
x_{std}	Standardized Variable for Normalized Uncertainty Window
y_i	Actual Value in Loss Function
y_i^p	Predicted Value in Loss Function

Greek Variables

μ	Mean
σ	Standard Deviation

Subscripts

i	Initial
*std*Standardized	

REFERENCES

Brashear, J.P., Becker, A.B. and Faulder, D.D. 2001. Where Have All the Profits Gone? *J Pet Technol* **53** (6): 20-73. DOI: https://doi.org/10.2118/73141-JPT.

Capen, E.C. 1976. The Difficulty of Assessing Uncertainty (Includes Associated Papers 6422 and 6423 and 6424 and 6425). *J Pet Technol* **28** (8): 843-850. DOI: https://doi.org/10.2118/5579-PA.

Cheng Y., e.a. 2010. Practical Application of a Probabilistic Approach to Estimate Reserves Using Production Decline Data. *SPE Econ & Mgmt* **2** (1): 19-31. DOI: http://dx.doi.org/10.2118/95974-PA.

Elliott, R.and Matthews, C.M. 2019. As Shale Wells Age, Gap between Forecasts and Performance Grows. *The Wall Street Journal*. Accessed 15 January 2022. DOI: https://www.wsj.com/articles/as-shale-wells-age-gap-between-forecasts-and-performance-grows-11577631601.

Fulford, D.S., Bowie, B., Berry, M.E., et al. 2016. Machine Learning as a Reliable Technology for Evaluating Time/Rate Performance of Unconventional Wells. *SPE Economics & Management* **8** (01): 23-39. DOI: 10.2118/174784-pa.

Gong, X., et al. 2014. Bayesian Probabilistic Decline-Curve Analysis Reliably Quantifies Uncertainty in Shale-Well-Production Forecasts. *SPE J* **19** (6): 1047-1057. DOI: https://doi.org/10.2118/147588-PA.

Harris, C. 2014. Potential Pitfalls in Exploration and Production Applications of Machine Learning. In *SPE Western North American and Rocky Mountain Joint Meeting*, All Days. DOI: https://doi.org/10.2118/169523-MS.

Jochen, V.A. and Spivey, J.P. 1996. Probabilistic Reserves Estimation Using Decline Curve Analysis with the Bootstrap Method. Paper presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado. DOI: http://dx.doi.org/10.2118/36633-MS.

Kuzma, H.A., Arora, N.S., and Farid, K. 2014. Generative Models for Production Forecasting in Unconventional Oil and Gas Plays. In *SPE/AAPG/SEG Unconventional Resources Technology Conference*, All Days. DOI: https://doi.org/10.15530/URTEC-2014-1928595.

Li, B., Billiter, T.C., and Tokar, T. 2021. Rescaling Method for Improved Machine-Learning Decline Curve Analysis for Unconventional Reservoirs. *SPE J* **26** (6): 1759-1772. DOI: https://doi.org/10.2118/205349-PA.

Li, Y., and Han, Y. 2017. Decline Curve Analysis for Production Forecasting Based on Machine Learning. Paper presented at the SPE Symposium: Production Enhancement and Cost Optimisation, Kuala Lumpur, Malaysia. DOI: https://doi.org/10.2118/189205-MS.

Lichtenstein, S. and Fischhoff, B. 1977. Do Those Who Know More Also Know More About How Much They Know? *Organizational Behavior & Human Performance* **20** (02): 159-183. DOI: https://doi.org/10.1016/0030-5073(77)90001-0.

McVay, D.A. and Dossary, M.N. 2014. The Value of Assessing Uncertainty. *SPE Economics & Management* **6** (02): 100-110. DOI: 10.2118/160189-pa

APPENDIX A – STATISTICS FROM GBRQ VS PDCA RUNS

	Metric	GBRQ	PDCA
Accuracy	RMSE Mean (STB)	3174.472	633.103
	RMSE P50 (STB)	2432.988	650.839
Probabilistic Reliability	Calibration Score	0.0015	0.0017
	Coverage Ratio	0.888	1.016
	Confidence Bias	0.112	-0.016
	Directional Bias	-0.014	4.19
Uncertainty	Average Normalized Uncertainty Window	0.4929	0.5677
Cost	Computational Time (min)	<1	324

Table A.1—GBRQ vs PDCA statistics - 24 months.

