Oil & Gas Science and Technology – Rev. IFP Energies nouvelles **75**, 7 (2020)

https://doi.org/10.2516/ogst/2019065

Available online at:
ogst.ifpenergiesnouvelles.fr

REGULAR ARTICLE OPEN ACCESS

Using data analytics to quantify the impact of production test uncertainty on oil flow rate forecast

Danielle D. Monteiro*, Maria Machado Duque, Gabriela S. Chaves, Virgílio M. Ferreira Filho, and Juliana S. Baioco

Universidade Federal do Rio de Janeiro, Av. Pedro Calmon, 550 – Cidade Universitária, Rio de Janeiro, RJ 21941-901, Brazil

Received: 14 August 2019 / Accepted: 9 December 2019

Abstract. In general, flow measurement systems in production units only report the daily total production rates. As there is no precise control of individual production of each well, the current well flow rates and their parameters are determined when production tests are conducted. Because production tests are performed periodically (e.g., once a month), information about the wells is limited and operational decisions are made using data that are not updated. Meanwhile, well properties and parameters from the production test are typically used in multiphase flow models to forecast the expected production. However, this is done deterministically without considering the different sources of uncertainties in the production tests. This study aims to introduce uncertainties in oil flow rate forecast. To do this, it is necessary to identify and quantify uncertainties from the data obtained in the production tests, consider them in production modeling, and propagate them by using multiphase flow simulation. This study comprises two main areas: data analytics and multiphase flow simulation. In data analytics, an algorithm is developed using R to analyze and treat the data from production tests. The most significant stochastic variables are identified and data deviation is adjusted to probability distributions with their respective parameters. Random values of the selected variables are then generated using Monte Carlo and Latin Hypercube Sampling (LHS) methods. In multiphase flow simulation, these possible values are used as input. By nodal analysis, the simulator output is a set of oil flow rate values, with their interval of occurrence probabilities. The methodology is applied, using a representative Brazilian offshore field as a case study. The results show the significance of the inclusion of uncertainties to achieve greater accuracy in the multiphase flow analysis of oil production.

1 Introduction

Several uncertainties exist in all segments of oil and gas industries including geology, reservoir, drilling, completion, and production. These uncertainties and their impacts on the segments have been studied. However, few studies have been done for the production segment as the uncertainties arise from multiphase flow correlations, heat transfer into the flow, reservoir data, measurements, and data computing; which influence the production test results.

In a production unit, the oil production measurement system usually reports only the daily total production rates. The flow rates and parameters of each well are measured only in the production test, which is done periodically (e.g., once a month). Thus, the data of the available wells are not extensive such that operational decisions are made using data that are not updated. Moreover, the production test data contain error sources including separator accuracy, equipment measurement, data obtained, and storage.

The production test data are used to calibrate well models to predict the expected production as it is the major source of profit. However, it is also important to predict the gas and water production, to achieve successful operations. The properties and parameters of the wells from the production test are used in the multiphase flow models to generate the oil, water, and gas flow rates.

Considering that production test data contains errors and is sometimes used for over a month without an update, it is important to understand that these parameters are not deterministic. Moreover, the data should be treated as stochastic while the impact of the uncertainties should be studied to obtain more reliable oil production predictions. Therefore, this study aims to introduce uncertainties into the prediction of the oil flow rate by identifying and quantifying uncertainties from the data obtained in production tests. It is then considered in production modeling and propagated on multiphase flow simulation. Moreover, this study is novel in that it combines two main study areas that are commonly used separately: multiphase flow simulation and data analytics, which comprises of statistical, machine learning, and stochastical methods. The merging of these

* Corresponding author: daniellemonteiro@poli.ufrj.br

areas yields a better understanding of the influence of the uncertainties of production test data on the prediction of oil flow and enables the development of more robust models.

The production tests are reported periodically during production from the well of a representative field. The test results are then transformed into a database for each well and used to study the parameter behavior. From the database to the oil prediction, this study goes through stages such as data pre-processing, stochastic modeling, sampling, multiphase flow simulation, and oil prediction.

In this study, data pre-processing includes techniques that remove outliers from the production historical data. In stochastical modeling, regression analysis computes the production parameter deviation using linear regression, segmented linear regression, polynomial regression, and Support Vector Regression (SVR) models. These deviations are then fitted by a suitable probability distribution, and more samples are generated by Monte Carlo and Latin Hypercube Sampling (LHS) methods. The multiphase flow model uses a black-oil model and multiphase flow correlation to describe the resulting fluid mixture and the pressure drop along the flow, respectively. In the oil prediction stage, the oil flow rate is predicted based on the uncertainty of the production parameters.

The remainder of this paper is structured as follows. Section 2 presents related literature. Section 3 describes the framework applied to quantify the impact of uncertainty on oil flow rate prediction. The methodology workflow is applied to the wells of a representative Brazilian offshore field and the case study is presented in Section 4. Finally, Section 5 concludes the paper.

2 Literature review

The oil industry usually considers the data as deterministic and computes the models accordingly. Several literatures have studied uncertainties and more areas are trying to model these uncertainties. However, most studies that consider the uncertainties in the oil industry are in the reservoirs segment. In this section, some works that correlate with the theme of this study are presented.

Charles *et al.* (2001) developed a general methodology that includes the uncertainties in an oil field, based on four steps. First, clarify the operational decision that will consider uncertainty. Second, reformulate the problem to include the uncertainties. Third, list the impacts of the parameter uncertainties on the study. Lastly, translate the uncertainty parameters into gross rock volume, oil-in-place, or reserves. They applied the study in four field cases and concluded that the approach could be used to translate parameter uncertainties into uncertainties on parameters of economic interest. The approach helps to increase the acceptance of projects. Meanwhile, they pointed out that an obstacle encountered to study uncertainties is the culture of engineers to defend one single technical choice.

On the other hand, Tyler *et al.* (1996) proposed a stochastic model to evaluate the uncertainty in reservoir engineering. They used an object-based stochastic model to generate reservoir heterogeneities that is combined with reservoir and Monte Carlo simulations to evaluate the uncertainty in Hydrocarbon Pore Volume, the pore volume in a reservoir formation available to hydrocarbon intrusion, which is used to estimate the reserves. As a result, they projected the uncertainties in project risk management and presented the risk in stochastic design. Chang and Lin (1999) proposed a stochastic method that analyzes production data to predict future performance and estimate probable reserves. They analyzed production data using regression, to obtain the parameters of the hyperbolic decline equation which calculates the residuals. Thus, PDF (triangular distribution) can be obtained, using a random generator. They concluded that the analysis results based on stochastic method are fully dependent on the production history data rather than subjective judgment. Dejean and Blanc (1999) proposed a simplified model to quantify the controlled and uncontrolled uncertainties in the reservoir predictions. They used experimental design, response surface methodology, and Monte Carlo statistical methods, and applied them to a field case. Corre *et al.* (2000) studied the impact of uncertainties in geophysics, geology, and reservoir engineering on project evaluation, to make better decisions in the risk analysis. They quantified the uncertainty impact on gross rock volume, oil originally in place, recoverable reserves, and production profiles. These parameters are very important at the project phase because they directly influence the economic project. However, Zabalza-Mezghani *et al.* (2004) reviewed the sources of uncertainties in reservoir management and concluded that the reservoir uncertainties influence the production forecast. This study proposes a statistical method, based on experimental design technique, to study uncertainties.

On the other hand, Maschio *et al.* (2010) proposed a new methodology to analyze the uncertainties in reservoir simulation models and reduce the number of reservoir uncertainty attributes, using observed data and LHS. Feraille and Marrel (2012) propose a statistical method to reduce uncertainties on the most influence parameters and to propagate these uncertainties to production forecast, using probability distributions.

Goda and Sato (2014) used Latin Hypercube to develop a new methodology to reduce the dimensionality and processing time required for history matching in petroleum reservoir studies with an Iterative Latin Hypercube. As Goda and Sato (2014), the work of Mahjour *et al.* (2019) also uses Latin Hypercube to characterize the uncertain existing in the reservoir models. The sampling method is used to combine different types of uncertain and generate set of reservoir models. As the last two works, (Costa *et al.*, 2019; Schiozer *et al.*, 2019) also use Latin Hypercube as the sampling method on its works, emphasizing that Latin Hypercube is as efficient method that reduces the number of necessary simulations.

Several more studies exist for reservoir uncertainties than for production uncertainties owing to the difficulty in obtaining reservoir and geology parameters, impact of the uncertainties on fields development projects, and stochastic nature of some parameters. However, the applied techniques can be used as an example of the production segment, such as MCS and LHS techniques.

Furthermore, some studies exist for uncertainties in wellbore stability. Morita (1995) conducted an uncertainty analysis based on statistical error analysis, to determine the parameters that influence the safe mud weight used in stabilizing the borehole. Moreover, Sheng *et al.* (2006) combined Monte Carlo uncertainty analysis and geomechanical modeling to predict the optimal mud weight window. Carpenter (2014) used Monte Carlo simulation to investigate the contribution of parameter uncertainties in fracture and collapse model predictions. Furthermore, Niño (2016) proposed a methodology based on sensitivity and uncertainty analyses for performing a wellbore stability analysis, using Monte Carlo simulation. As a result the required mud weight is obtained as a probability function, (Niño, 2016).

