Optimizing initial oil production of horizontal Wolfcamp wells utilizing data analytics

A. Alzahabi¹ · A. Alexandre Trindade² · A. Kamel¹ · A. Harouaka¹

Received: 9 April 2020 / Accepted: 3 June 2020 / Published online: 14 June 2020
 © The Author(s) 2020

Abstract

Machine learning techniques have fundamentally altered how oil and gas industry practitioners design fracture operations. In this paper, we perform data analytics utilizing response surface methodology (RSM), a group of statistical techniques that develop a functional relationship between an output variable of interest and several associated input variables, to optimize the output. We apply RSM to optimize horizontal well production based on initial production (IP) of horizontal oil wells for 180 days (IP180 Oil), as a function of five input variables: reservoir type, fracturing fluid (gal/ft), proppants (lbm/ft), cluster spacing, and stage length (ft). The RSM model correlates the initial production of each well to the input variables via a single equation, thus allowing for exploration of the fitted response surface in order to maximize production. Although the choice of the five inputs is made based primarily after consultation with industry professionals, we validate our selection by also applying an assortment of data-analytics-based methods that attempt to rank variable importance and thereby identify completion variables that may be predictive of initial production. The findings rank all five variables above the 50th percentile, thus indicating that the chosen variables have merit. This procedure is applied to a dataset of 201 horizontal wells from the Wolfcamp formations. The model fits reasonably well, with $R^2 = 61\%$, a very significant F-statistic p value, and a predicted versus observed values scatterplot indicating a good fit. The RSM analysis suggests that, within the feasible space defined by this dataset, maximum values of IP 180 Oil may be obtained by setting the fracturing fluid in gal/ft at approximately 1972, while simultaneously maximizing the remaining input variables (proppant loading, cluster spacing, stage length). The outcome indicates the possible directions to be taken in seeking a global optimum initial production for the setting of completion variables. Iteration of this scheme may lead to a near-optimum global solution. The real utility of this work may be indicating the way different studies may be designed to optimize production, each with its own selection of inputs, and ultimately be combined in a meta-analysis.

Keywords Initial oil production · Wolfcamp · Horizontal wells · Completion variables · Data analytics · Response surface methodology

List of symbols

K_f	Fracture permeability	IP	The amount of oil produced by a new oil well, measured in B/D (barrels of oil per day) or BOE/D (barrels of oil equivalent per day)
w_{fp}	Propped fracture closed width	Proppant Lbs.	Mass of proppants
		Age	The difference in number of days between the time the well was completed and 01/01/2001
		Avg. Prop. Concentration	The lbs of proppant pumped per fluid gallons pumped
		Avg. Rate	The average rate at which the mixture is pumped downhole to create the fractures

✉ A. Alzahabi
 alzahabi_a@utpb.edu

¹ Petroleum Engineering Department, University of Texas-Permian Basin, 11105 West Highway 191, Midland, TX 79705, USA

² Department of Mathematics and Statistics, Texas Tech University, Broadway and Boston, Lubbock, TX 79409, USA

Cluster spacing	The distance between the clusters within the stage	Perfs/Clusters	The number of perforations used in each cluster
Clusters per stage	The number of clusters used in each stage	Perfs	The number of holes created from the charges that are inserted downhole in sets of perforations called clusters designed in different patterns to achieve optimal completion
Clusters	A set of perforations arranged in a certain pattern to achieve the optimal completion		
Comp Date	The date the well was completed	Prop. Lbs./Cluster	The amount of proppant pumped per cluster
Completed feet	The calculated distance between the top perf and bottom perf where the fractures occur	Prop. Lbs./Ft.	The amount of proppant pumped per completed foot
County Variable	A numeric variable to distinguish between counties	Prop. Lbs./Perf	The amount of proppant pumped per perforation
County	The county where the well was drilled	Prop. Lbs.	The amount of proppant pumped with the fluid to keep the created fractures open
Fluid Bbls	The amount of fluid pumped downhole to initiate fracture and place proppant	Rate/Cluster Rate/Ft.	The average rate per cluster The average rate per completed foot
Fluid Gal/Cluster	The amount of fluid pumped per cluster	Rate/Perf Fluid Gal/Ft.	The average rate per perf Fluid volume in gallons per foot
Fluid Gal/Ft.	The amount of fluid pumped per completed foot	Reservoir	The formation in which the lateral was drilled (i.e., Wolfcamp A, Wolfcamp LA, Wolfcamp MA)
Fluid Gal/Perf	The amount of fluid pumped per perforation		
IP	Initial production rates	Reservoir Variable	A numeric variable distinguishing between reservoirs Wolfcamp C–D and Wolfcamp A
IP 30, 90, & 180	Calculations taking a rolling average by number of days described (30, 90, or 180) and then using the maximum value obtained for oil, gas and water	Reservoir	The formation that the lateral was drilled (i.e., Wolfcamp A, Wolfcamp LA, Wolfcamp MA)
ISIP/Ft.	Instantaneous shut-in pressure once a stage has been completed and frac pumps are shut down	Slickwater	Water with chemicals added to speed up the rate at which it can be pumped to create more fractures
Linear Gel	A fracturing fluid supplemented with different polymers which increase its ability to carry proppant	Stage length	The length of each stage, a good indicator for normalizing stages for lateral length
Max Prop. Concentration	The proppant concentration begins with mostly fluid and then is built up to a maximum concentration of pounds of proppant per fluid gallon	TVD	The furthest vertical depth drilled
Max. Rate	The maximum rate achieved pumping mixture into fractures	UWI	A unique well identifier for every well; each set of numbers stands for a unique location and well identifier (i.e., county, state, horizontal, pilot)
Number of stages	Whenever a plug is set, perforations are created, the reservoir is fractured with fluid, and then another plug is set, it is called a stage	GOR Ratios	A metric used to determine the amount of gas produced per oil produced (SCF/STB)

