Journal of Petroleum Exploration and Production Technology (2019) 9:1339–1354
https://doi.org/10.1007/s13202-018-0581-x

REVIEW PAPER – EXPLORATION ENGINEERING

# Mud loss estimation using machine learning approach

Abo Taleb T. Al-Hameedi[1] · Husam H. Alkinani[1] · Shari Dunn-Norman[1] · Ralph E. Flori[1] · Steven A. Hilgedick[1] · Ahmed S. Amer[2] · Mortadha Alsaba[3]

Received: 14 July 2018 / Accepted: 2 November 2018 / Published online: 28 November 2018

## Abstract

Lost circulation costs are a significant expense in drilling oil and gas wells. Drilling anywhere in the Rumaila field, one the world's largest oilfields, requires penetrating the Dammam formation, which is notorious for lost circulation issues and thus a great source of information on lost circulation events. This paper presents a new, more precise model to predict lost circulation volumes, equivalent circulation density (ECD), and rate of penetration (ROP) in the Dammam formation. A larger data set, more systematic statistical approach, and a machine-learning algorithm have produced statistical models that give a better prediction of the lost circulation volumes, ECD, and ROP than the previous models for events. This paper presents the new model, validates the key elements impacting lost circulation in the Dammam formation, and compares the predicted outcomes to those from the older model. The work previously presented by Al-Hameedi et al. (http://www.onepetro.org, 2017a; http://www.AADE.org, 2017b) provided a platform for predicting the severity of lost circulation incidents in the Dammam formation. Using the new models, the predictions closely track actual field incidents of lost circulation. When new lost circulation events were compared with predictions from the old and new models, the new model presented a much tighter prediction of events. Three equations for optimizing operations were developed from these models focusing on the elements that have the highest degree of impact. The total flow area of the nozzles was determined to be a significant factor in the ROP model indicating that nozzle size should be chosen carefully to achieve optimal ROP. Good modeling of projected lost circulation events can assist in evaluating the effectiveness of new treatments for lost circulation. The Dammam formation is a significant source of lost circulation in a major oilfield and warrants evaluation of the effectiveness of lost circulation treatments. These techniques can be applied to other fields and formations to better understand the economic impact of lost circulation and evaluate the effectiveness of various lost circulation mitigation efforts.

**Keywords** Lost circulation · Machine learning · Partial least squares · Iraq

### Abbreviations

| | |
|---|---|
| PV | Plastic viscosity |
| YP | Yield point |
| ECD | Equivalent circulation density |
| SPM | Strokes per minutes |
| MW | Mud weight |
| RPM | Revolutions per minute |
| TFA | Total flow area of the nozzles |
| WOB | Weight on bit |

## Background

Drilling fluid losses and problems associated with lost circulation while drilling represent a major expense in drilling oil and gas wells, by industry estimates, more than 2 billion USD is spent to combat and mitigate this problem each year (Arshad et al. 2015).

The materials of the drilling fluid are so expensive, companies spent $7.2 billion in 2011 and it is expected to reach $12.31 billion in 2018 as the global market for drilling fluid indicates, which shows a vigorous yearly maximize by 10.13% (Transparency Market and Research 2013). The cost of the drilling mud is equivalent to averages 10% of total well costs; however, drilling fluid can extremely impact the ultimate expenditure (Darley and Gray 1988). Lost circulation events, defined as the loss of drilling fluids into the formation, are known to be one of the most challenging

✉ Abo Taleb T. Al-Hameedi
ata2q3@mst.edu

[1] Missouri University of Science and Technology, Rolla, Missouri, USA

[2] Newpark Technology Center, Katy, Texas, USA

[3] Australian College of Kuwait, Kuwait City, Kuwait

problems to be prevented or mitigated during the drilling phase. The severity of the consequences varies depending on the loss severity; it could start as just losing the drilling fluid and it could end in a blowout (Messenger 1981). Among the top ten drilling challenges facing the oil and gas industry today is the problem of lost circulation. Major progress has been made to understand this problem and how to combat it. However, most of the products and guidelines available for combating lost circulation are often biased towards advertisement for a service company.

Lost circulation is a common drilling problem especially in highly permeable formations, depleted reservoirs, and fractured or cavernous formations (Nayberg and Petty 1986). The range of lost circulation problems begins in the shallow, unconsolidated formations and extends into the well-consolidated formations that are fractured by the hydrostatic head imposed by the drilling mud (Moore 1986). Two conditions are both necessary for lost circulation to occur downhole: (1) the pressure in the wellbore must exceed the pore pressure and (2) there must be a flow pathway for the losses to occur (Osisanya 2002). Subsurface pathways that cause, or lead to, lost circulation can be broadly classified as follows:

- Induced or created fractures (fast tripping or underground blow-outs).
- Cavernous formations (crevices and channels).
- Unconsolidated or highly permeable formations.
- Natural fractures present in the rock formations (including non-sealing faults).

The rate of losses is indicative of the lost pathways and can also give the treatment method to be used to combat the losses. The severity of lost circulation can be grouped into the following categories (Basra Oil Company 2012):

- Seepage losses: up to 1 $m^3$/h lost while circulating.
- Partial losses: 1–10 $m^3$/h lost while circulating.
- Severe losses: more than 15 $m^3$/h lost while circulating.
- Total losses: no fluid comes out of the annulus.