However, Fonseca Junior *et al.* (2009) proposed a methodology to evaluate the uncertainty in multiphase flow simulators using Monte Carlo simulation. In the first stage, it was considered a uniform distribution to uncertainty estimation of simulator input data. A sensitivity analysis was then performed to identify the input variables that have more impact on the output data. They concluded that the oil flow rate was the output variable that was more sensitive to the input data uncertainty.

However, an increasing number of studies on data-driven models for petroleum production focus on the prediction of production parameters and flow rates. Jahanandish *et al.* (2011) presented an artificial neural network model to predict the bottom-hole pressure and the pressure drop in vertical multiphase wells. The results indicated that the model has a high correlation coefficient and 3.5% absolute average percent error.

Grimstad *et al.* (2016) developed a completely data-driven study for real-time decision support. The methodology adopted can be divided into a few main parts: data pipeline; transform data to operational advice. One application of the operational advice is estimation which includes rate estimation and rate allocation. The study emphasized that it is important to estimate flow rates using data-driven methods when a well test could not be performed or when a multiphase flow meter is under maintenance or fails.

Monteiro *et al.* (2017) developed a methodology of uncertainty analysis for production forecasting in oil wells. Firstly, a statistical analysis was performed to identify and quantify the uncertainties. Monte Carlo simulation was then used to propagate these uncertainties into the multiphase flow simulator, based on beta distribution.

On the other hand, Balaji *et al.* (2018) reviewed the status of data-driven methods and their application in the oil and gas industry. The main areas of the industry that use data-driven methods are reservoir management and simulation, production and drilling optimization, drilling automation, and facility maintenance. The study emphasized the significant reluctance of oil companies to adopt these methodologies, despite their numerous advantages.

Furthermore, Spesivtsev *et al.* (2018) proposed a methodology for constructing a predictive model using a machine learning approach for bottom-hole pressure. The methodology used a neural network with two hidden layers. At last, in the production area, Sales *et al.* (2018) use a Monte Carlo simulation approach to deal with uncertainties for the field design layout problem. The sampling method objective is to generate many production scenarios of an oil field in order to assign the wells flow rates a probability distribution. As a conclusion, the authors emphasize that considering the uncertainties that exist on the process allow to know the possible outcomes and helps the decision making about the field design.

Table 1 summarizes the literature review. The works are listed based on the focus area and main methodology used.

The last row of Table 1 summarizes how this article relates to the literature review. Its main focus is production and its methodology is a physical model, using multiphase flow simulation and data analytics, which includes statistical models, stochastic methods, and machine learning models.

3 Framework for quantifying the impact of production test uncertainty on oil flow rate forecast

The framework used in this work to quantify the impact of production test uncertainty in oil flow rate forecast is summarized in Algorithm 1. This algorithm shows an overview of the steps presented in this study as well as the main approaches and techniques used. To study parameter uncertainty and its impacts on oil flow rate predictions, the production test data are statistically treated and regression techniques are applied to model the parameters previously selected. The best distribution is then fitted using the parameter model and samples are generated by sampling techniques. These samples and multiphase flow model are employed to incorporate a stochastic behavior on the production parameter. Finally, the multiphase flow model results are used to quantify the parameter uncertainty on the oil flow rate forecast. Each of these steps is explained in the subsections in more details.

Algorithm 1: Quantifying the impact of production test uncertainty on oil flow rate forecast

Input: Production test data
1) Data pre-processing
   <Obtain> Treated data
2) Stochastical modeling
   <Obtain> Regression model
   <Obtain> Probability distribution
3) Sample
   <Obtain> Set of random variables
4) For each sample:
   a. Simulate
   End For
   <Obtain> Flow rate stochastic distribution
5) Analyze
End

Table 1. Literature review summary.

Reference	Focus	Methodology
Charles *et al.* (2001)	Reservoir	Generic
Tyler *et al.* (1996)	Reservoir	Stochastic methods
		Monte Carlo
Chang and Lin (1999)	Reservoir	Stochastic methods
Dejean and Blanc (1999)	Reservoir	Monte Carlo
Corre *et al.* (2000)	Reservoir	Generic
Zabalza-Mezghani *et al.* (2004)	Reservoir	Statistical method
Maschio *et al.* (2010)	Reservoir	Latin Hypercube
Feraille and Marrel (2012)	Reservoir	Statistical method
Goda and Sato (2014)	Reservoir	Latin Hypercube
Mahjour *et al.*, (2019)	Reservoir	Latin Hypercube
Costa *et al.* (2019)	Reservoir	Latin Hypercube
Schiozer *et al.* (2019)	Reservoir	Latin Hypercube
Morita (1995)	Wellbore stability	Statistical error analysis
Sheng *et al.* (2006)	Wellbore stability	Monte Carlo
Carpenter (2014)	Wellbore stability	Monte Carlo
Niño (2016)	Wellbore stability	Monte Carlo
Fonseca Junior *et al.* (2009)	Production	Monte Carlo
Jahanandish *et al.* (2011)	Production	Artificial neural network
Grimstad *et al.* (2016)	Production	Data-driven method
Monteiro *et al.* (2017)	Production	Statistical analysis Monte Carlo
Balaji *et al.* (2018)	Data-driven methods	Literature review
Spesivtsev *et al.* (2018)	Production	Machine learning
Sales *et al.* (2018)	Production	Monte Carlo
This article	Production	Physical model
		Data analytics

3.1 Data pre-processing

Data is selected from well production tests by gathering test information and creating a dataset with measured variables that characterize the production of a well with time. A variable having a great influence on the oil flow rate should be selected from the dataset as the input variable for the following procedure. The impact of this input variable on oil flow rate forecast can then be observed at the end of the procedure. The variable is chosen based on two criteria: operational characteristics of the wells considered and statistical support by correlation analysis. In the first, physical conditions of the wells and the expertise of the production engineer are considered. The second criterion uses a linear correlation between each variable of the dataset and the oil flow rate to identify variables having a high correlation with oil flow.

From the selected variable, outliers are identified and removed in a subsequent procedure. Anomalous points resulting from different sources including typing errors and monitor equipment failures are usually observed when real data is used, which can lead in incorrect analysis results and difficulty in the modeling process. Therefore, it is necessary to treat the dataset by applying methods to identify and remove inconsistent observations referred to as outliers.

Hence, modified Z-score method (Iglewicz and Hoagling, 1993) was applied to identify the outliers in the series. For each sample, a parameter M_i is calculated using equation (1):

$$M_i = \frac{0.6745}{\text{MAD}}\,(x_i - \tilde{x}) \tag{1}$$

where x_i is the value of sample i, $\tilde{x}$ is the median of the samples, and MAD is the median absolute deviation, given by median $\{|x_i - \tilde{x}|\}$. Using modified Z-score, the outliers are identified when $|M_i| > D$, where $D = 3$. The advantage of this method is that it uses median and MAD rather than mean and standard deviation, respectively. Median and MAD are robust measures of central tendency and dispersion, respectively. Moreover, methods that use median have better performance for a small dataset, which is the case of the dataset used in this study.

To perform data cleaning procedure using the modified Z-score method, it is necessary to fit the data of the analyzed variable using a suitable model. The deviations are then calculated based on the difference between the actual

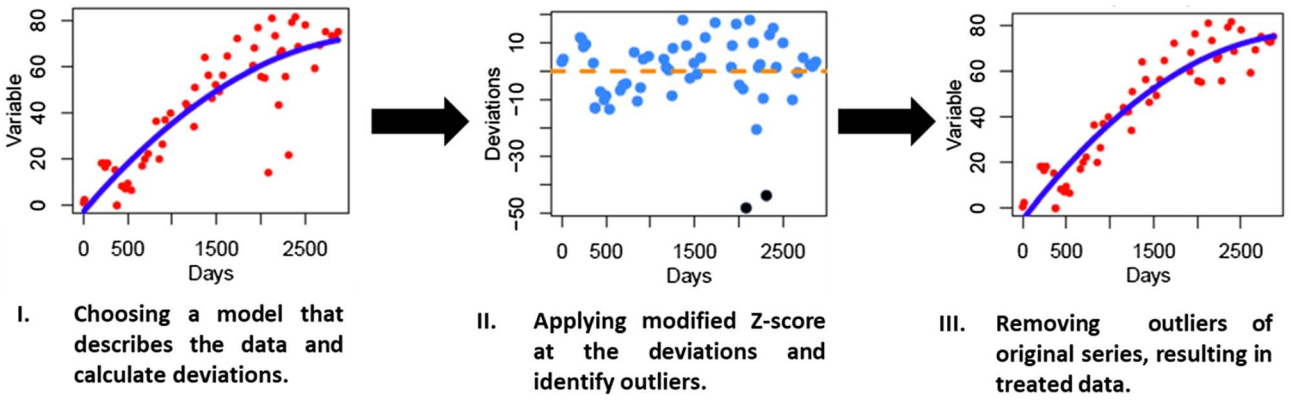

Fig. 1. Outlier detection methodology.

data and the value adjusted by the model to remove temporality of the series. The modified Z-score method is applied to these deviations to identify the outliers, as illustrated in Figure 1. The anomalous points are then removed from the actual series and the process is repeated until no outliers remain in the data.