	Metric	GBRQ	PDCA
Accuracy	RMSE Mean (STB)	4351.587	1872.888
	RMSE P50 (STB)	3267.756	1823.169
Probabilistic Reliability	Calibration Score	0.0007	0.0025
	Coverage Ratio	0.927	0.924
	Confidence Bias	0.073	0.076
	Directional Bias	0.115	1
Uncertainty	Average Normalized Uncertainty Window	0.5227	0.6242
Cost	Computational Time (min)	<1	222

Table A.2—GBRQ vs PDCA statistics - 36 months.

	Metric	GBRQ	PDCA
Accuracy	RMSE Mean (STB)	6296.343	4015.468
	RMSE P50 (STB)	4661.871	4134.244
Probabilistic Reliability	Calibration Score	0.0003	0.0024
	Coverage Ratio	0.944	0.918
	Confidence Bias	0.056	0.082
	Directional Bias	0.069	0.993
Uncertainty	Average Normalized Uncertainty Window	0.6373	0.7242
Cost	Computational Time (min)	<1	223

Table A.3—GBRQ vs PDCA statistics - 60 months.

	Metric	GBRQ	PDCA
Accuracy	RMSE Mean (STB)	9883.658	9619.81
	RMSE P50 (STB)	7285.771	9735.615
Probabilistic Reliability	Calibration Score	0.0001	0.0029
	Coverage Ratio	0.972	0.89
	Confidence Bias	0.029	0.109
	Directional Bias	0.417	0.75
Uncertainty	Average Normalized Uncertainty Window	0.7683	0.9096
Cost	Computational Time (min)	<1	313

Table A.4—GBRQ vs PDCA statistics - 120 months.

	Metric	GBRQ	PDCA
Accuracy	RMSE Mean (STB)	14151.951	16544.374
	RMSE P50 (STB)	10621.609	18425.043
Probabilistic Reliability	Calibration Score	0.0007	0.0016
	Coverage Ratio	0.934	0.911
	Confidence Bias	0.066	0.089
	Directional Bias	0.055	0.617
Uncertainty	Average Normalized Uncertainty Window	0.7877	1.1093
Cost	Computational Time (min)	<1	228

Table A.5—GBRQ vs PDCA statistics - 240 months.

	Metric	GBRQ	PDCA
Accuracy	RMSE Mean (STB)	19565.12	60617.775
	RMSE P50 (STB)	15590.773	57914.065
Probabilistic Reliability	Calibration Score	0.0004	0.0023
	Coverage Ratio	1.007	0.901
	Confidence Bias	-0.007	0.099
	Directional Bias	1.333	0.455
Uncertainty	Average Normalized Uncertainty Window	0.8281	1.2764
Cost	Computational Time (min)	<1	228

Table A.6—GBRQ vs PDCA statistics - 360 months.

	Metric	GBRQ	PDCA
Accuracy	RMSE Mean (STB)	21953.519	42675.656
	RMSE P50 (STB)	18645.714	48392.494
Probabilistic Reliability	Calibration Score	0.0005	0.0025
	Coverage Ratio	0.939	0.903
	Confidence Bias	0.06	0.097
	Directional Bias	0	0.381
Uncertainty	Average Normalized Uncertainty Window	0.7791	1.3388
Cost	Computational Time (min)	<1	141

Table A.7—GBRQ vs PDCA statistics - 440 months.