The Rumaila field in Iraq is one of the largest oilfields in the world. Wells drilled in this field are highly susceptible to lost circulation problems when drilling through the Dammam formation. Lost circulation events range from seepage losses to complete loss of the borehole and are a critical issue in field development. Figure 1 shows the Rumaila field location.

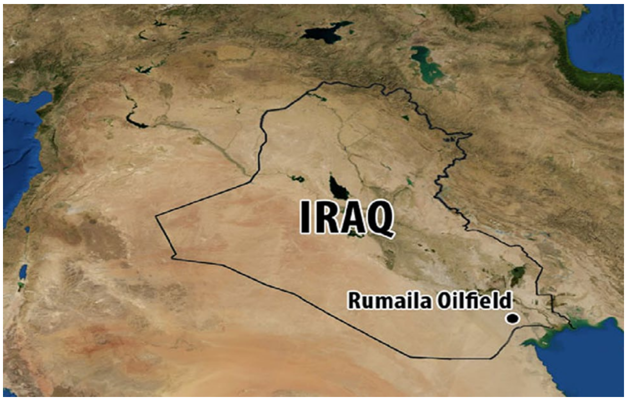


**Fig. 1** Rumaila field (Parks 2010)

The Dammam formation is a very shallow formation prone to mud losses and is continuous across the Rumaila field. The top of this zone is found between 435 and 490 m; thus, all of the wells in the field must be drilled through this zone. The interval is composed of interbedded limestone and dolomite, which is generally 200–260 m thick. The top of Dammam was eroded after burial and is karstified at depth. The karst features are believed to lead to the characteristic mud losses seen while drilling through this interval (Al-Hameedi et al. 2017a). Because of the persistent losses in the Dammam formation in such a valuable and large oilfield, it is worth studying the lost circulation issues of the Dammam formation to determine the effectiveness of treatments and mitigate efforts. Figure 2 provides a summary of the stratigraphic column and primary geological formations in Basra's oil fields. Formations, where loss circulation has occurred, include the Dammam, Hartha, and Shuaiba formations. The aim of this work is to provide estimation models for volume loss, equivalent circulation density (ECD), and rate of penetration (ROP) that can be used prior to drilling the Dammam formation using advanced machine learning approach. Also, the proposed models can be used in reverse to set up the key drilling parameters to avoid or at least mitigate mud losses.

## Modeling lost circulation in the Dammam formation

Data analysis has become a very popular topic nowadays. This is due to the large data sets that are recorded and available. Utilizing data analysis methods/techniques will help to evaluate and understand the performance of the particular process and will help to optimize the future outcomes. Data analysis has been used in most industries, for petroleum engineering particularly; data analysis has been utilized mostly in reservoir engineering to evaluate enhanced oil recovery methods (Al-Dhafeeri et al. 2005; Alvarado et al. 2008; Baker et al. 2012; Aldhaheri et al. 2016). However, in drilling, there is a gap in data analysis, since most drilling data are confidential and owned by companies.

Leite Cristofaro et al. (2017) utilized artificial intelligence strategies to minimize lost circulation NPT in deep water Brazilian wells. Several predictive data-mining techniques used such as Naive Bayes, Instance-Based and Neural Network to