3.2 Stochastical modeling

Stochastical modeling is employed in this study to identify a regression model that better describes the behavior of the variable with time and to obtain a suitable continuous distribution probability to quantify the uncertainty of the variable. Meanwhile, it is important to model the variable data used. Based on an analysis of the dataset, six regression models were considered in this paper: simple linear regression, segmented linear regression with one and two breakpoints, polynomial regression of second and third order, and SVR. Figure 2 illustrates an example of these six models. Meanwhile, segmented regression and SVR are explained below.

Segmented linear regression models are used when the relationship between the response and one or more explanatory variables are piecewise linear, *i.e.*, represented by two or more straight lines connected by breakpoints (Muggeo, 2008). At these points, regression lines change direction. The breakpoints and coefficients of the regression were estimated using Segmented package (Muggeo, 2008) in R statistical software. For this paper, segmented linear regression analysis was performed considering one and two breakpoints as these two cases were effective for the available data.

SVR (Vapnik, 1995) is a supervised learning method that executes nonlinear regressions *via* kernel functions. These functions transform the original data into a higher dimension by using nonlinear mapping for linear separation. At the new dimension, SVR attempts to find the best linear approximation by controlling errors based on the Structural Risk Minimization (SRM) principle, which expresses a trade-off between empirical risk and the complexity of the approximating function. Epsilon Support Vector Regression (ε-SVR) is one of the most recurrent versions of SVR and it is applied in this work. In the ε-SVR approach, an ε-insensitive loss function is used, which is zero if the difference between the observed and predicted values is less than ε, and loss will just be counted for deviations above ε. In this study, SVR was implemented using e1071 package in R statistical software (Meyer *et al.*, 2015), considering a radial kernel.

The suitable regression model depends on the variable, well, and the elapsed production time of the analysis. The choice of the best model among the six models considered in this paper is made using graphical support (Fig. 2) and by comparing three statistical measures for each regression model j such as (i) Residual Standard Error (RSE_j), (ii) Mean Absolute Error (MAE_j), and (iii) Adjusted-R^2 ($R^2_{\mathrm{Adj},j}$).

RSE_j and MAE_j are error measures whose values must be close to zero for a suitable model. The values of these statistical measures are given by equations (2) and (3), respectively:

$$\mathrm{RSE}_j = \sqrt{\frac{1}{n-2} \times \sum_{t=1}^{n} \left(Y_{\mathrm{tj}} - \widehat{Y}_{\mathrm{tj}}\right)^2} \tag{2}$$

$$\mathrm{MAE}_j = \frac{1}{n} \times \sum_{i=1}^{n} \left| Y_{\mathrm{tj}} - \widehat{Y}_{\mathrm{tj}} \right| \tag{3}$$

where Y_{tj} is the real value, $\widehat{Y}_{\mathrm{tj}}$ is the adjusted value by the model j, and n is the number of data point samples of the variable for the analysis.

Adjusted-R^2 ($R^2_{\mathrm{Adj},j}$) is another error measure used. However, a value close to one usually indicates that the model is an adequate estimator of the data. It is calculated using equation (4):

$$R^2_{\mathrm{Adj},j} = 1 - \left(\frac{\sum_{t=1}^{n} \left(Y_t - \widehat{Y}_t\right)^2 / \mathrm{df_e}}{\sum_{t=1}^{n} \left(Y_t - \overline{Y}\right)^2 / \mathrm{df_t}} \right) \tag{4}$$

where $\mathrm{df_e}$ indicates the degree of freedom of the estimated error variance of the variable sample, $\mathrm{df_t}$ is the degree of freedom of the variance of the dependent variable Y that is the analyzed variable, and $\overline{Y}$ is the mean value of the sample variable.

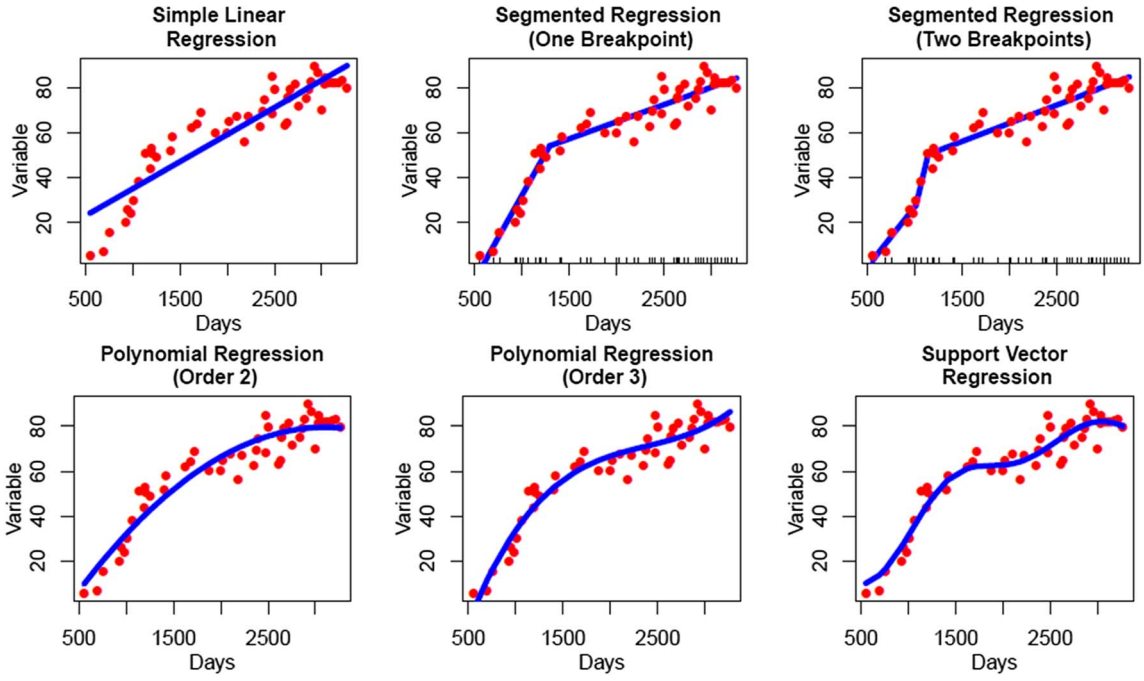

Fig. 2. Regressions models.

Moreover, validation tests are conducted to support the model choice. To perform the tests, the data used to parameterize the methodology are randomly split in the ratio of p and $(1 - p)$ into two groups: calibration and validation sets, respectively. The main idea is to use the calibration set to fit the models discussed in this section and evaluate these models using the validation set. Validation tests are performed n rounds for each well. For each model j, the mean values of $R^2_{\text{Adj},j}$ and RSE_j of the calibration set are computed. Furthermore, the mean value of MAE_j is calculated for the validation set to evaluate the result and define how models that predict values differ from real variable data.

After the most appropriate regression model that fits the variable in the analysis is chosen, the deviations are calculated based on the difference between the actual data and the adjusted values. The uncertainty quantification process is performed by identifying and adjusting a continuous probability distribution on the chosen model deviations, using the FitDistrPlus package (Delignette-Muller and Dutang, 2015) in R. The identification of the distribution parameters that represent the deviation can provide information on how the actual values will be scattered in relation to the adjusted model, considering that the model represents the variable behavior with time.

3.3 Sampling

Sampling is done as follows:

I. Sampling methods: Two sampling methods such as Monte Carlo Sampling (MCS) and Latin Hypercube Sampling (LHS) were used to propagate the uncertainties of the analyzed variable. Both methods are non-intrusive-based methods that use input variable distribution to generate random samples. Meanwhile, MCS is more common as it is generally applied in many areas including petroleum industry. However, it may be inefficient for intensive computational problems, as large samples are required to achieve good accuracy. On the other hand, LHS usually reduces the number of iterations to obtain the same level of accuracy as MCS. The expected range for the uncertainty distribution variable is discretized in segments and at least one sample is considered for each segment to obtain a better representation of the cumulative probability distribution with fewer samples. LHS reduces the variance of statistical estimators in comparison with classical Monte Carlo estimates (Shields and Zhang, 2016). Thus, both methods are applied using the parameters of deviation distribution adjusted to produce n samples of deviations.

II. Constructing the set of random variables: The set of random variables is created based on the sum of the deviations sampled by MCS or LHS and the variable predicted by the regression model, using its equation. For example, if segmented regression is chosen, the linear equation of the last segment will be used to make the prediction.