APPENDIX B – CALIBRATION PLOTS FROM PDCA RUNS

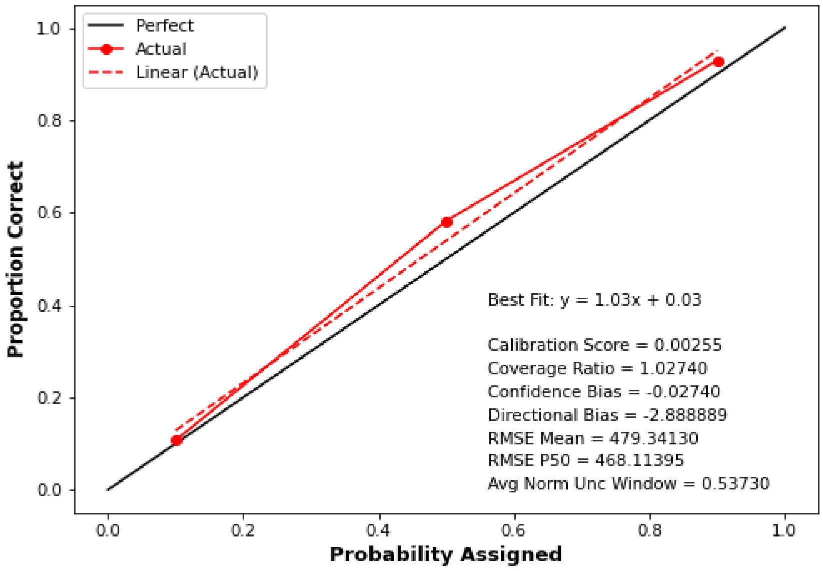


Fig. B.1—PDCA calibration plot - 21 months.

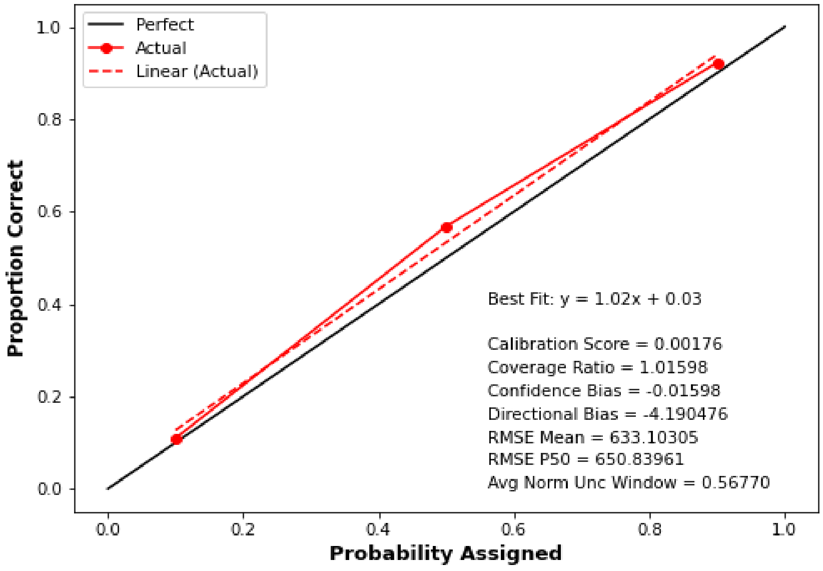


Fig. B.2—PDCA calibration plot - 24 months.

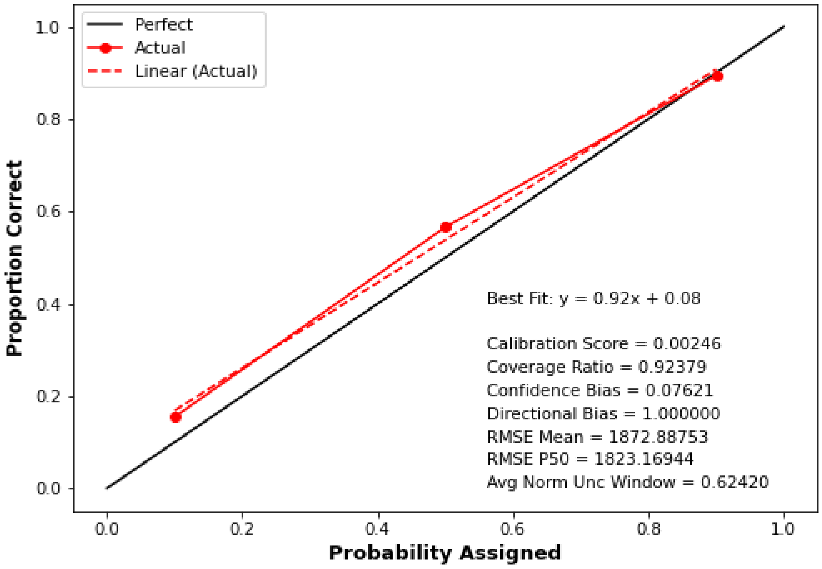


Fig. B.3—PDCA calibration plot - 36 months.