Yield	Condensate yield, MMSCF/STB
WCUT	Water cut; a metric used to determine the amount of water produced per oil produced. (Water/(Water + Oil))
Oil, Gas, MBOE EUR	The estimated ultimate recovery of oil and gas, MBOE – Oil + Gas/6
Degradation	The amount of overlap divided by total rectangular area

Introduction

Challenges to oil production in the Wolfcamp range from low productivity and quick decline to high variability in the response of wells due to fracture treatment. Initial production of wells is used as a measure to assess well performance and completion effectiveness, and to compare wells of a given company to those of its competitors, and production performance of one shale play to that of another. However, the challenge remains to determine which of these initial production metrics should be used. The O & G industry recognizes peak rate (the starting point of production decline); initial production (IP30) as a rolling average for 30-day

production; and IP60, IP90, IP180, and IP365 as a measure of production performance. Some schools of thought use cumulative production in the first months along with EUR to judge performance and profitability of oil wells.

A key question is, what are production metrics?

Production metrics may be represented in initial production (IP) IP30, through IP365 initial production for a rolling average of 30 through 365 days; cumulative production for the first number of days or months of the life of the well, and Estimated Ultimate Recovery (EUR). The production metrics listed in Table 1 show benefits and challenges associated with each.

The IP rate of a well is established based on well type; reservoir considerations; historical and analogous performance; management and commercial strategy; operational considerations; target and economic reserves, and nodal analysis results (Olufemi and Ambastha 2017).

Empirical techniques to forecast oil production include decline curve analysis and capacitance/resistance modeling (CRM). CRM is a data-driven method for characterizing conventional reservoirs and optimizing oil production without complex reservoir simulations (Yousef et al. 2006).

Recent industry efforts established multistage fracture design optimization for unconventional shale oil and gas basins to locate the number of fracture stages that maximize NPV (Rashid et al. 2014). The effect of both spacing

Table 1 Oil and gas production metrics

Metrics	Scale of opportunity	Benefits	Challenges
Peak rate	Starting point of production decline	Comparative scale of wells	Not dependent on time needed to reach the peak Volume recovered is not correlated Highly affected by constraints (choke) Reported peak may not be the actual peak
IP 30	Peak 30-day average of production routine and quick scale	Earlier scale	Not representative of well performance due to flow-back period and clean-up process
IP60	Peak 60-day average of production	Normalizes peak rates to a 60-day average, taking out some early time production issues and draw-down	Does not take into account entire well life
IP90	Peak 90-day average of production	Normalizes peak rates to a 90-day average, taking out some early time production issues and draw-down	Does not take into account entire well life
*IP180	Peak 180-day average of production	Normalizes peak rates to a 180-day average, taking out some early time production issues and draw-down	Does not take into account entire well life
EUR	Quantity of oil or gas that is potentially recoverable or has already been recovered from a reserve or well	Excellent measure of a well's overall performance based on well production and offset well production	Difficult to estimate early time response
Cumulative production in <i>N</i> months	Quantity of total production accumulated to a certain point in time	Normalizes for different draw-down techniques and provides an excellent representation of overall well performance	NA

*In this work we will use (IP180) as representative of well initial performance because it extends time beyond flowback and cleanup periods

and perforation friction on the propagation of concurrent hydraulic fractures from a horizontal well was studied (Izadi et al. 2015). The optimization indicated that while limited-entry perforating could help equalize the lengths of fractures simultaneously growing from multiple clusters, it did not equalize their widths, but equalized placement of proppant across the entire series of perforation clusters.

The data-based technique, which relies on empirical correlations, also correlates the well production to reservoir characteristics. Due to the currently used downhole tool restrictions currently in place, multistage fracturing is often conducted consecutively, stage by stage from well toe to heel without real optimization (Xia et al. 2016; Roussel and Sharma 2011). Each stage usually contains several perforation sets, and thus, several fractures propagate simultaneously if each perforation set produces only one fracture (Wu and Olson 2014).

There difference is significant between data-driven reservoir modeling and empirical technologies. First, a model with a good predictivity is necessary; then internal interactions between the variables can be checked to see if the model makes sense. The physical and geological interactions can tell the prediction consistency of the built model (Mohaghegh 2017). A carefully worked-out effort to develop the unconventional reservoirs may change the trend of manufacturing approach (uniform spacing of wells). Utilization of stimulation software such as “Kinetix-Intersect” and comparison of the modeling outcome of fracture simulator for a selected fracture stage and rate-transient analytical results led to variable Enhanced Oil Recovery (EOR) (Rodriguez 2019). Acknowledgment of variable recovery and pressure depletion associated with the producers’ drainage volume should be the basis for the planning of EOR techniques (Rodriguez 2019). Spacing and placement of perforation clusters, stress distribution, and fracture mechanics are the most important input variables of production performance (Cheng 2012). A case study implemented an engineering approach to optimize multiple sets of parallel horizontal wells in the oil segment of the Eagle Ford Shale. The data was used to design stimulation-stage intervals and evaluate surface injection responses during fracture treatments (Stegent et al. 2013).

The correlation between the style of completions and the initial production rate is well established (Griffin et al. 2013; Lafollette et al. 2012a, b). Initial production is a crucial metric to measure profitability of wells. Estimated ultimate recovery of wells (EUR) usually increases with the number of stages. As the number of fracture stages increases, the efficiency of incremental stages decreases in the Bakken Shale formation (Ran and Kelkar 2015). Beyond a certain number, therefore, the incremental cost would exceed the incremental benefits. All the efforts led to a good understanding of the effect of individual variables on production

without agreement on an optimum methodology. Microseismic monitoring in the unconventional monitoring has proven to be highly effective in many field runs (McClure 2012).