3.4 Multiphase flow simulation

Nodal analysis is widely used in the oil industry to perform iterations between the reservoir, wellbore, and flowline system. The production flow rate is a result of the equilibrium between the available and required pressures at the chosen

node (in this case, bottom hole). The available pressure is the capability of the reservoir to produce whereas the required pressure is the pressure needed to produce the fluids from the reservoir to the separator.

In this work, the reservoir is represented by Inflow Performance Relationship (IPR), although IPR can be considered as a simple model, it is a common practice in industry use it as a input data of multiphase flow model since the reservoir and multiphase flow models are decoupled. Empirical correlations are used to evaluate the pressure drop along the well and pipeline system. Moreover, the fluid mixture is represented by a black-oil model. The nodal analysis is executed at a multiphase flow simulator developed using Python. Meanwhile, the flow rates are the target outputs.

To obtain the target output, namely a set of oil flow rate values, a sensitivity analysis is conducted for the chosen production variable using the created set of random values. Thus, multiphase model used is tuned to the data of the last well production test available. The model calibration is done to characterize the behavior exhibited by the well in the last production test, *i.e.*, the resulting flow rate of the model must be the same as the measurement during the production test, given the measured pressures.

The multiphase flow simulation results in a set of oil flow rate values as a response to the uncertainty of the chosen production variable.

3.5 Production forecast analysis

The analysis of the production forecast was made based on the probability distribution of the resulting oil flow rate values. To represent this distribution, three main measures were selected based on widely used terms in the industry such as: the P-10, P-50, and P-90 percentiles.

The P-10 percentile represents a pessimistic condition as there is a 90% chance that the value of the oil flow rate is greater than this value. Meanwhile, P-50 represents the situation where there is an equal chance of the flow being lower or being higher than the measure, while P-90 represents the optimistic condition, *i.e.*, there is a 10% chance that the oil flow rate value is greater than P-90. The probabilities of occurrence of the different flow rates are calculated by an empirical probability distribution using the set of ordered oil flow rate values.

During analysis, the actual value of the oil flow rate should exist between the P-10 and P-90 percentiles. If this occurs, the production forecast is considered successful.

4 Case study

The methodology developed was applied to 13 production wells with production test data measured at a representative Brazilian offshore field to evaluate the proposed procedure. The total amount of analyzed data from the 13 wells consisted of above 1000 production tests with 13 production variables each, emphasizing once again the use of data analytics tools on this work. The analysis was performed for the uncertainty of only one production variable to observe its impact on the oil flow rate. As production tests contain different variables, a correlation matrix (Fig. 3) was constructed with the data of representative production variables to identify the variables that have greater influence on oil flow rate response.

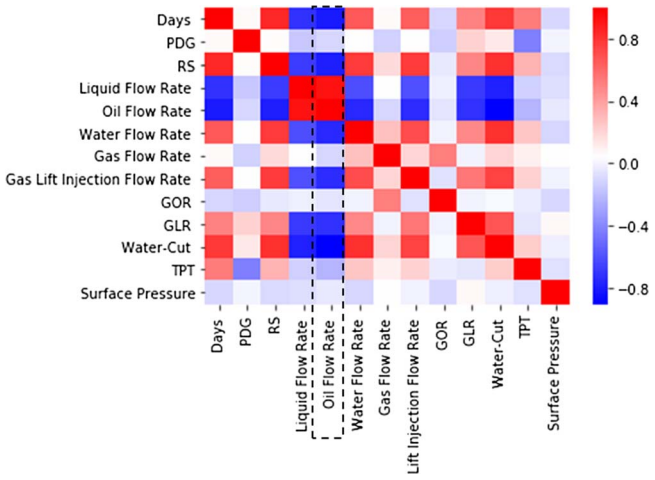

Fig. 3. Correlation matrix.

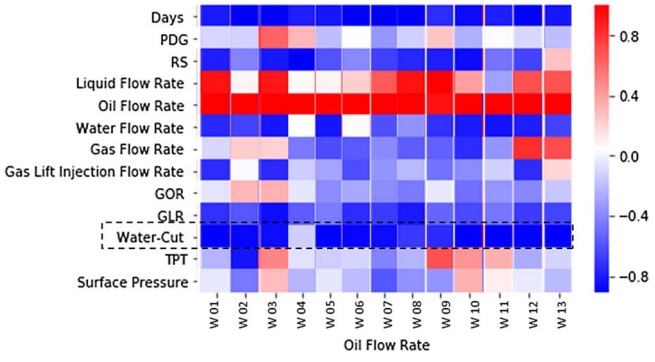

Fig. 4. Oil flow rate correlation matrix.

The production variables that were used to construct the correlation matrix were: all pressures available: PDG – bottom-hole pressure, TPT – wellhead pressure and surface pressure; all flow rates available: oil, water, gas and gas injection flow rates; RS – solution gas-oil ratio; GOR – gas-oil ratio; GLR – gas–liquid ratio; days in production; and water-cut.

A correlation matrix indicates associations between variables and computes the corresponding linear statistical correlation.

Because this study focuses on the impact of uncertainty on oil flow rate response, Figure 4 was constructed with thirteen correlation matrix slices corresponding to the oil flow rate columns for each well. Each column is a different well and each line corresponds to a production variable.

In Figure 4, a dark red color indicates a strong positive correlation and a dark blue color indicates a strong negative correlation between two variables. Thus, the oil flow rate has a strong correlation with water-cut in almost all wells, except the water-cut value in a well is lower than the others.

Hence, water-cut was chosen as the uncertainty variable for further analysis. Table 2 summarizes the water-cut range of the thirteen production wells used for this case study.

Although little can be done during production in relation to water-cut value, since there are uncertainties inherent on measuring its value and its future value and these uncertainties strongly affect oil production, the in-depth water-cut analysis is beneficial to more robust oil total production prediction.

For each well, the dataset was separated into two groups: a set containing the last four tests was reserved for validation while others were used to parameterize the methodology, *i.e.*,

(i) outliers of water-cut data were removed at the preprocessing phase,
(ii) a suitable regression model was adjusted to the variable and the distribution of the deviations was identified,
(iii) sampling techniques were applied to generate representative samples of deviations,
(iv) four water-cut values were predicted using the adjusted statistical model at the same dates of the test set,
(v) for each water-cut value predicted, deviations randomly generated using a sampling technique were summed to this value to generate a cloud of water-cut data,
(vi) the water-cut data cloud was used as input parameters of the multiphase flow simulator while a representative set of oil flow rates was produced as output,
(vii) the values obtained by the forecast methodology are compared with the real values of oil flow rate contained in the test set.

Table 2. Properties of the production wells.

Well	Water-cut range (%)
Well 01	75–95
Well 02	30–55
Well 03	20–40
Well 04	5–15
Well 05	60–95
Well 06	50–70
Well 07	30–60
Well 08	30–50
Well 09	60–80
Well 10	60–80
Well 11	70–90
Well 12	70–90
Well 13	50–80

4.1 Regression modeling validation

The regression model used should be carefully selected. Because the tendency of water-cut variable can be characterized during the production time, an unsuitable model can lead to erroneous results. Thus, validation tests were conducted to analyze the performance of regression models using water-cut data and to support the model choice. To perform the tests, the data used to parameterize the methodology were randomly split into two groups (calibration and validation set) in the ratio 0.7 and 0.3. Ten rounds of validations tests were performed for each well. The process for fourth round validation test of Well 11 is shown to illustrate the test: Figure 5 illustrates the three models

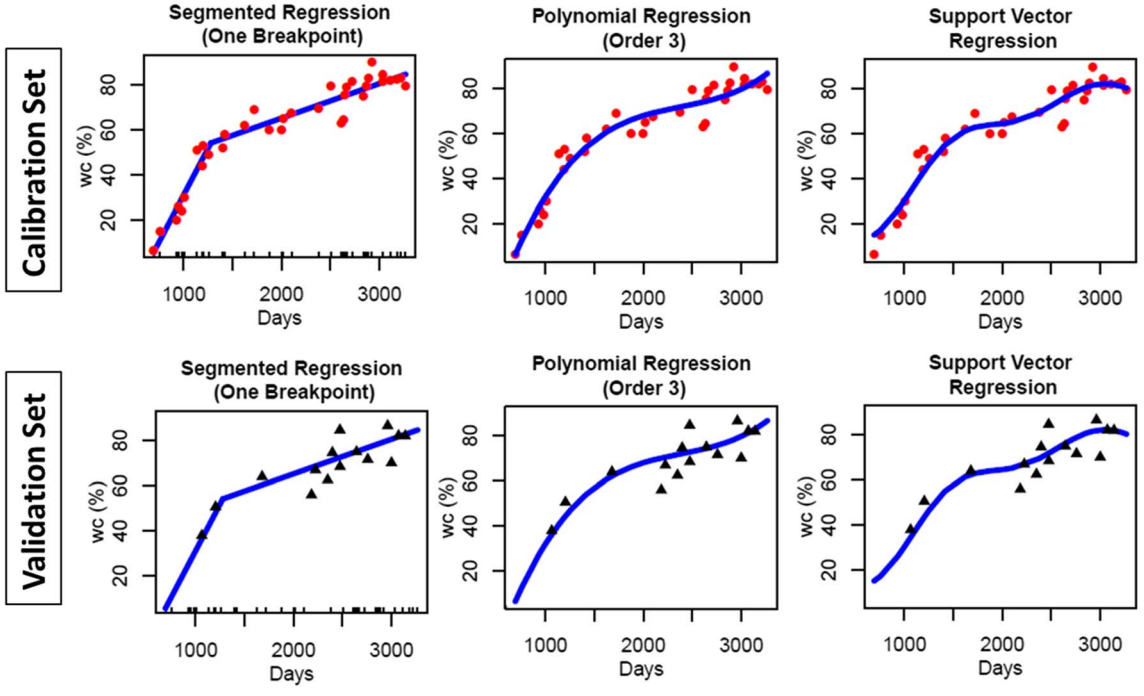

Fig. 5. Fourth round validation test for Well 11.