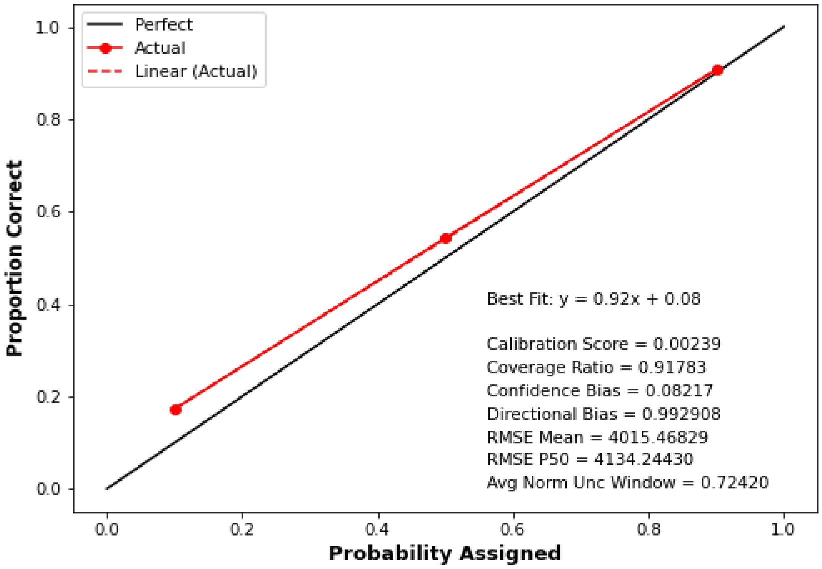


Fig. B.4—PDCA calibration plot - 60 months.

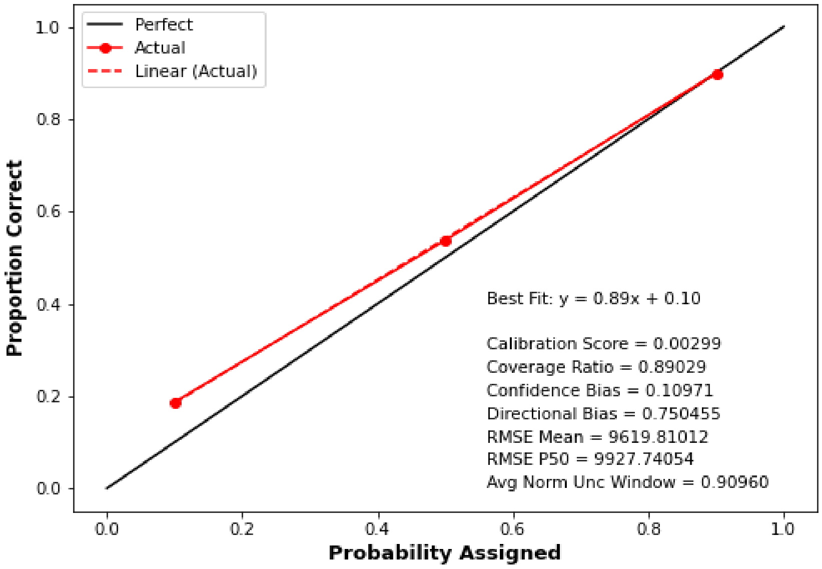


Fig. B.5—PDCA calibration plot - 120 months.

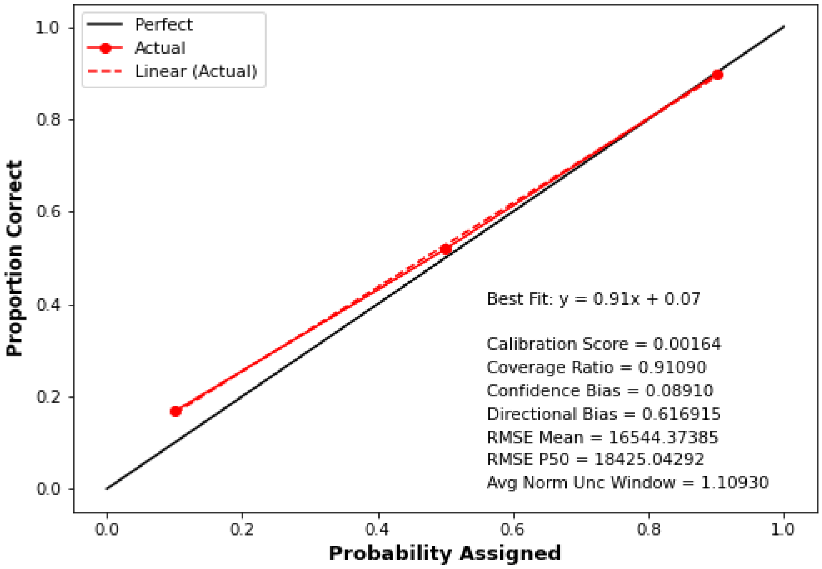


Fig. B.6—PDCA calibration plot - 240 months.