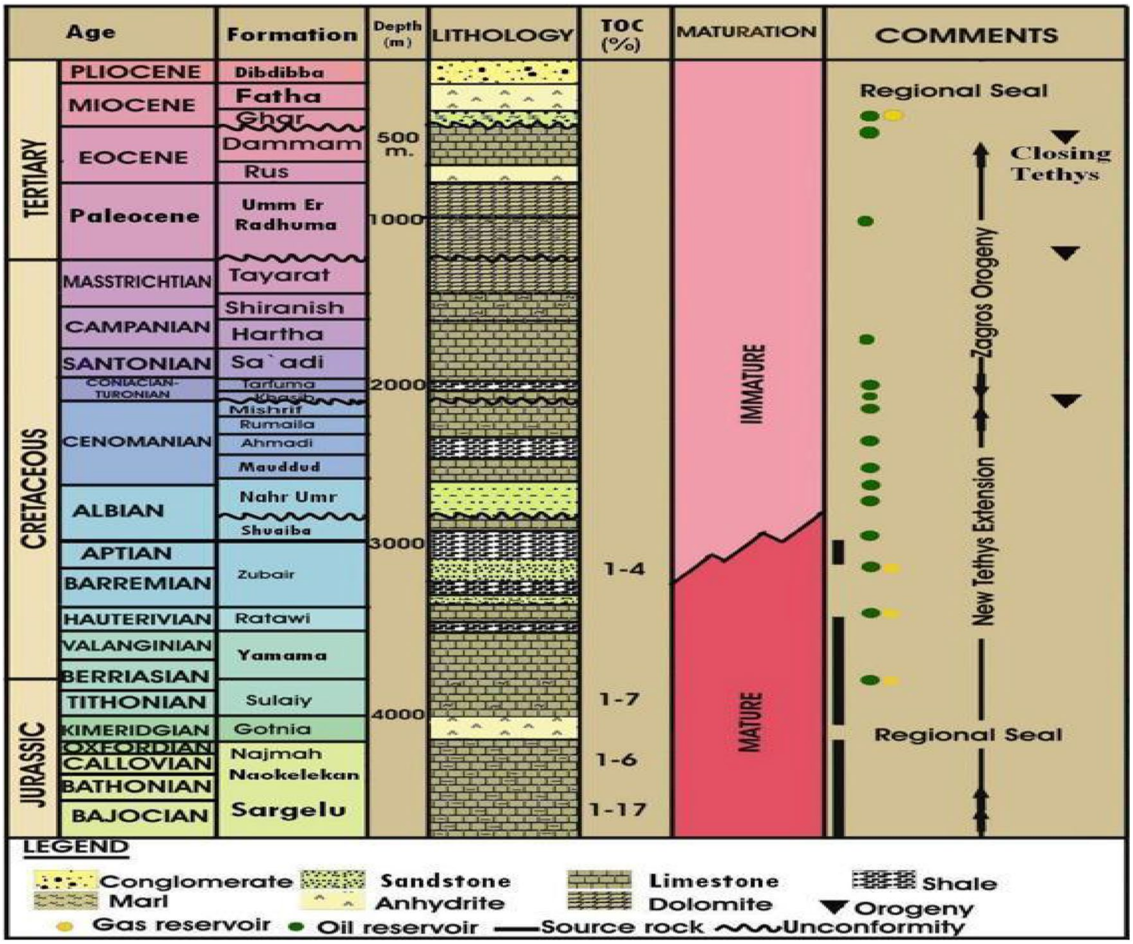


**Fig. 2** Stratigraphic column of the Rumaila field (Al-Ameri et al. 2011)

predict losses and choose the best treatment for losses prior to entering the losses zone. Hegde et al. (2015a, b) used principal component regression, least squares regression, and Ridge and Lasso regression as well as bootstrapping, trees, bagging, and the random forecast with data of to predict ROP. Wallace et al. (2015) developed a statistical learning model to predict and optimize the real-time drilling performance.

This paper shows the application of new models developed to study volume loss, equivalent circulating density (ECD), and rate of penetration (ROP) for lost circulation events within the Dammam formation in the Rumaila field in Iraq. The resulting models are compared to the previous models developed by Al-Hameedi et al. (2017a, b) for the Dammam formation in the same field. The old models presented by Al-Hameedi et al. (2017a, b) used 75 wells within the Dammam formation in the South Rumaila field in Iraq to develop three statistical models based on volume loss, ECD, and ROP. Least squared multi-linear regression was utilized to build the old models.

This study focuses on mud loss and lost circulation information extracted from drilling data of over 500 wells in the Rumaila field in Iraq and the lost circulation screening criteria developed for the Dammam formation, based on the historical mud loss and lost circulation problems. Data from over 500 wells were utilized to train the models, and another set of data of over 200 wells was used to test the models. Three mathematical models are created to evaluate the mitigation of lost circulation events—volume loss, ECD, and ROP. Lost circulation events are categorized according to the total volume of fluid lost during the event. The volume of mud losses depends on number of factors, including formation properties, drilling-fluid properties, operational drilling parameters, and formation breakdown pressure.

The aim of this new work was to develop a more systematic approach utilizing advanced machine learning algorithm to estimate mud losses prior to drilling to choose the best operational drilling parameters to limit volume loss. The new models use a significantly larger data set (over 500 wells compared to 75 wells on the old models) and utilize advanced machine learning algorithm (new models used partial least squares (PLS) regression and the old model used simple least square multi-linear regression). In addition, more operational drilling parameters were included in the new models such as plastic viscosity (PV) and the total flow area of the nozzles (TFA).

## PLS regression algorithm

Principal component analysis (PCA) is an unsupervised learning method, it is the very common approach to modeling to get low-dimensional features from a large set of variables. In other words, PCA is a mathematical tool that transforms a number of correlated variables to a smaller number of uncorrelated variables called principal components. PCA honors the variability of the data; thus, the first principal component will tend to represent most of the variability of the predictors (*X*-variables). This means that the response (*Y*-variable) is not utilized in identifying the principal components; this is why, it is called the unsupervised method. Unlike unsupervised learning techniques, supervised methods can be tested to see whether the model is estimating the response (*Y*-variable) with an acceptable range of error; such a test can be done with data that were not utilized on creating the model. Thus, partial least square (PLS), a supervised method alternative to PCA, used in this study. PLS will compute a new set of latent factors that are the linear combination of the original data. Then, using the new set of latent factors, a model will be fitted via least squares. Unlike, PCA, PLS will identify these new latent factors in a supervised technique—that is, it honors the response (*Y*-variable) as well as the predictors (*X*-variables). In other words, PLS will use both the predictors and the response to find the best direction that best explains the variability of the predictors as well as honoring the *Y*-variable (James et al. 2013). PLS was chosen from other regression techniques, because it is very efficient with large data sets and a large number of variables with collinearity. In addition, it is recommended for use in the petroleum industry (Tufféry 2011).

The first step of the algorithm of the PLS regression is centering and scaling the data. This means that the response and the predictor will be centered and scaled to have a mean and standard deviation of zero and one, respectively. Centering the data is important, since the variable mean and its variation around the mean will be both involved in constructing the latent factors. In other words, this will allow a change in one standard deviation of a predictor to be equivalent to the change of one standard deviation of another predictor. The data are scaled and centered before forming the interaction term. To illustrate how the interaction term is calculated, assume that there are two predictors ($X_1$ and $X_2$), the interaction term can be calculated from Eq. 1 (SAS 2008):

$$\text{Interaction term} = \left(\frac{X_1 - \text{mean}(X_1)}{\text{STD}(X_1)}\right) \times \left(\frac{X_2 - \text{mean}(X_2)}{\text{STD}(X_2)}\right). \tag{1}$$