Table 3. Validation test results for Well 11.

Mean statistics measures	Segmented with one breakpoint	Polynomial regression (order 3)	Support vector regression	Set
$R^2_{\text{Adj},j}$	0.932	0.918	0.927	Calibration set
RSE_j	5.495	6.003	5.611	
MAE_j	5.309	5.658	5.515	Validation set

Table 4. Summary of statistical analyses.

Well	Model case choice	Mean $R^2_{\text{Adj},j}$ before removal of outliers	Mean $R^2_{\text{Adj},j}$ after removal of outliers	Mean RSE_j	Mean MAE_j
Well 01	SVR	0.903	0.968	5.312	4.235
Well 02	SVR	0.554	0.963	3.213	3.197
Well 03	Linear (simple)	0.728	0.866	4.014	3.485
Well 04	Linear (one breakpoint)	−0.012	0.962	0.523	0.396
Well 05	SVR	0.910	0.963	4.401	5.440
Well 06	Linear (two breakpoints)	0.708	0.961	4.838	3.217
Well 07	Linear (two breakpoints)	0.934	0.977	2.834	1.942
Well 08	Linear (one breakpoint)	0.532	0.981	1.281	0.839
Well 09	Polynomial (order 3)	0.917	0.920	7.213	5.853
Well 10	SVR	0.931	0.956	5.181	4.099
Well 11	Linear (one breakpoint)	0.819	0.932	5.495	5.309
Well 12	SVR	0.912	0.976	4.725	3.521
Well 13	Linear (two breakpoints)	0.737	0.878	8.796	8.129

having the best fit. The graphs at the top are models fitted by calibration set while the graphs at the bottom show the location of the validation set relative to these models. The statistical results obtained for the Well 11 are summarized in Table 3.

The choice of the best model(s) is made using the following criteria: First, the graphics of the models are displayed to identify the models that are consistent with the physical process of the water-cut variable. The possible models are then selected and their statistics, which are obtained using the validation test, are compared. The model that has the best measures is selected. Table 4 summarizes the statistical parameters obtained for the best model of each well. The table also shows $R^2_{\text{Adj},j}$ before and after removing the outliers for the chosen model. Analyzing the table, it is possible to observe improvements to the data after removing outliers, as $R^2_{\text{Adj},j}$ increased in all cases.

4.2 Selection of probability distribution of the deviations

The deviation data is fitted into a continuous probability distribution. Meanwhile, the probability distribution that best fits the deviation data obtained in the previous section was obtained using Stat::Fit software. This software uses maximum likelihood estimation to reject or accept each possible probability distribution. Among the 13 wells, each regression model has data composed of 78 water-cut deviations datasets. For each dataset, the software output is a rank from the probability distributions not rejected to the rejected ones. The analyses of the results obtained are illustrated in Figures 6 and 7. Figure 6 illustrates the overall result of the top 5 not reject probability distributions for the 78 datasets. Figure 7 illustrates the same results separated using regression model.

Hence, this work fits the water-cut deviation using beta distribution. In general (Fig. 6) this distribution is widely not rejected, with the same percentage than the Weibull distribution, 87%. In the results using regression model (Fig. 7), this distribution proportion of not reject is below 70% for the segmented regression with two breakpoints, however, any other distribution obtained a better result.

4.3 Comparison of sampling methods

To select one of the sampling methods (Monte Carlo or Latin Hypercube), the histogram results of water-cut and oil flow rate distributions were compared. The objective of this comparison is to replace 1000 MCS set points that are mostly used in studies with a smaller set of Latin Hypercube sample points.

Both methods used sampled sets of 100, 500 and 1000 points, and Figures 8 and 9 illustrate the results for 2 of the 13 wells tested. The results of Well 05 and Well 08 were

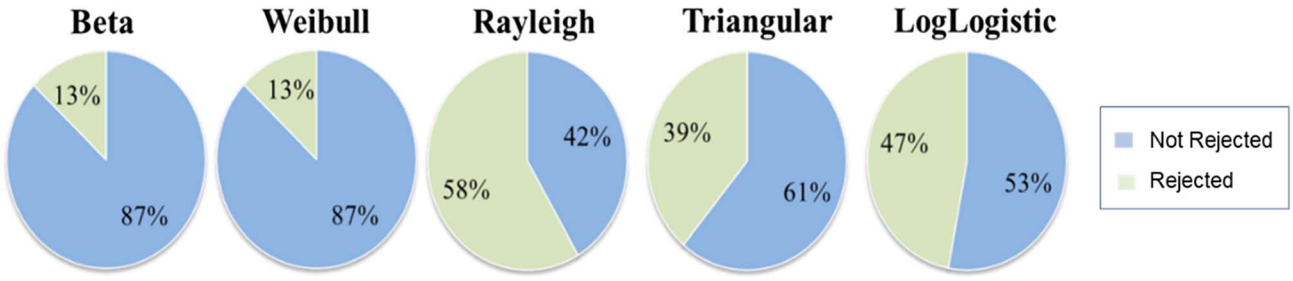

Fig. 6. Data analysis of general probability distribution.

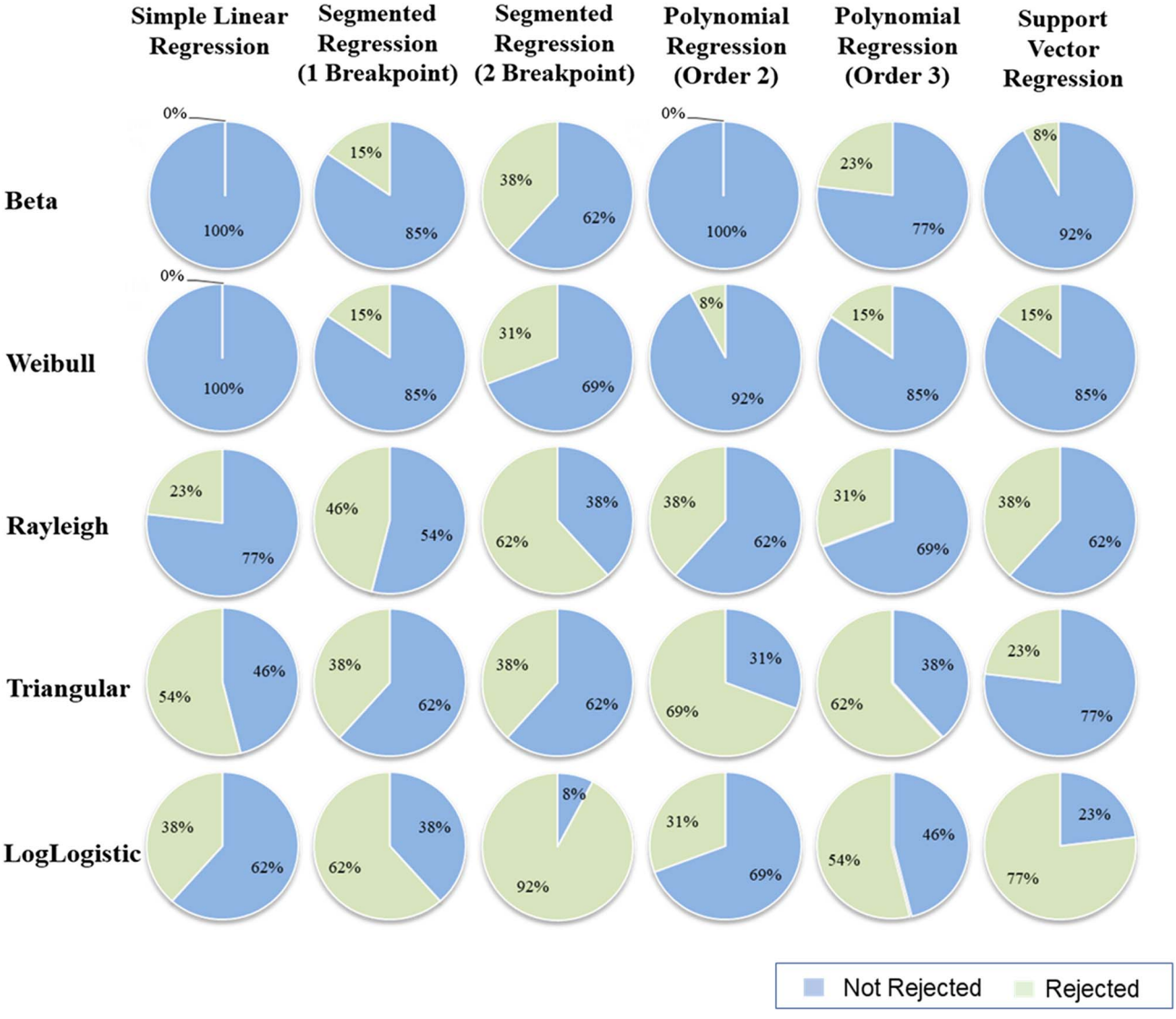

Fig. 7. Data analysis of probability distribution using regression model.

displayed because all the wells considered in this case study presented the same trend of results.