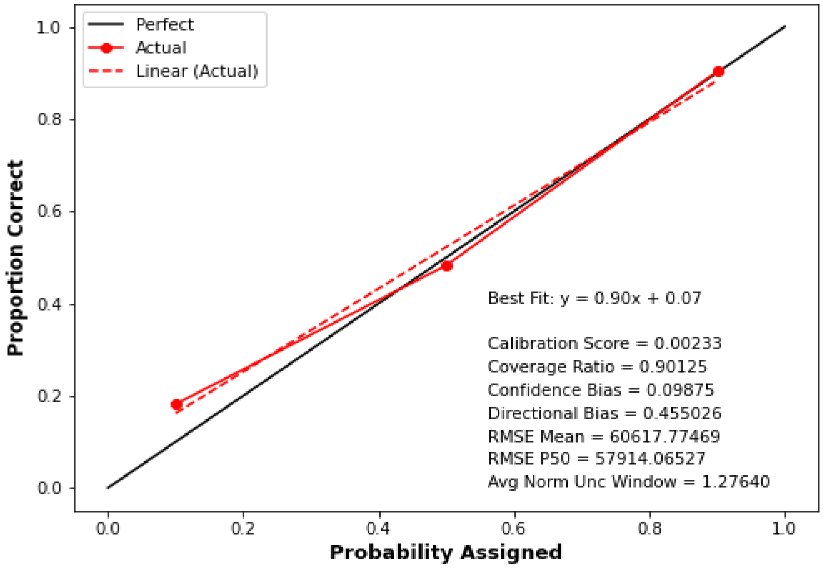


Fig. B.7—PDCA calibration plot - 360 months.

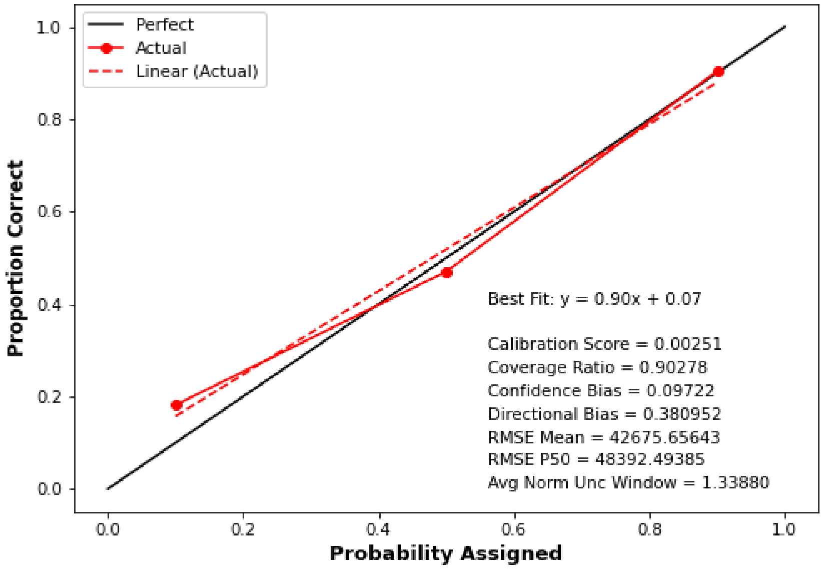


Fig. B.8—PDCA calibration plot - 440 months.