The most common PLS algorithm is called non-linear iterative partial least square (NIPALS). The NIPALS works by extracting one factor at a time. Let $Y = Y_o$ be the scaled and centered matrix of the response value and $X = X_o$ be the centered and scaled matrix of the predictors. A linear combination $t = X_o w$ of the predictors will be created, where $t$ is the score vector and $w$ is its associated weight vector. The PLS algorithm predicts both $X_o$ and $Y_o$ by regression on t, as shown in Eqs. 2 and 3:

$$\hat{X}_o = tp, \quad \text{where } p' = (t't)^{-1}t'X_o \tag{2}$$

$$\hat{Y}_o = tc', \quad \text{where } c' = (t't)^{-1}t'Y_o. \tag{3}$$

The vectors $c$ and $p$ are called the *Y*- and *X*-loadings, respectively. The linear combination $t = X_o w$ will be chosen to maximize the covariance $t'u$ with some response (*Y*-variable) combination $u = Y_o q$. In addition, the *X*- and *Y*-weights, $w$ and $q$, are proportional to the first eigenvectors of $X_o'Y_oY_o'X_o$ and $Y_o' X_oX_o'Y_o$, respectively. This accounts of the extraction of the first latent factor of the PLS regression. The second latent factor can be extracted in the same way by replacing $X_o$ and $Y_o$ with the *X*- and *Y*-residuals from the first latent factor, as shown in Eqs. 4 and 5 (SAS 2008):

$$X_1 = X_o - \hat{X}_o \tag{4}$$

$$Y_1 = Y_o - \hat{Y}_o. \tag{5}$$

The same extraction process of the score vectors is repeated for as many latent factors as are desired (SAS 2008).

### Cross validation

After computing all latent factors, a cross validation has to be performed to decide how many latent factors should be included in the model. The number of latent factors has to be chosen to meet two goals; the first one is to capture the variation in *X*-variables and to honor the predictive (*Y*-variable); the second one, however, is to avoid overfitting (avoid utilizing large number of latent factors that may result in overfitting). The root mean of the predicted residual sum of squares (PRESS) and scores plots are typically used to perform cross validation, the process of finding the root mean PRESS for a specific number of factors *A*, is summarized in Fig. 3 (SAS 2008).

### Variable importance in projection

One of the most important steps in the PLS algorithm is to choose the *X*-variables that are important to the model and eliminate the *X*-variables that are not important. This can be done using variable importance in projection (VIP). The VIP is computed for each *X*-variable, and then, a threshold

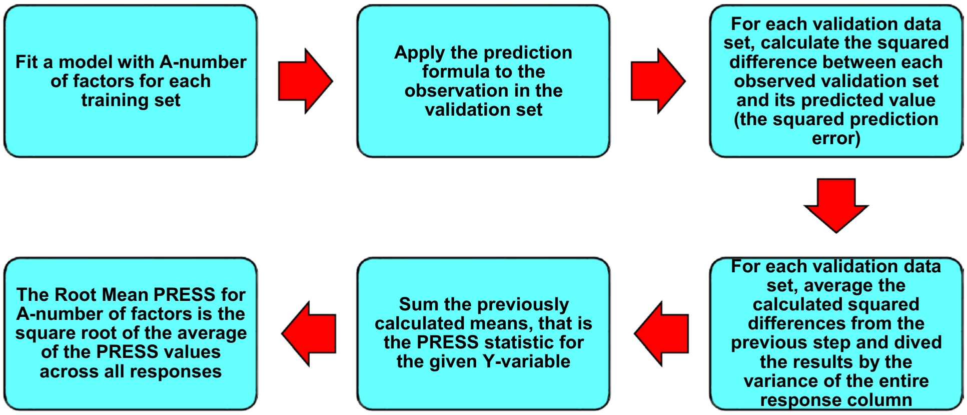

Fit a model with A-number of factors for each training set
Apply the prediction formula to the observation in the validation set
For each validation data set, calculate the squared difference between each observed validation set and its predicted value (the squared prediction error)
For each validation data set, average the calculated squared differences from the previous step and dived the results by the variance of the entire response column
Sum the previously calculated means, that is the PRESS statistic for the given Y-variable
The Root Mean PRESS for A-number of factors is the square root of the average of the PRESS values across all responses

**Fig. 3** Process of finding root mean of PRESS

has to be chosen to eliminate the variables that are below the chosen threshold. In general, a threshold of 0.8 is considered to be low; thus, any variable has a VIP less than 0.8 will be eliminated (Eriksson et al. 2006).

For each $X$-variable ($j_1, j_2 \ldots j_n$), the VIP for each $X$-variable can be calculated as follows (Wold et al. 1993; Tran et al. 2014):

$$\mathrm{VIP}_j = \sqrt{d \sum_{k=1}^{A} v_k (w_{kj})^2 / \sum_{k=1}^{A} v_k}, \tag{6}$$

where $d$ is the number of variables, $A$ is the number of latent factors, and $v_k$ is the variance of $X$ which can be expressed as follows:

$$v_k = c_k{}^2 t'_k t_k, \tag{7}$$

where $C_k$ is calculated for each column of the $t$ score vector and for the predicted response $y$ as follows:

$$c_k = \frac{t'_k y(k)}{t'_k t_k}. \tag{8}$$

Figure 4 shows a summary of the PLS algorithm.