In Figure 8, each row represents 100, 500 and 1000 points, respectively, while each column represents a sampling method. All the histograms exhibited similar forms with the execution of the MCS with 100 points.

Moreover, the same behavior observed for the water-cut analysis was also observed for oil flow rate. All the

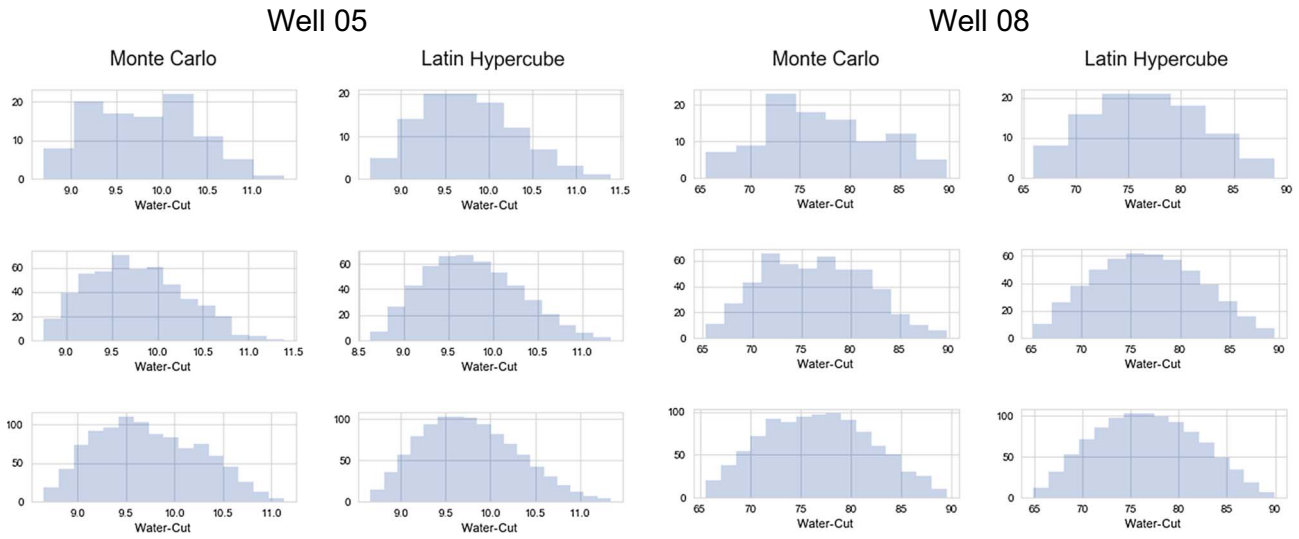

Fig. 8. Comparison of Monte Carlo and Latin Hypercube methods for water-cut.

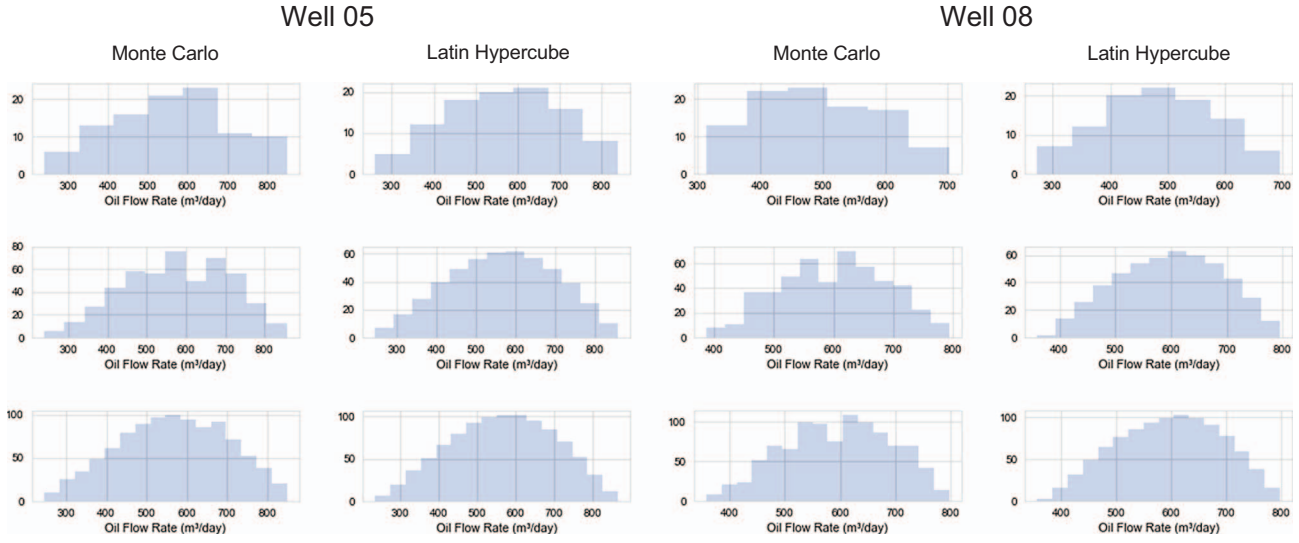

Fig. 9. Comparison of Monte Carlo and Latin Hypercube methods for oil flow rate.

histograms exhibited similar distributions while MCS with 100 points exhibited a different pattern.

Hence, the histogram analyses for all wells were quite similar for both sampling methods, in particular when compared the largest set of Monte Carlo to all sets of LHS.

Therefore, LHS with 100 points was used in oil flow rate forecast analysis, as it has a lower computational time because fewer simulation points were required.

4.4 Oil flow rate forecast analysis

To validate the proposed methodology, it was applied on a set of 13 wells to predict the production flow rate for four future dates.

Table 5 summarizes the detailed results for all wells. The first column shows the last production test date available for the well while the second column comprises of the four future production dates to be predicted. The subsequent columns represent the water-cut value, P-10 percentile, P-50 percentile, P-90 percentile, the production oil flow rate, and if the prediction was successful or not. In this case, the prediction was successful if the oil flow rate value was between the P-10 and P-90 percentiles. However, it is desirable that the production value should be close to the P-50 percentile.

Based on the 52 forecasted dates in Table 5, 42 presented accurate results, indicating 80.7% of total dates. Among all wells simulated, two were chosen for the detailed analysis owing to their representative responses to the problem: Well 05 and Well 06.

For Well 05, only the first prediction date was inaccurate, and the production oil flow rate values were higher than P-10 probability. In the first date, the behavior of

Table 5. Oil flow rate forecast results.