APPENDIX C – CALIBRATION PLOTS FROM GBRQ RUNS

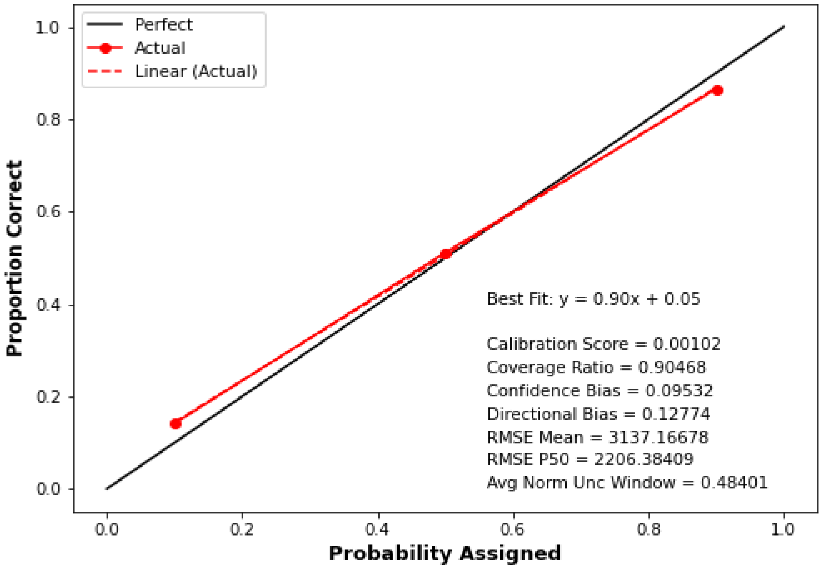


Fig. C.1—GBRQ calibration plot - 21 months.

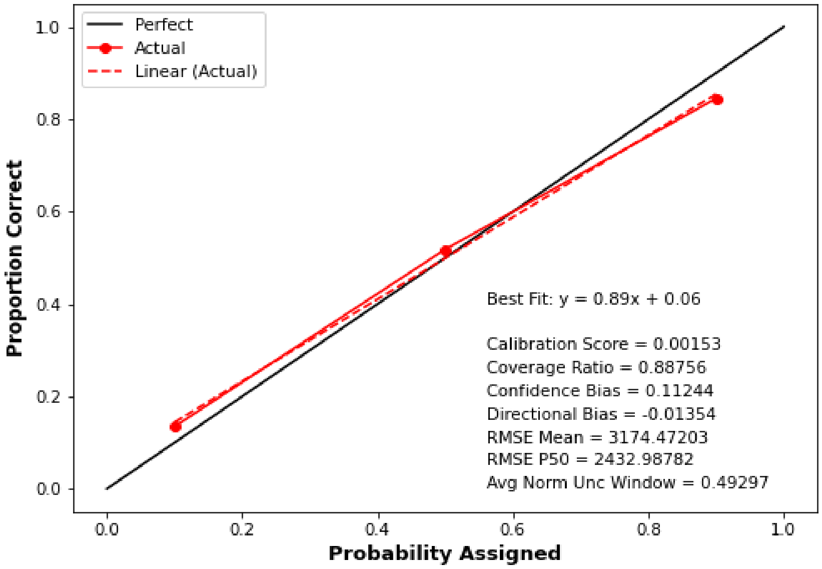


Fig. C.2—GBRQ calibration plot - 24 months.

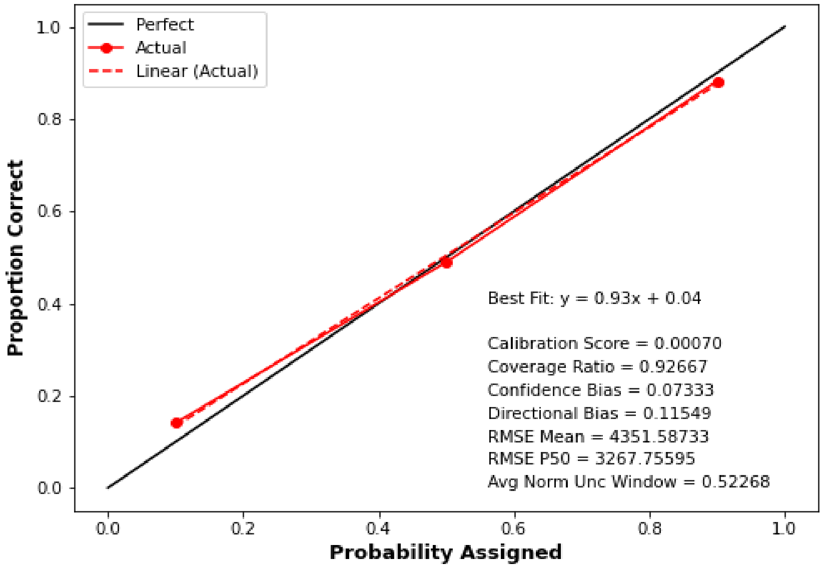


Fig. C.3—GBRQ calibration plot - 36 months.