## Approach

Given the number of drilling parameters that affect mud loss and the complex interrelationship between some of the drilling parameters, a drilling engineer is challenged to select the optimum value for each parameter that will optimize the entire situation. The purpose of this work is to develop advanced regression models to estimate mud loss, ECD, and ROP using advanced statistical techniques. These models are then tested with new data and compared with previous regression models developed for the Dammam formation (Al-Hameedi et al. 2017a).

Data of key drilling parameters [e.g., ROP, ECD, mud weight (MW), yield point (Yp), plastic viscosity (PV), flow rate (Q), strokes per minutes (SPM), revolutions per minutes (RPM), weight on bit (WOB), pressure losses ($\Delta P_{\mathrm{losses}}$), and total flow area (TFA) of the nozzles] for more than 500 wells were gathered from daily drilling reports, technical reports, final wells reports, and drilling programs. Partial least squares (PLS) regression was used to develop these models. All key drilling parameters were tested to find which parameters were significant and should be included in the models. The variable importance in projection (VIP) was used to test the key drilling parameters. The VIP threshold is assumed to be 0.8, and any key drilling parameter has a VIP greater than 0.8 will be included in the model. Finally, a sensitivity analysis is conducted for the parameters influencing mud loss, ECD, and ROP using Visual Basic® for Application (VBA) in Excel®.[1]

The purpose of the sensitivity analysis is to examine which parameter has the highest influence in each model and to test the effect of each parameter in each model. Figure 5 summarizes the methodology of this paper.

## Volume loss model

The process of creating the model involves the section of the number of latent factors. Score plots help to select the optimum number of latent factors that will be used in the model. Unlike the principal component analysis (PCA), the PLS scores plots are calculated to explain the variation in $x$ and $y$ and to maximize the relationship between $x$- and $y$-variables. Choosing the optimum number of latent factors is a complicated process and requires trial and error until reaching the optimum number of latent factors. Using too many latent factors will lead to overfitting the model which in return will flip the sign of some variables and make the model unrealistic. On the other hand, using a very low number of latent factors will not explain

[1] Visual Basic and Excel are registered trademarks of Microsoft.

**Fig. 4** PLS algorithm summary

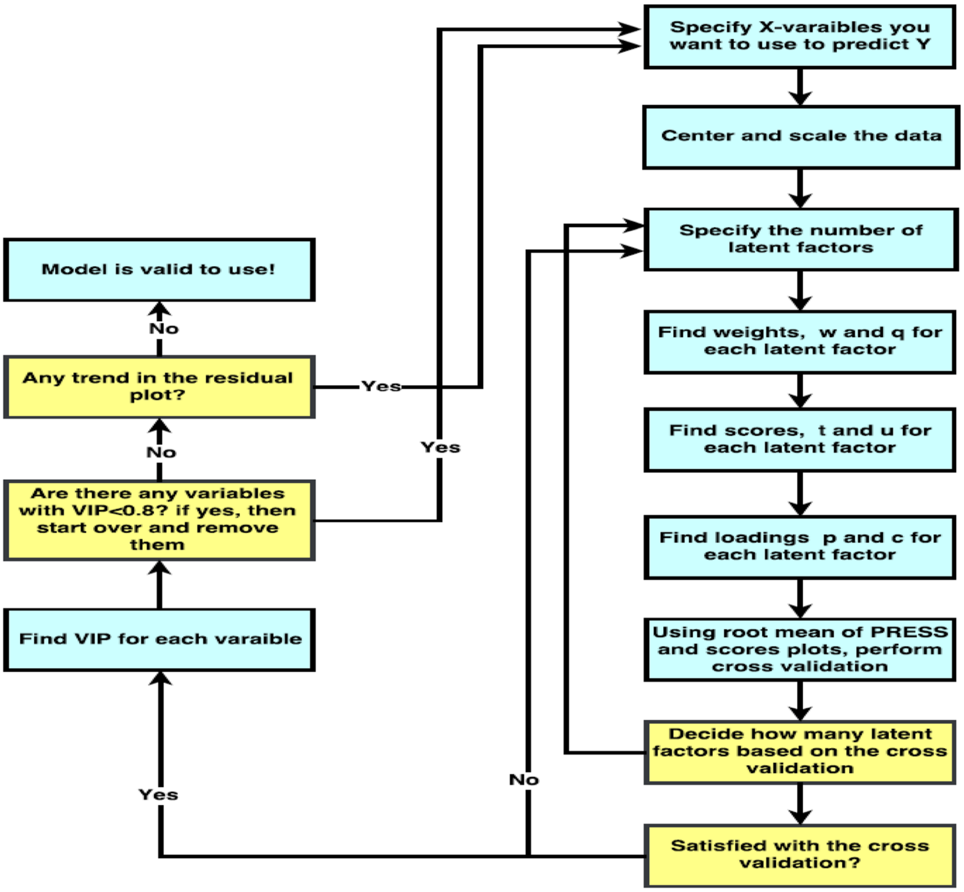
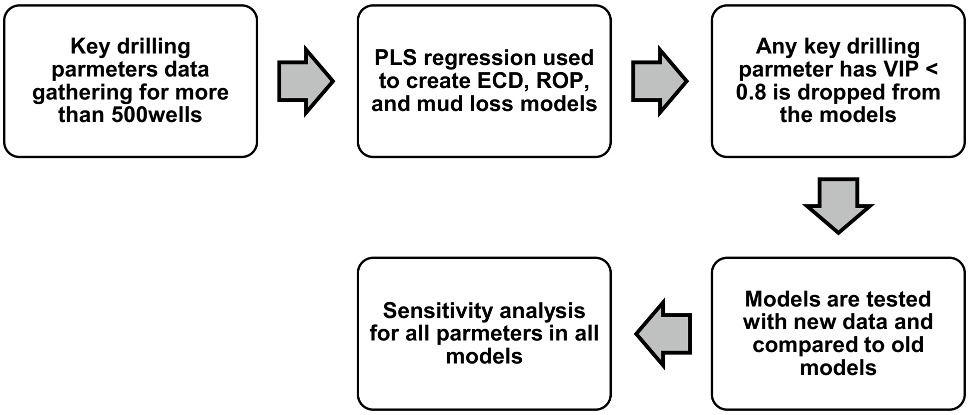