Table 1 lists the benefits and challenges of commonly used production metrics.

Why do O&G operators need production metrics?

- To measure performance of new wells versus old wells.
- To compare their own wells versus competitors’ wells, shales, plays.
- To correlate production metrics with the EUR of wells.

Hydraulic fracture operations lead to an increase in production. This increase may depend on various metrics (shown in Table 2): completed length of the horizontal well; quantity of perforations; number of clusters per stage; cluster spacing in length; size of proppant in lbm; concentration of proppant in fluid; fluid in gal/ft; fluids in bbls; average injection rate; fluid type, and proppant type and size, along with production metrics (e.g., Initial Production, Yield).

The data ranges used to develop the model are summarized in Table 2, which also represents completion variables that might affect initial production performance of horizontal wells.

The paper is organized as follows:

- Description of problem
- Data analytics methodology to select optimum completion variables
- Development of model
- Testing and validation of model.

For lack of standard metrics, optimum completion variables are determined by trial and error approach (e.g., amount of fluid and proppant pumped per completed foot, cluster spacings and fracture stage length). One unanswered question is optimum number of completion variables needed per well for highest initial production. Another area of uncertainty is the degree to which each metric affects estimated ultimate recovery (EUR) from an individual well. One school of thought suggests the more, the better. Another suggests that fracture length and dimensionless fracture conductivity are the two primary variables that control the productivity index of a fractured well and thus initial production. For years, industry researchers have applied accepted methods to estimate post-fracture production increase. These approaches depend on a variety of fracture lengths and fracture conductivities. The procedures involve a combination of propped fracture lengths and conductivities that optimize post-fracture production response.

Graphical solutions rely on the interdependence between formation permeability, half-length of the fracture, and effective conductivity ($K_f \times w_{fp}$). Veatch et al. (2017) showed

Table 2 Range of completion parameters used to develop the model

No.	Completion metrics	Parameters
1	Completed feet	3600–11,210
2	Number of stages	10–58
3	Stage length (ft.)	156–404
4	Perforations per clusters	3–18
5	Perforations	408–2192
6	Clusters	30–798
7	Cluster spacing(ft.)	13–153
8	Fluid BBLs	60–655,118
9	Fluid Gal/Ft.	27,400–180,378
10	Proppant Lbs.	2,677,278–28,539,969
11	Proppant Lbs./Ft.	724–2839
12	Proppant Lbs./Clusters	35,100–133,759
13	Proppant Lbs./Perfs	3385–21,542
14	Avg. Prop. concentration (ppg)	1–209
15	Avg. rate (bpm)	38–98
16	Avg. pressure (psia)	4208–10,289
17	Max. pressure (psia)	6959–12,567
18	Initial shut-in pressure (ISIP)/ft	0.91–170
19	Fluid type	Slickwater, linear gel, composite fluid, XL borate, hydrochloric acid, mud acid
20	IP production variables (IP oil) (BBL/day)	130–3261
21	GOR (scf/STB)	1130–50,000
	Ip 30 Yield (BBLs)	37–4870
	WCUT	0.4–0.9
22	ROA (rectangular overlap area) between wells (ft ²)	0–1,590,320
23	Initial reservoir pressure (psia)	3900–8900

five commonly used charts for predicting semi-steady-state post-fracture production fold of increase, i.e., the ratio of production rate post-fracture cleanup to pre-fracture rate. The methods include Prats; McGuire and Sikora; Holditch; Tinsley et al., and Tannich and Nierode graphs. These charts may provide guidelines for determining approximate treatment size (fluid systems and proppants), injection rates and proppant staging schedules (Veatch et al. 2017). Economic analysis based on IP alone can be misleading for gas producers considering lateral placement in the reservoir at the bottom or midway (Taylor et al. 2011). On the other hand, ultimate production (EUR) from Marcellus wells was precisely predicted using initial well production (Male et al. 2016). The wells in the study, which showed a boundary-dominated flow, were used to forecast EUR of 5275 wells in Pennsylvania and West Virginia over a period of 25 years.

This paper focuses on use of data from the Wolfcamp play of the Permian Basin to introduce an effective model to predict initial production as a production performance metric from completion variables. The work utilizes data analytics for effective selection of key completion

variables that lead to optimum IP. For a list of various analytics tools used in unconventional reservoirs and development, see Table 6 in “Appendix 2”.

As far as we know, no simple model is available to estimate production metrics utilizing key completion variables such as fluid in gal/ft, cluster spacing, proppant in Lbs./ft, reservoir type and stage length of horizontal wells. For this paper, we are optimizing the initial production of oil for 180 days (IP 180 Oil).

Methodology

In this section, we first describe data sources, as well as the selection of input variables for optimizing IP 180 Oil. Summary measures commonly used in the machine learning literature to rank the relative merits of predictor variables are computed. Finally, a quadratic response surface model is fitted.

Data and nomenclature

The dataset consisted of measurements taken from 201 horizontal wells. The 209 originally listed variables of particular interest were analyzed. A list of 32 variables was chosen as given in Table 3. From this list, the following variables were chosen after the analyses described in the next subsection.