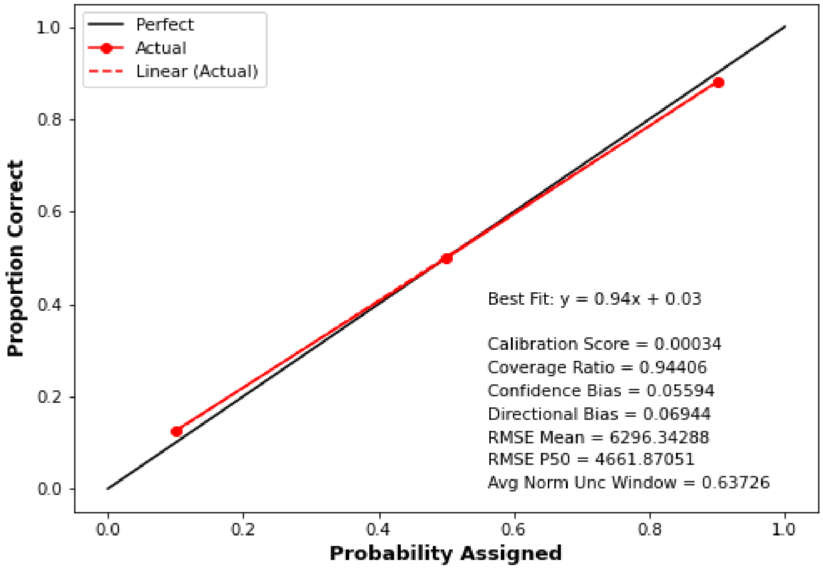


Fig. C.4—GBRQ calibration plot - 60 months.

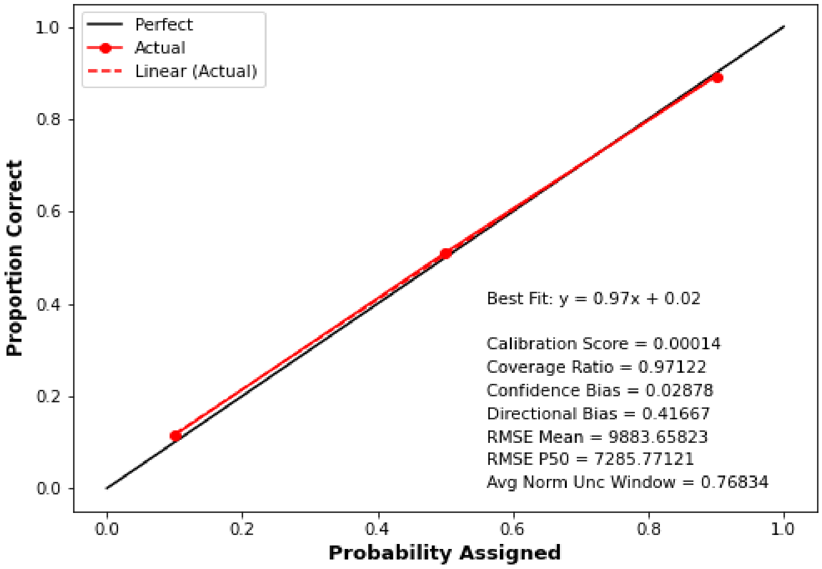


Fig. C.5—GBRQ calibration plot - 120 months.