Specify X-varaibles you want to use to predict Y
Center and scale the data
Specify the number of latent factors
Find weights, w and q for each latent factor
Find scores, t and u for each latent factor
Find loadings p and c for each latent factor
Using root mean of PRESS and scores plots, perform cross validation
Decide how many latent factors based on the cross validation
Satisfied with the cross validation?
No
Yes
Find VIP for each varaible
Are there any variables with VIP<0.8? if yes, then start over and remove them
Yes
No
Any trend in the residual plot?
Yes
No
Model is valid to use!

**Fig. 5** Approach


Key drilling parmeters data gathering for more than 500wells
PLS regression used to create ECD, ROP, and mud loss models
Any key drilling parmeter has VIP < 0.8 is dropped from the model
Models are tested with new data and compared to old models
Sensitivity analysis for all parmeters in all models

the variability of $x$ and $y$. Two criteria are used to select the optimum number of latent factors. The first one is by minimizing at the root mean of the predicted residual sum of squares (PRESS). The second one is by inspecting the score plots of $x$- and $y$-variables, and each latent factor will have one score plot of $x$ versus $y$. If there is a trend in the score plot, it means that the latent factor should be included in the model. However, if no trend is presented in the score plot, the latent factor should be ignored (Tufféry 2011).

Figure 6 shows the root mean of PRESS versus number of latent factors. The figure is plotted for ten latent factor to inspect the optimum number of latent factors. By applying the first criteria, it is easy to see that having two or more latent factors will minimize the root mean of PRESS. However, it is not clear how many latent factors should be chosen. This is where the scores plots come to play, Fig. 7 shows the score plot for six latent factors. By applying the second criteria, having more than two latent factors will not add any valuable information to the model, since there is almost no

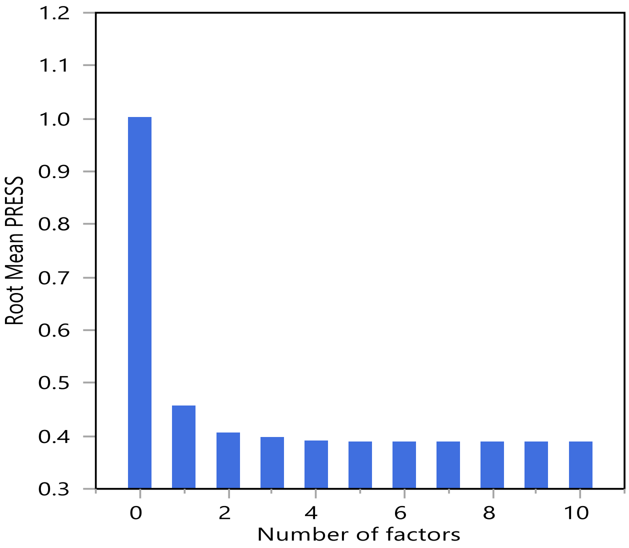


**Fig. 6** Number of latent factors versus root mean PRESS for volume loss model

relationship between $x$ scores and $y$ scores after two factors. Thus, two factors are chosen for this model.

Figure 8 shows the VIP versus the coefficients plot for each drilling parameter. Any drilling parameter below the 0.8 VIP line will be dropped, since it is considered not significant to the model. After dropping those parameters, the coefficients on the model will be changed, and Fig. 9 shows the new VIP versus coefficient plot after removing the key drilling parameters that have less than 0.8 VIP.

Figure 10 shows the correlation loading plot. This plot shows the relationship between the variables, the strong relationship between the variables can be indicated from their distance from each other, the closer the variable to another variable indicates a strong relationship and vice versa. Another thing can be observed from this plot is the percentage circles (25%, 50%, 75%, and 100%); the significant variables should be between the 50 and 100%. This is a useful check for the significant variables that might be missed by the VIP test. Moreover, the $R$-squared of each latent factor is shown in the plot, and the cumulative $R$-squared for this model is the sum of the $y$ R-squared of each latent factor, which is 0.83. Volume loss can be estimated using Eq. 9 prior to drilling the Dammam formation:

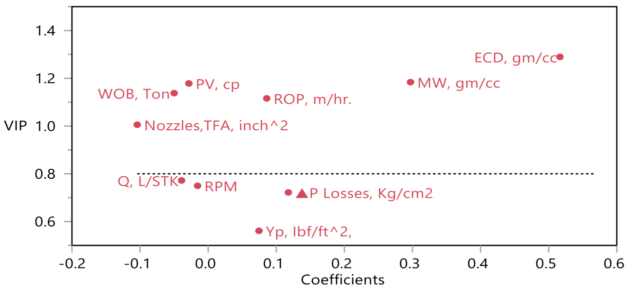


**Fig. 8** VIP versus coefficient for volume loss model (before refining)

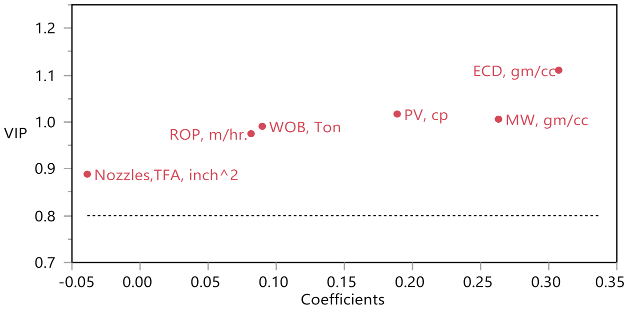


**Fig. 9** VIP versus coefficient for volume loss model (after refining)

$$\begin{aligned}\text{Volume loss} = & -1088.52 + 509.76 \times \text{ECD}\left(\frac{\text{g}}{\text{cc}}\right) \\ & + 504.35 \times \text{MW}\left(\frac{\text{g}}{\text{cc}}\right) \\ & - 0.492 \times \text{Nozzels, TFA (in.}^2) \\ & + 0.93 \times \text{PV(cp)} + 0.86 \times \text{ROP}\left(\frac{\text{m}}{\text{h}}\right) \\ & + 0.6 \times \text{WOB (ton).} \end{aligned} \tag{9}$$