Well	Last test	Days	Water-cut (%)	Percentile P-10	Percentile P-50	Percentile P-90	Production oil flow rate	Forecast
Well 01	3282	3330	57.0	433.8	548	678	669.3	✓
		3384	64.2	455.9	570.5	705.2	524.8	✓
		3404	58.0	463.2	578.6	714.8	599.3	✓
		3458	61.1	483.1	600.6	734.1	463.5	×
Well 02	3023	3068	42.4	313.6	328.4	417.7	350.3	✓
		3135	41.4	313.5	328.1	417.8	320.4	✓
		3174	37.1	313.7	329	418.1	368.4	✓
		3213	39.0	312.5	356.7	418.1	358.4	✓
Well 03	3262	3300	30.4	650.7	707.8	771.5	718.3	✓
		3338	32.4	647.2	703.2	768.3	663.8	✓
		3391	30.4	651.4	657.7	700.9	694.4	✓
		3438	34.8	634.8	691	758.2	640.9	✓
Well 04	3176	3215	11.5	431.1	435.9	441.8	440.6	✓
		3251	11.6	435.6	440.2	445.7	437.5	✓
		3285	10.9	434.1	438.6	447.5	446.9	✓
		3324	16.1	432.7	437.3	443.3	419.6	×
Well 05	3125	3163	72.8	276.6	457.8	617.2	632.6	×
		3220	73.9	288.4	471.0	624.5	614.6	✓
		3253	85.7	296.7	478.5	635.4	338.9	✓
		3289	85.7	308.4	491.1	648.1	343.7	✓
Well 06	3298	3338	59.6	737.8	906.7	1093.7	531.2	×
		3416	59.4	723.5	893.7	1077.7	853.3	✓
		3452	61.2	715.5	887.7	1071.7	790.7	✓
		3466	62.9	713.1	884.1	1069.9	764.1	✓
Well 07	3255	3295	40.3	410.1	470.4	509.5	459.2	✓
		3334	46.1	422.3	479.4	569.5	546.4	✓
		3373	40.4	389.0	445.1	531.1	597.4	×
		3412	50.7	400.2	479.2	535.7	516.6	✓
Well 08	3281	3327	39.0	263.9	279.9	290.1	279.0	✓
		3366	38.0	259.7	275.1	285.5	246.6	×
		3403	39.0	248.2	270.7	272.1	247.1	×
		3431	33.0	267.2	271.1	277.6	251.1	×
Well 09	3166	3205	76.4	312.8	438.2	692.7	307.5	×
		3245	71.3	317.7	439.2	691.1	414.8	✓
		3280	68.6	317.1	437.7	692.4	458.7	✓
		3325	67.5	313.5	438.9	692.9	456.0	✓
Well 10	3109	3130	69.5	343.5	462.2	565.2	464.3	✓
		3169	72.6	351.5	468.6	573.7	442.6	✓
		3250	66.8	370.1	492.9	599.3	504.2	✓
		3290	75.6	383.8	505.4	611.3	368.5	×
Well 11	3266	3346	81.4	159.5	369.7	575.5	492.2	✓
		3366	83.0	160.7	365.2	567.5	450.3	✓
		3406	84.9	156.8	356.8	554.1	388.3	✓
		3423	84.0	154.6	354.9	552.3	412.5	✓
Well 12	3250	3277	79.7	411.9	555.5	688.1	589.2	✓
		3315	78.5	415.6	551.3	689.7	619.8	✓

(Continued on next page)

Table 5. (Continued)

Well	Last test	Days	Water-cut (%)	Percentile P-10	Percentile P-50	Percentile P-90	Production oil flow rate	Forecast
		3349	76.5	413.8	553.7	687.3	681.0	✓
		3441	79.7	416.9	550.8	689.4	584.1	✓
Well 13	2610	2667	68.9	181.3	294.7	485.6	356.8	✓
		2730	75.0	178.7	291.1	484.9	294.3	✓
		2805	73	174.5	288.5	478.3	317.1	✓
		2827	72.6	173.7	285.6	480.9	329.4	✓

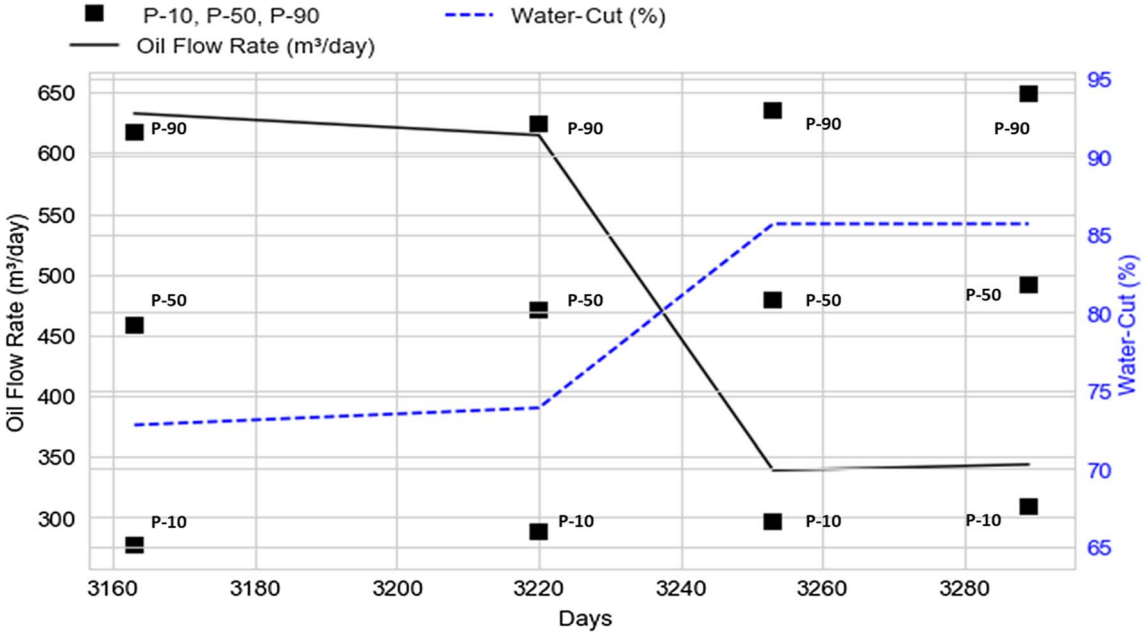

Fig. 10. Oil flow rate forecast results for Well 05.

the production flow rate could not be predicted, as the production values were higher than P-90 probability, as illustrated in Figure 10.

This phenomenon was influenced by the water-cut values (Fig. 11) which decrease more than expected such that the oil production interval could not enclose the value of the production flow rate. In other words, the maximum deviation sampled by Latin Hypercube method was lower than the water-cut production value such that the value of the oil flow rate was higher than P-90 percentile.

However, it is important to highlight the importance of considering uncertainties for the prediction of oil flow rate. While the interval of water-cut variation could be considered small (~10%), its impact on the oil flow rate exhibited a significant variation of ~50%.

Figure 12 illustrates the results for Well 06. The forecast exhibited excellent results for three of the four predicted dates. However, only the first predicted date could not be made.

In this case, the production oil flow rate value in the first date was lower than P-10 percentile. This was not caused by a change in water-cut value, as it did not exhibit a significant variation between the four dates.

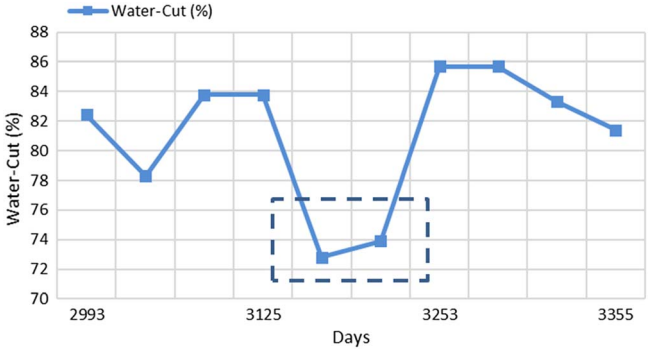

Fig. 11. Trend of water-cut values for Well 05.

Thus, it is necessary to investigate other operational variables such as GOR, gas lift injection rate, and wellhead pressure that may have caused the decrease in the

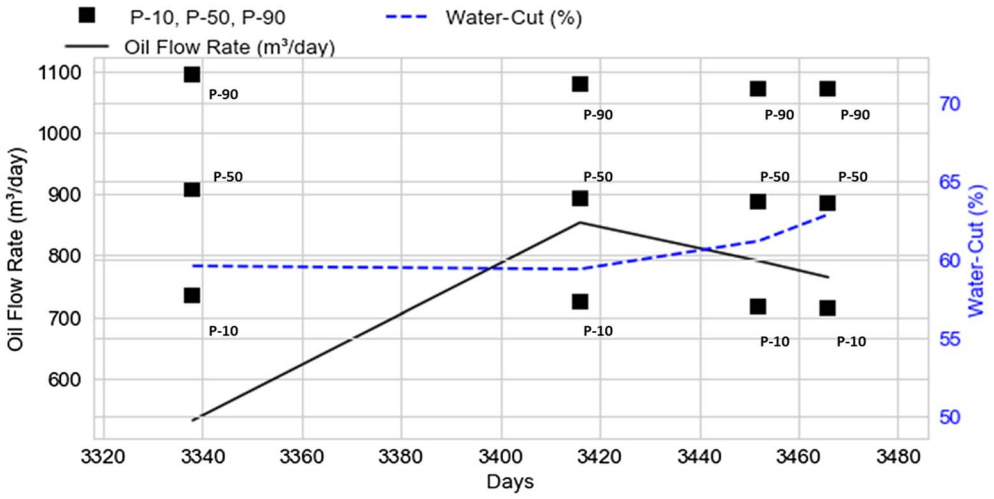

Fig. 12. Oil flow rate forecast results for Well 06.

production flow rate. However, if other production variables could explain this phenomenon, there is an indication that the oil flow rate measurement is wrong, thereby necessitating another round of production tests.

5 Conclusion

The purpose of this study was to develop an analysis tool that could characterize, analyze, and quantify the uncertainties of operational variables and propagate them for the prediction of oil production flow rate.

The water-cut was selected as the uncertainty variable owing to its importance in oil production process and its high significance in oil flow rate response.