1. Oil: natural logarithmic of IP 180 of Oil (output variable). Range of values: 4.860–8.011.
2. Reservoir type: reservoir name (acts as blocking factor with 9 levels). Levels: WOLFCAMP, WOLFCAMP A, WOLFCAMP C, WOLFCAMP D, WOLFCAMP D4, WOLFCAMP LA, WOLFCAMP MA, WOLFCAMP UA, WOLFCAMP UAY.
3. Fluid.gal/ft: amount of fluid pumped per completed foot. Range: 643–2687.

4. Prop.lbs/ft: amount of proppant pumped per completed foot. Range: 724–2839.
5. Cluster spacing: distance between the clusters within the stage. Range: 13–132.
6. Stage length: length of each stage. Range: 156–404.

Selection of input variables

The selection of the five input variables discussed above is based primarily on expert opinion after extensive consultations with industry professionals. As a form of validation, we sought also a variety of purely machine-learning-based methods to rank the relative variable importance in predicting IP 180 Oil. The Random Forest Package in R, which fits regression tree predictive models, provides two such measures (Liaw and Weiner 2018): %IncMSE (Fig. 1) and IncNodePurity (Fig. 2). The third measure is the absolute value of the (Pearson) correlation coefficient of each variable with IP 180 Oil (Fig. 3).

Table 3 List of chosen 32 variables

Age of well on production since completion
Reservoir type, (the formation in which the lateral was drilled (i.e., Wolfcamp A, Wolfcamp LA, Wolfcamp MA))
Well total vertical depth, ft
Completed feet
Number of stages
Stage length
Clusters per stage
Cluster spacing
Perfs per cluster
Total clusters
Total perfs
Injected fluid, Bbls
Fluid volume in gal/ft
Fluid in gal/cluster
Fluid in gal/perf
Proppants in Lbs
Prop. Lbs/Ft=the amount of proppant pumped per completed foot
Prop. Lbm/Cluster=the amount of proppant pumped per cluster
Prop. Lbs/perf=the amount of proppant pumped per perforation
Avg. prop. concentration=the Lbs of proppant pumped per fluid gallons pumped
Maximum prop. concentration=the Lbs of proppant pumped per fluid gallons pumped
Average rate at which the mixture is pumped downhole to create the fractures
Maximum rate at which the mixture is pumped downhole to create the fractures
Rate per cluster
Rate per perf
Avg pressure
Max pressure
IP 180 oil (initial production of oil rolling average over 180 days)
Initial.producing.OGR (oil gas ratio)
Initial.producing.GOR (oil gas ratio)
Critical pressure
Initial reservoir pressure

Fig. 1 The first measure of importance calculated from regression trees

Fig. 2 The second measure of importance calculated from regression trees

Note that, unlike the regression tree-based methods, correlations cannot be calculated between a categorical variable (in this case reservoir type) and a quantitative output.

We note that the five selected input variables, although not at the very top of these various rankings, do come out as being important. Namely: Reservoir is ranked 3rd in Fig. 1; Fluid gal/ft is ranked 12th in Fig. 3; Prop.lbs/ft; is ranked 7th in Fig. 3; Cluster spacing is ranked 9th in Fig. 3; Stage length is ranked 19th in Fig. 3.

Fitting a response surface model

Response surface methodology (RSM) is a group of statistical procedures used in the development of a flexible functional relationship between a relevant output variable, y , and a number of related input variables, x_i , with a view toward optimization of the output (Khuri and Mukhopadhyay 2010). If we have just one categorical input variable x_0 with L levels, and k numerical input variables $x_1, \dots, x_k$, such a relationship is usually approximated by a low-degree polynomial model of the form (Eq. 1) listed in “Appendix 4”. Since the model is of the form of a multiple linear regression, it is straightforward to estimate the variables via least-squares and assess its goodness of fit. Denoting these estimates by $\hat{B}_i$ and $\hat{B}_{i,j}$, the resulting fitted model at a given value of the vector of input variables $\mathbf{x}$ is given by Eq. 2 listed in “Appendix 4”, which is the optimal predictor of y at $\mathbf{x}$ (in the least-squares sense).

The main goal of RSM is determination of the optimum model-predicted output $\hat{y}(\mathbf{x})$, over all $\mathbf{x}$ in the data region D . As a result and for our data, this region would be the 4-dimensional rectangle delineated by the intervals (see above summary statistics):

$$643 < \text{fluid.gal/ft} < 2687$$

$$724 < \text{prop.lbs/ft} < 2839$$

$$13 < \text{cluster spacing} < 132$$

$$156 < \text{stage length} < 404$$

For ease of interpretation, we normalize these variables by subtracting the midpoint of the interval and dividing the result by the interval half-width:

$$x_1 = \frac{\text{fluid.gal/ft} - 1665}{1022}$$

$$x_2 = \frac{\text{prop.lbs/ft} - 1781.5}{1057.5}$$

$$x_3 = \frac{\text{cluster spacing} - 72.5}{59.5}$$

$$x_4 = \frac{\text{stage length} - 280}{124}$$

whence the (normalized) data region D_{norm} becomes the 4-dimensional cube delineated by the intervals $-1 < x_i < 1$,

Fig. 3 Abs correlations of IP 180 Oil with each of the 30 variables used in the study

for $i = 1, \dots, 4$. On this normalized scale, the fitted model coefficients are given in Table 5 of “Appendix 1”.

This model fits reasonably well when applied to our dataset with a (typical) quadratic order polynomial, resulting in $R^2 = 61\%$ and an F-statistic p value < 0 . The predicted versus observed value plot in Fig. 4 shows relatively good in-sample predictions.