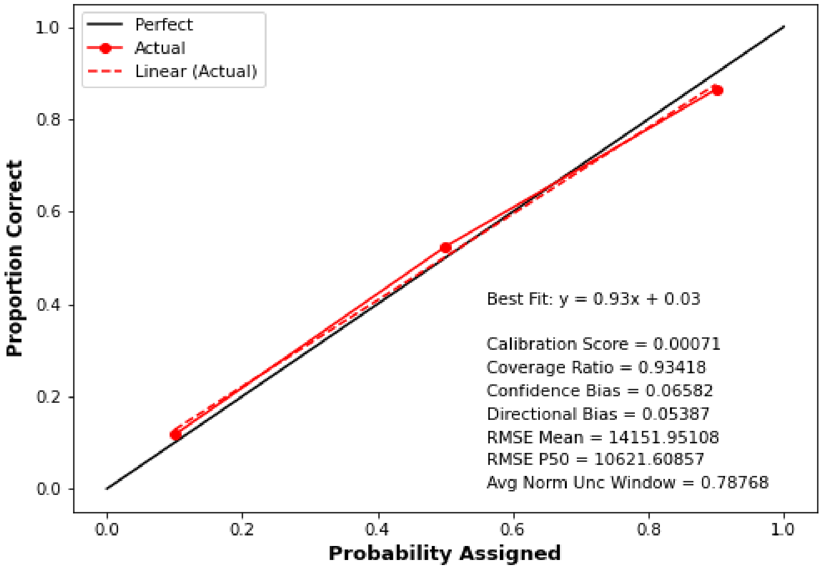


Fig. C.6—GBRQ calibration plot - 240 months.

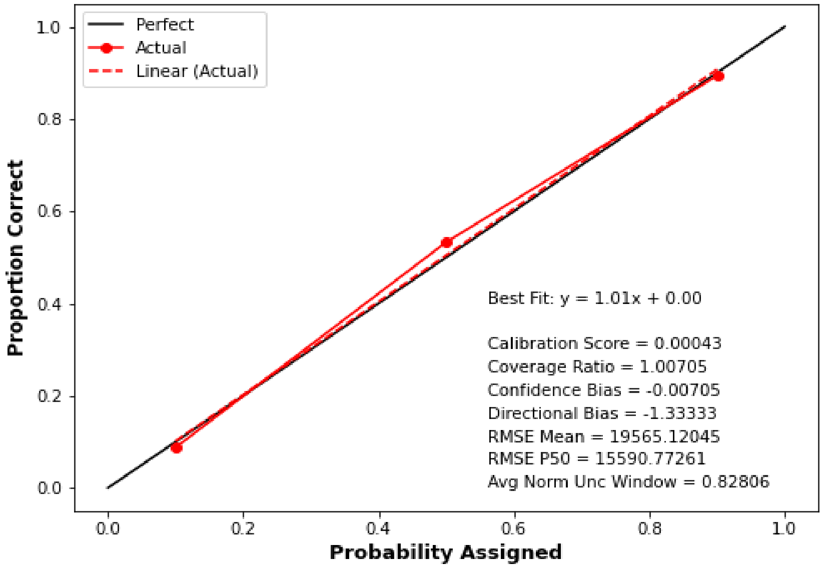


Fig. C.7—GBRQ calibration plot - 360 months.

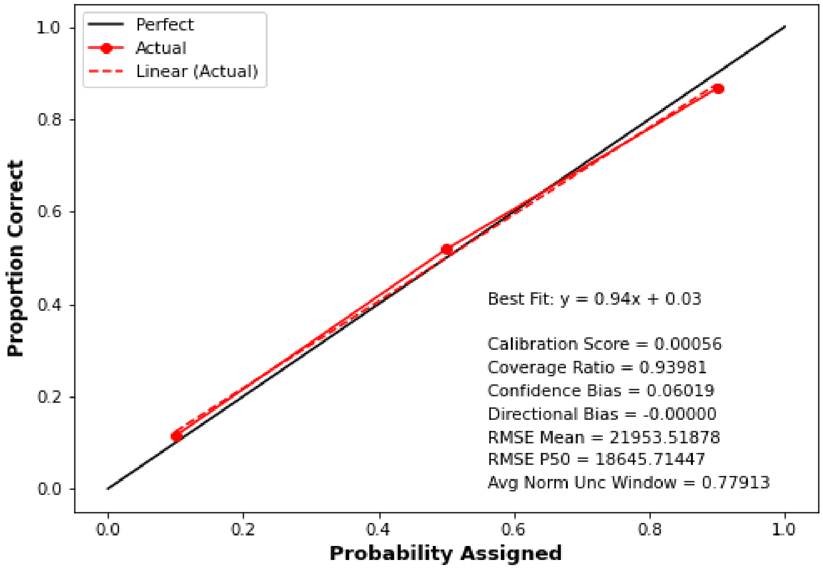


Fig. C.8—GBRQ calibration plot - 440 months.