Hence, data treatment for water-cut uncertainties is crucial because the removal of outliers will reduce the uncertainty of input data for multiphase flow modeling simulators. Additionally, as the water-cut trend changes with production time, the choice of a more appropriate model to represent this trend changes are necessary for better uncertainty analyses.

Segmented regression model presented good results when applied in the statistical analysis, so as the nonlinear model SVR especially when linear and polynomial regressions not presented satisfactory results.

The comparison between MCS and LHS methods allowed to choose the second methodology. Both methods were equivalent as no significant difference was observed when their histogram distributions were compared. Meanwhile, LHS had a shorter computational time as it required a smaller set of data for accuracy, hence its selection.

The prediction of the oil flow rate was a probabilistic range of occurrence rather than a single deterministic value, owing to the methodology applied. This makes it possible to determine if a new test value is beyond the expected range.

In addition, the results of the forecast is effective as a support tool once the actual value is beyond the expected prediction range, because it is possible to observe significant changes in the water-cut value or other production variables that caused the change in the oil flow rate.

However, operational data have problems that need to be interpreted along with the statistical analysis. The reduction of these errors is of great importance to the industry as any improvement in the operational data collection can cause a huge benefit in oil production curves.

Finally, the uncertainty analysis tool could be used for production optimization, to study the influence of uncertainties on the optimal decisions. Gas lift allocation problem, well routing, well opening, subsea layout arrangement, and many other problems that occur with the simulations of some operational variable could be improved if uncertainties analyses were considered.

References

Balaji K., Rabiei M., Suicmez V., Canbaz H., Agharzeyva Z., Tek S., Bulut U., Temizel C. (2018) Status of data-driven methods and their applications in oil and gas industry, *80th EAGE Conf. Exhib.*, 11–14 June, Copenhagen, Denmark, Society of Petroleum Engineers.

Carpenter C. (2014) Uncertainty evaluation of wellbore-stability-model predictions, *J. Pet. Technol.* **66**, 1, 91–92. doi: 10.2118/0114-0091-JPT.

Chang C.-P., Lin Z.-S. (1999) Stochastic analysis of production decline data for production prediction and reserves estimation, *J. Pet. Sci. Eng.* **23**, 149–160. doi: 10.1016/S0920-4105(99)00013-3.

Charles T., Guemene J.M., Corre B., Vincent G., Dubrule O. (2001) *Experience with the Quantification of Subsurface Uncertainties*, *SPE Asia Pacific Oil and Gas Conference and Exhibition*, 17–19 April, Jakarta, Indonesia, Society of Petroleum Engineers. doi: 10.2118/68703-ms.

Corre B., Thore P., Feraudy V.De, Vincent G., Elf T. (2000) Integrated uncertainty assessment for project evaluation and risk analysis, *SPE European Petroleum Conference*, 24–25 October, Paris, France, Society of Petroleum Engineers.

Costa L.A.N., Maschio C., Schiozer D.J. (2019) Evaluation of an uncertainty reduction methodology based on Iterative Sensitivity Analysis (ISA) applied to naturally fractured reservoirs, *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles* **74**, 40. doi: 10.2516/ogst/2019013.

Dejean J., Blanc G. (1999) Managing uncertainties on production predictions using integrated statistical methods, *SPE Annual Technical Conference and Exhibition*, 3-6 October, Houston, Texas, Society of Petroleum Engineers. doi: 10.2523/56696-ms.

Delignette-Muller M.L., Dutang C. (2015) fitdistrplus: An R package for fitting distributions, *J. Stat. Softw.* **64**, 1–34. doi: 10.18637/jss.v064.i04.

Feraille M., Marrel A. (2012) Prediction under uncertainty on a mature field, *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles* **67**, 193–206. doi: 10.2516/ogst/2011172.

Fonseca Junior R.D., Gonçalves M.D.A.L., Azevedo L.F.A. (2009) Consideration of uncertainty in simulations of elevation and flow, *An. do IV Semin. Elev. Artif. e Escoamento* (in Portuguese: Consideração de Incerteza nas Simulações de Elevação e Escoamento).

Goda T., Sato K. (2014) History matching with iterative Latin hypercube samplings and parameterization of reservoir heterogeneity, *J. Pet. Sci. Eng.* **114**, 61–73. doi: 10.1016/j.petrol.2014.01.009.

Grimstad B., Gunnerud V., Sandnes A., Shamlou S., Skrondal I.S., Uglane V., Ursin-Holm S., Foss B. (2016) A simple data-driven approach to production estimation and optimization, *SPE Intell. Energy Int. Conf. Exhib.*, 6–8 September, Aberdeen, Scotland, UK. doi: 10.2118/181104-MS.

Iglewicz B., Hoagling D. (1993) Volume 16: How to detect and handle outliers, *ASQC Basic Ref. Qual. Control Stat. Tech.*

Jahanandish I., Salimifard B., Jalalifar H. (2011) Predicting bottomhole pressure in vertical multiphase flowing wells using artificial neural networks, *J. Pet. Sci. Eng.* **75**, 336–342. doi: 10.1016/j.petrol.2010.11.019.

Mahjour S.K., Correia M.G., de Souza dos Santos A.A., Schiozer D.J. (2019) Developing a workflow to represent fractured carbonate reservoirs for simulation models under uncertainties based on flow unit concept, *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles* **74**, 15. doi: 10.2516/ogst/2018096.

Maschio C., de Carvalho C.P.V., Schiozer D.J. (2010) A new methodology to reduce uncertainties in reservoir simulation models using observed data and sampling techniques, *J. Pet. Sci. Eng.* **72**, 110–119. doi: 10.1016/j.petrol.2010.03.008.

Meyer D., Dimiitriadou E., Kurt H., Weingessel A., Leisch F., Cang C.-C., Lin C.-C. (2015) *Misc functions of the department of statistics, probability*, TU Wien, Vienna, Austria.

Monteiro D.D., Ferreira Filho V.M., Chaves G.S., De Santana R.S., Duque M.M., Granja-Saavedra A.L., Baioco J.S., Vieira B.F., Teixeira A.F. (2017) Uncertainty analysis for production forecast in oil wells, *SPE Lat. Am. Caribb. Pet. Eng. Conf.*, 17-19 May, Buenos Aires, Argentina. doi: 10.2118/185550-MS.

Morita N. (1995) Uncertainty analysis of borehole stability problems, *SPE Annual Technical Conference and Exhibition*, 22–25 October, Dallas, Texas, Society of Petroleum Engineers, 533–542. doi: 10.2118/30502-ms.

Muggeo V.M.R. (2008) Segmented: An R package to fit regression models with broken-line relationships, *R News* **8**, 20–25. doi: 10.1002/sim.1545.

Niño F.A.P. (2016) Wellbore stability analysis based on sensitivity and uncertainty analysis, *SPE Annual Technical Conference and Exhibition*, 26–28 September, Dubai, UAE, Society of Petroleum Engineers.

Sales L.P.A., Pitombeira-Neto A.R., de Athayde Prata B. (2018) A genetic algorithm integrated with Monte Carlo simulation for the field layout design problem, *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles* **73**, 24. doi: 10.2516/ogst/2018017.

Schiozer D.J., de Souza dos Santos A.A., de Graça Santos S.M., von Hohendorff Filho J.C. (2019) Model-based decision analysis applied to petroleum field development and management, *Oil Gas Sci. Technol. - Rev. IFP Energies nouvelles* **74**, 46. doi: 10.2516/ogst/2019019.

Sheng Y., Reddish D., Lu Z. (2006) Assessment of uncertainties in wellbore stability analysis, in: Wu W., Yu H.S. (eds), *Modern trends in geomechanics*, vol. **106**, Springer, Berlin, Heidelberg.

Shields M.D., Zhang J. (2016) The generalization of latin hypercube sampling, *Reliab. Eng. Syst. Saf.* **148**, 96–108. doi: 10.1016/j.ress.2015.12.002.

Spesivtsev P., Sinkov K., Sofronov I., Zimina A., Umnov A., Yarullin R., Vetrov D. (2018) Predictive model for bottomhole pressure based on machine learning, *J. Pet. Sci. Eng.* **166**, 825–841. doi: 10.1016/j.petrol.2018.03.046.

Tyler K., Sandsdalen C., Maeland L., Aasen J.O., Siring E., Barbieri M. (1996) Integrated stochastic modeling in reservoir evaluation for project evaluation and risk assessment, *SPE Annual Technical Conference and Exhibition*, 6–9 October, Denver, Colorado, Society of Petroleum Engineers. doi: 10.2118/36706-ms.

Vapnik V. (1995) *The nature of statistical learning theory*, Springer-Verlag New York. doi: 10.1007/978-1-4757-3264-1.

Zabalza-Mezghani I., Manceau E., Feraille M., Jourdan A. (2004) Uncertainty management: From geological scenarios to production scheme optimization, *J. Pet. Sci. Eng.* **44**, 11–25. doi: 10.1016/j.petrol.2004.02.002.