Results and discussion

The selection of an *appropriate* pool of input (or predictor) variables is a perennial problem in supervised machine learning applications. One may begin to understand the difficulty of the issue by realizing that the exact functional

Fig. 4 Visualization of the model fit, predicted versus measured log (IP 180 Oil) of the Wolfcamp horizontal wells. The diagonal line indicates where the predicted and observed log IP 180 Oil values would be equal

Table 4 Coordinates of the optimum point in the data region

Fluid gal/ft	Prop lbs/ft	Cluster space, ft	Stage length, ft	Optimal value of oil (log (IP 180)) at maximum point
1971.6	2839 (= max)	132 (= max)	404 (= max)	9.7 (for reservoir WOLFCAMP)

relationship between a predictor and the outcome of interest is (and forever will be) unknown. Moreover, predictors are typically correlated, thus further confounding the issue of what is predicting what. And how should one define “appropriate”? In classical regression analysis (Neter et al. 2004), one can begin to untangle these inter-related issues by measuring the decrease in prediction error variance (the so-called MSE or “mean squared error”) when each variable by itself is added to the model in turn: important predictors should lead to large decreases in MSE. But this is not the full story, as the results depend on assumed functional relationships (typically linear), and the amount of “dependence” (correlation) among the predictors.

The selection of the five input variables listed in the previous section that were chosen for this study, was based primarily on expert opinion after extensive consultations with industry professionals. As a form of validation, the three metrics displayed in Figs. 1, 2 and 3 provide different instances of purely machine-learning-based rankings of importance. The two tree-based measures, %IncMSE (Fig. 1) and IncNodePurity (Fig. 2), consist of drop-in-MSE type of metrics that use frequency of splits along the variable to determine rank (see “Appendix 3” for a summary). The measure in Fig. 3 is the usual correlation coefficient. A plethora of other measures could be used, e.g., one of the many statistical variable selection information-based criteria, such as AIC or BIC, in connection with the fitted RSM model. Although the five inputs we chose are not the top 5 by any of these measures, they do rank in the top 50th% according to all of them. In any case, the aim of the present study is not to definitively “solve” this difficult ranking problem, but rather to sketch out a procedure for optimizing production that could be implemented following the selection of a few inputs based on pragmatic choices borne out of field experience.

Translation of this optimization goal into the RSM of “Methodology” section, converts into determination of the maximum model-predicted output $\hat{y}(\mathbf{x})$, over all $\mathbf{x}$ in the data region D . Interestingly, the optimal $\mathbf{x}$ is obtained by maximizing all the variables, except fluid in gal/ft as shown in Table 4. Note that different reservoir types simply act as an intercept in the model, so that anything beyond WOLFCAMP increases this value; e.g., the largest increase is 1.0436 for WOLFCAMP UAY (see Table 5 of “Appendix 1”). Thus the location of the optimal point is unaffected

by reservoir type, and all discussion and plots pertain to WOLFCAMP.

Figure 5 shows contour plots of $\hat{y}(\mathbf{x})$ in the vicinity of this optimal point (red dot). The variables not displayed on the axes are fixed at their optimal point values. Overlaid are the data points displayed in blue. The top left panel would seem to indicate that the optimal point is in the vicinity of the data, an error that is dispelled in the bottom left panel where the red dot is actually seen to be quite far from the data cloud.

Finally, Fig. 6 is a zoomed-out version of Fig. 5 plotted on the normalized variable scale. The data region D_{norm} is displayed as a red square, where the optimal point is still the red dot.

From this we can see that the red dot is in a flat region, thereby suggesting that no dramatic improvements appear to be feasible beyond this point. (The fact that the global optimal solution simultaneously sets *all* inputs at infinity is not of interest here).

Summary

This paper devises an optimization procedure to maximize the initial production of horizontal wells considering a pool of reservoir variables, summarized as follows:

1. Choose an initial oil production metric, e.g., IP 180 Oil.
2. Select a pool of “important” predictors/variables, e.g., number of days in production; horizontal well completion configurations; proppant size; fluid type; stages count, and lateral completed interval, etc. This selection could (perhaps ideally) be made by combining pragmatic field experience with state-of-the art data analytics.
3. Use RSM to model the functional relationship between the pool of predictors and the metric.
4. Based on the fitted RSM model, determine the optimum model-predicted output $\hat{y}(\mathbf{x})$, over all $\mathbf{x}$ in the feasible data region.

Response surface methodology suggests that, within the feasible space defined by this dataset, maximum values of IP 180 Oil may be obtained by setting fluid in gal/ft at approximately 1972, while simultaneously maximizing the remaining inputs (prop in lbs/ft, cluster spacing, stage length).

Fig. 5 Contour plots of log (IP 180) versus 2D-slice of various input variables

Note that this statement is not to be taken as absolute truth, but rather as suggestive of possible directions to be taken in seeking a global optimum value for the setting of inputs. More refined estimates may be obtained by collecting new data in the vicinity of this optimum, followed by a new analysis. Iteration of this scheme may lead to a near-optimum global solution.

The value of the new model proposed in the paper lies in its simplicity and the lower number of input variables needed for a practicing engineer to optimize oil production from shale horizontal wells. More importantly, it is based on completion design variables that may be chosen relatively quickly. It is also possible that the IP 180 Oil calculated from this model can be correlated with EUR.

It is projected that the calculated IP 180 Oil from this model may be used in several applications. First, it may work as a comparative scale to measure well performance of a key play versus that of a competitor. Second, it may be used as a tool to enable quick decision on certain completion variables needed for horizontal wells, which can

facilitate further estimation of fluids gal/ft, proppants lbs/feet, cluster spacing and fracture stage length. Third, in case of multiple leases with a variety of IPs, the model may work as a quality check tool to optimize production of a lease. Estimation of IP will be used to differentiate among different reservoirs (Wolfcamp A, B, C, D, MA and LA).

Conclusion and recommendation

The objective of the paper was to propose a methodology for identifying the values of key completion variables that may lead to optimum production of horizontal wells. The study proceeded as follows.

- Extensive literature review to identify all production metrics, understand the relationships between them, compare performance, and suggest which may be used.
- Description of the problem.

Fig. 6 Contour plots of normalized log (IP 180) versus 2D-slice of various input variables

- The usage of data analytics methodology to select optimum values of key completion variables, after elimination of non-effective variables such as Avg. Rate (bpm) from the available pool.
- Testing and validation of the procedure on publicly available data.

The following conclusions may be drawn.

1. Optimal production may be obtained by maximizing with respect to some pairs of variables (e.g., cluster spacing and proppants).
2. Maximum IP over domain is obtained by maximizing all the variables (e.g. proppants lbs/ft of 2839, cluster spacing of 132 ft, and stage length of 404 ft), except fluid in gal/ft.
3. The proposed model may be used to predict initial production of horizontal wells of the Wolfcamp using few input completion variables with reasonable accuracy.
4. Additional data might benefit the work for further testing of global areas of shale plays.

Acknowledgements The authors would like to acknowledge the kind assistance through fruitful discussion with Wade Baustian from Camino Natural Resources and Cimarex for providing the data used in this work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix 1: Model coefficient details

See Table 5.

Table 5 Model coefficients

	Estimate	SE	<i>T</i> value ^a	Pr(> <i>t</i>) (level of significance)
(Intercept)	5.876303	0.207053	28.3807	< 2.2e–16***
Reservoir WOLFCAMP A	0.679090	0.155122	4.3778	2.008e–05***
Reservoir WOLFCAMP C	0.299941	0.170744	1.7567	0.08064
Reservoir WOLFCAMP D	0.440165	0.142056	3.0985	0.00225
Reservoir WOLFCAMP D4	0.294721	0.184463	1.5977	0.11182
Reservoir WOLFCAMP LA	0.725409	0.152426	4.7591	3.923e–06
Reservoir WOLFCAMP MA	0.981948	0.183175	5.3607	2.463e–07
Reservoir WOLFCAMP UA	0.912737	0.161668	5.6457	6.144e–08
Reservoir WOLFCAMP UAY	1.043565	0.222151	4.6975	5.139e–06
x1	–0.222704	0.351589	–0.6334	0.52725
x2	1.228031	0.484916	2.5325	0.01216
x3	0.176712	0.480894	0.3675	0.71369
x4	0.328183	0.352960	0.9298	0.35369
x1:x2	1.874229	1.392720	1.3457	0.18004
x1:x3	–0.515561	1.207539	–0.4270	0.66991
x1:x4	–0.136812	0.937562	–0.1459	0.88414
x2:x3	1.618066	1.074972	1.5052	0.13398
x2:x4	0.052065	0.706816	0.0737	0.94136
x3:x4	0.536093	0.625130	0.8576	0.39225
x1 ²	–1.678645	0.914885	–1.8348	0.06815
x2 ²	0.010602	0.516644	0.0205	0.98365
x3 ²	–0.163539	0.632406	–0.2586	0.79623
x4 ²	–0.121142	0.237280	–0.5105	0.61028

^aA measure of the size relative to 0 scaled by the variance of the data. For instance, large absolute *t* values indicate a large difference from 0

Appendix 2: Table of data analytics use in completion strategies in unconventional resources

See Table 6.

Table 6 Data analytics tools used for unconventional reservoirs

Shale Play	Data Analytic Tool	Findings	References
Permian Basin Wolfcamp and Spraberry shales	Workflow	Examined several well completions aspects and effect on production Denser cluster spacing and shorter stage length show production uplift	Jaripatke et al. (2018)
Utica shale play of Pennsylvania and WV	Data analytic workflow (neural networks)	Tried to understand performance indicators, optimize location selection, well spacing, horizontal drilling and completion designs	Passman (2018)
N/A	Predictive models using proxies Random forest Grad Boost Linear regression Decision tree AdaBoost MF OLS MF NN	Demonstrated a new workflow coupling earth sciences with data analytics to operationalize well completion optimization	Pankaj et al. (2018)
Gulf of Mexico (GOM) deepwater	Workflow	Utilized visualizations to draw additional insights from drilling rigs	Bolen et al. (2018)
Bakken formations	Gradient boosting regressors (GBR) provided	Analyses showed that pad drilling and increased hydraulic fracture intensity are apparently associated with small increases in initial production rates but have led to larger terminal decline rates	Male et al. (2018)
Wolfcamp Basin (PB)	Machine learning (pattern recognition, multivariate analysis)	EUR determination through integration of geology, and machine learning (pattern recognition, multivariate analysis)	Gaurav (2017)
North Sea, UK	Iterative big data “discovery workflows”	Use of big data (multi-dimensional analysis) in optimizing drilling parameters	Johnston and Guichard (2015)
Barnett Shale, TX (gas)	Self-organized maps (SOM)	Highlighted completion and reservoir parameters (proppant, lb.) directly affecting production Investigated Barnett completion strategy from 1993 to 2002 The results show that SWF-type simulation of the Barnett outperformed to varying degrees XLG treatments for the five reservoir types used in this evaluation	Grieser et al. (2006)
N/A		A top-down modeling approach Employed data mining techniques to analyze shale production	Mohaghegh et al. (2017)
GOM and multiple basins	Real-time data	Explored potential applications of the technology with an eye toward speeding time to first oil, reducing costs and increasing the safety of drilling	Febowitz (2013)
Barnett Shale, TX (gas)	Self-organizing mapping (SOM)	SWF-type stimulation of the Barnett outperformed to varying degrees XLG treatments for the five reservoir types used in this evaluation	Shelley et al. (2011)
Wolfcamp shale formation (PB)	Regression modeling (random forest)	Identified factors separating good wells from poor performers	Schuetter et al. (2015)
Unconventional sandstone	Regression model	Used of nonlinear regression method and variable structure regressions (VSR) to predict cumulative well production (performance) Input of the model are Frac stages Average length TOC (total organic carbon) Proppant to fluid ratio Rock brittleness	Bakshi et al. (2017)
N/A	Time series	Introduced a workflow combining both ANN and SOM to predict production from unconventional resources	Gupta et al. (2014)

Table 6 (continued)

Shale Play	Data Analytic Tool	Findings	References
Permian Basin	Logistic nonlinear regression method	Use of proppant loading as a wide metric for well comparison should be in context of other completion parameters such as cluster spacing and number of stages Used public information while drilling down to specific completion parameters to assist in data-based decisions	Olaoye and Zakhour (2018)
Bone springs sands and Wolfcamp shale formation (PB)	Neural network models	Using plug-and-perf leads to a better perforation efficiency The solution involved using a technique to adjust perforation placement to achieve the lowest pressure required to open as many perforations as possible	Lehman et al. (2016)

Appendix 3: Regression tree-based measures for ranking predictors

1. %Inc Mean Squared Error (MSE)

It is calculated from the well data for each tree, the prediction error on the test (randomly rearranged values of the variable) is documented as MSE.

$$\text{MSE} = \text{mean} \left((y_{\text{measured}} - y_{\text{predicted}})^2 \right)$$

2. Inc Node Purity

It is considered from the test data as well, total reduction in node impurities from splitting on the variable, averaged over all trees. Then impurities are evaluated by residual sum of squares.

Appendix 4: Response surface model

$$y = \int_{\ell=1}^L \beta_{0,\ell} x_{0,\ell} + \int_{i=1}^k \beta_i x_i + \int_{i<j} \beta_{i,j} x_i x_j + \int_{i=1}^k \beta_{i,i} x_i^2 + \varepsilon, \quad (1)$$

where $x_{0,\ell}$ is a logic denoting the presence ($x_{0,\ell} = 1$) or absence ($x_{0,\ell} = 0$) of the factor at level ℓ in the model, $\beta_{0,0}, \dots, \beta_{0,L}, \beta_1, \dots, \beta_k, \beta_{1,1}, \beta_{1,2}, \dots, \beta_{k,k}$ are unknown coefficients to be estimated, and ε is a random error term.

$$\hat{y}(X) = \sum_{l=1}^L \hat{B}_{0,l} x_{0,l} + \sum_{l=1}^k \hat{B}_l x_l + \sum_{i<j} \hat{B}_{i,j} x_i x_j + \sum_{i=1}^k \hat{B}_{i,i} x_i^2 \quad (2)$$

References

- Bakshi A, Uniacke E, Korjani M (2017) A novel adaptive non-linear regression method to predict shale oil well performance based on well completions and fracturing data. In: SPE-185695-MS presented at the SPE western regional meeting, Bakersfield, California, USA, pp 23–27. <https://doi.org/10.2118/185695-MS>
- Bolen M, Crkvenjakov V, Converset J (2018) The role of big data in operational excellence and real time fleet performance management—the key to deepwater thriving in a low-cost oil environment. Society of Petroleum Engineers. <https://doi.org/10.2118/189603-MS>
- Cheng Y (2012) Impacts of the number of perforation clusters and cluster spacing on production performance of horizontal shale gas wells. SPE Reserv Eval Eng 15(01):31–40. <https://doi.org/10.2118/138843-PA>
- Febowitz J (2013) Analytics in oil and gas: the big deal about big data. Society of Petroleum Engineers. <https://doi.org/10.2118/163717-MS>
- Gaurav A (2017) Horizontal shale well EUR determination integrating geology, machine learning, pattern recognition and multivariate statistics focused on the Permian Basin. Society of Petroleum Engineers. <https://doi.org/10.2118/187494-MS>
- Grieser WV, Shelley RF, Johnson BJ, Fielder EO, Heinze JR, Werline JR (2006) Data analysis of Barnett shale completions. Society of Petroleum Engineers. <https://doi.org/10.2118/100674-PA>
- Griffin LG, Pearson CM, Strickland S, McChesney J, Wright CA, Mayer J, Weijers L (2013) The value proposition for applying advanced completion and stimulation designs to the Bakken Central Basin. Society of Petroleum Engineers. <https://doi.org/10.2118/166479-MS>
- Gupta S, Fuehrer F, Jeyachandra BC (2014) Production forecasting in unconventional resources using data mining and time series analysis. In: SPE-171588-MS presented at the SPE/CSUR unconventional resources conference, 30 September–2 October, Calgary. <https://doi.org/10.2118/171588-MS>
- <https://www.investopedia.com/terms/e/estimated-ultimate-recovery.asp>
- <https://www.verdazo.com/blog/what-production-performance-measure-should-i-use/>
- Izadi G, Settgaest R, Moos D, Baba C, Jo H (2015) Fully 3D hydraulic fracture growth within multi-stage horizontal wells. International Society for Rock Mechanics and Rock Engineering
- Jaripatke OA, Barman I, Ndungu JG, Schein GW, Flumerfelt RW, Burnett N, Bello HD, Barzola GJ (2018) Review of permian completion designs and results. Soc Pet Eng. <https://doi.org/10.2118/191560-MS>
- Johnston J, Guichard A (2015) New findings in drilling and wells using big data analytics. In: Offshore technology conference. <https://doi.org/10.4043/26021-MS>
- Khuri AI, Mukhopadhyay S (2010) Response surface methodology. Wiley Interdiscip Rev Comput Stat 2(2):128–149

- Lafollette R, Holcomb WD, Aragon J (2012a) Impact of completion system, staging, and hydraulic fracturing trends in the Bakken formation of the Eastern Williston Basin. Society of Petroleum Engineers. <https://doi.org/10.2118/152530-MS>
- Lafollette R, Holcomb WD, Aragon J (2012b) Practical data mining: analysis of Barnett shale production results with emphasis on well completion and fracture stimulation. Society of Petroleum Engineers. <https://doi.org/10.2118/152531-MS>
- Lehman LV, Jackson K, Noblett B (2016) Big data yields completion optimization: using drilling data to optimize completion efficiency in a low permeability formation. Soc Pet Eng. <https://doi.org/10.2118/181273-MS>
- Liaw A, Weiner M (2018) “Randomforest”: Breiman and Cutler’s random forests for classification and regression (R package version 4.6-14). R Foundation for Statistical Computing, Vienna
- Male F, Marder MP, Browning J et al (2016) Marcellus wells’ ultimate production accurately predicted from initial production. In: SPE-180234-MS paper presented at the SPE low perm symposium held in Denver, Colorado, USA
- Male F, Aiken C, Duncan IJ (2018) Using data analytics to assess the impact of technology change on production forecasting. Society of Petroleum Engineers. <https://doi.org/10.2118/191536-MS>
- McClure MW (2012) Modeling and characterization of hydraulic stimulation and induced seismicity in geothermal and shale gas reservoirs. Ph.D. thesis. Stanford University, Stanford, California
- Mohaghegh SD (2017) Data-driven reservoir modeling: top-down modeling (TDM): a paradigm shift in reservoir modeling, the art and science of building reservoir models based on field measurements. Society of Petroleum Engineers, Richardson
- Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (2004) Applied linear statistical models, 5th edn. McGraw-Hill/Irwin, New York
- Olaoye O, Zakhour N (2018) Utilizing a data analytics workflow to optimize completion practices while leveraging public data: a Permian Basin case study. Society of Petroleum Engineers. <https://doi.org/10.2118/191416-MS>
- Olufemi A, Ambastha A (2017) Initial production rate estimation: impact on recovery and interplay of recovery process. In: SPE-189078-MS paper presented at the Nigeria annual international conference and exhibition held in Lagos, Nigeria
- Pankaj P, Geetan S, MacDonald R, Shukla P, Sharma A, Menasria S et al (2018) Need for speed: data analytics coupled to reservoir characterization fast tracks well completion optimization. In: SPE Canada unconventional resources conference. Society of Petroleum Engineers, Alberta, Canada
- Passman A (2018) Modeling and data analytics workflow from delineation to development: a case study utilizing limited data in the extensional deep dry Utica shale play of Pennsylvania. Society of Petroleum Engineers. <https://doi.org/10.2118/191468-MS>
- Ran B, Kelkar M (2015) Fracture stages optimization in Bakken shale formation. In: Unconventional resources technology conference, no. July: 20–22. <https://doi.org/10.15530/URTEC-2015-2154796>
- Rashid K, Bailey W, Couet B, Hudgens P, Heim R (2014) Sector-wide multi-stage hydraulic fracture optimization. Society of Exploration Geophysicists
- Rodriguez A (2019) Inferences of two dynamic processes on recovery factor and well spacing for a shale oil reservoir. Society of Petroleum Engineers. <https://doi.org/10.2118/197089-MS>
- Roussel NP, Sharma MM (2011) Strategies to minimize frac spacing and stimulate natural fractures in horizontal completions. In: SPE annual technical conference and exhibition, 30 October–2 November, Denver, Colorado, USA. <https://doi.org/10.2118/146104-MS>
- Schuetter J, Mishra S, Zhong M, LaFollette R (2015) Data analytics for production optimization in unconventional reservoirs. In: Unconventional resources technology conference. <https://doi.org/10.15530/URTEC-2015-2167005>
- Shelley RF, Lehman LV, Shah K (2011) Survey of more than 1,000 fracture stage database with net pressure in the Barnett shale. Part 2. Society of Petroleum Engineers. <https://doi.org/10.2118/143330-MS>
- Stegent NA, Wagner A, Stringer C, Tompkins R, Smith N (2013) Engineering approach to optimize development strategy in the oil segment of the Eagle Ford shale: a case study. Society of Petroleum Engineers. <https://doi.org/10.2118/158846-PA>
- Taylor RS, Barree R, Aguilera R (2011) Why not to base economic evaluations on initial production alone. In: SPE 148680 paper presented at the Canadian unconventional resources conference held in Calgary, Alberta, Canada
- Veatch RW, King GE, Holditch SA (2017) Essentials of hydraulic fracturing: vertical and horizontal wellbores. PennWell Corporation, Tulsa
- Wu K, Olson JE (2014) Simultaneous multifracture treatments: fully coupled fluid flow and fracture mechanics for horizontal wells. SPE J 20(2):334–346. <https://doi.org/10.2118/167626-PA>
- Xia K, Fonseca E, Jones R, Zhan L (2016) A new perspective on multistage stimulation of multiple horizontal wells. In: The 50th US rock mechanics/geomechanics symposium, 26–29 June, Houston, Texas, USA
- Yousef AA, Gentil PH, Jensen JL, Lake LW (2006) A capacitance model to infer interwell connectivity from production and injection rate fluctuations. SPE Reserv Eval Eng J 9(5):630–646. <https://doi.org/10.2118/95322-pa>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.