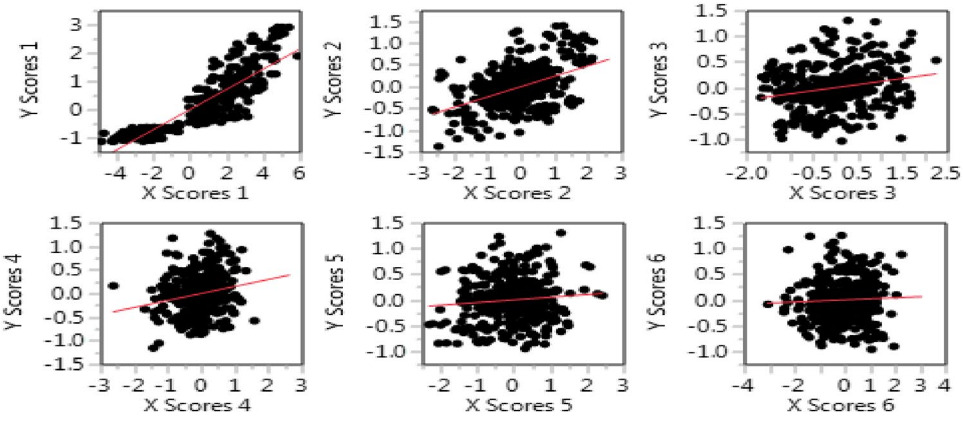


**Fig. 7** Scores plots for volume loss model

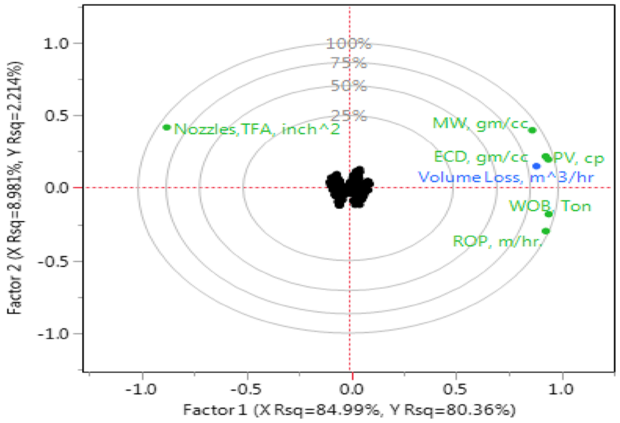


**Fig. 10** Correlation loading plot of volume loss model

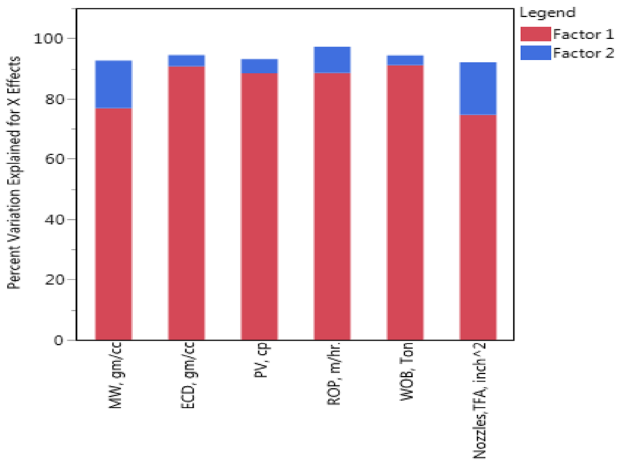


**Fig. 11** Percent of variation explained by each factor for *X*-variables (volume loss model)

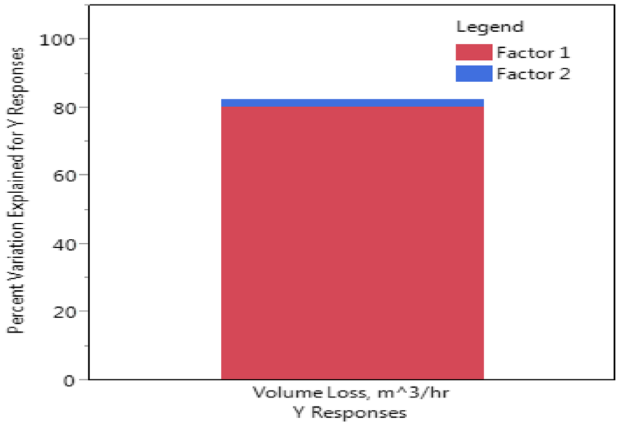


**Fig. 12** Percent of variation explained by each factor for *Y*-variable (volume loss model)

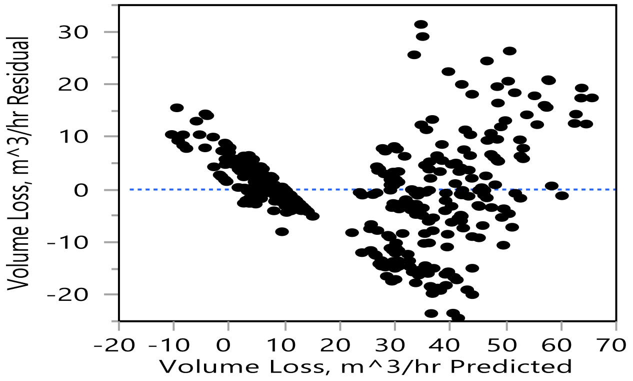


**Fig. 13** Residual plot of volume loss model

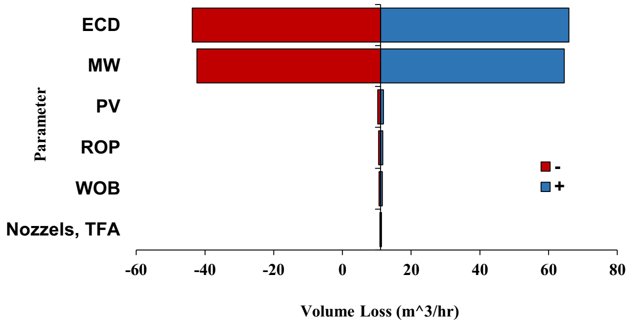


**Fig. 14** Tornado chart for volume loss model

Figures 11 and 12 show the percent variation explained by each factor for *x*-variables and *y*, respectively. It is easy to see that the second factor is not contributing to the model as much as the first factor. Choosing more than two factors will not add anything to the model. Thus, two factors are the optimum number of factors for this model. Figure 13 shows the residual plot of the volume loss model. A residual plot is a plot of residual (actual minus predicted) versus the predicted. No trend is observed in the residual plot. Thus, the model is valid.

## Tornado chart of volume loss model

Figure 14 presents a tornado chart for the significant parameters of volume loss model. A 10% sensitivity is used in this analysis. The base parameters are as follows: ECD = 1.075 g/cc, MW = 1.07 g/cc, TFA = 4.42 in.$^2$, PV = 9 cp, ROP = 7 m/h, and WOB = 8 ton. Figure 14 shows that volume loss is highly influenced by ECD, and least influenced by TFA.

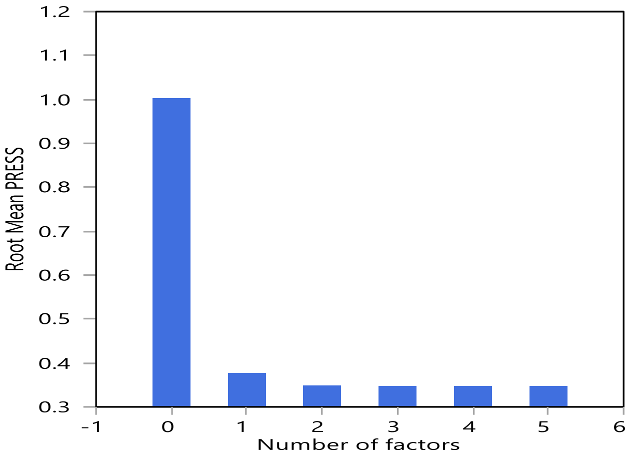


**Fig. 15** Number of latent factors versus root mean PRESS for ECD model

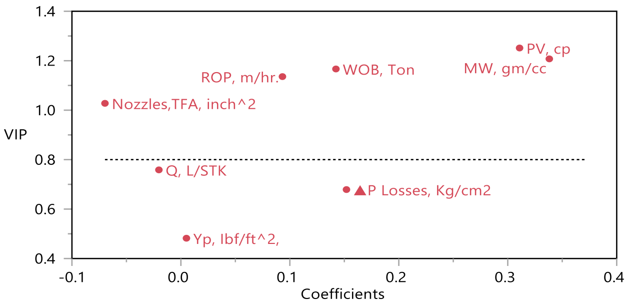


**Fig. 17** VIP versus coefficient for ECD model (prior to refining)

## ECD model

The same procedure is applied to develop a model to estimate ECD. Figure 15 shows the root mean of PRESS versus number of latent factors for the ECD model. Using Fig. 15, the number of factors that will minimize the root mean PRESS is 4 factors. Figure 16 shows the scores plots for the ECD for 4 factors. Looking closely at Fig. 16, an argument can be made about factor 4 which is not contributing much to the model. If only the scores plots are used, then the 4th factor can be eliminated. However, deleting this factor will flip and signs of the model and the minimization of the root mean PRESS will not be obtained. Thus, the 4th factor should be kept in the model.

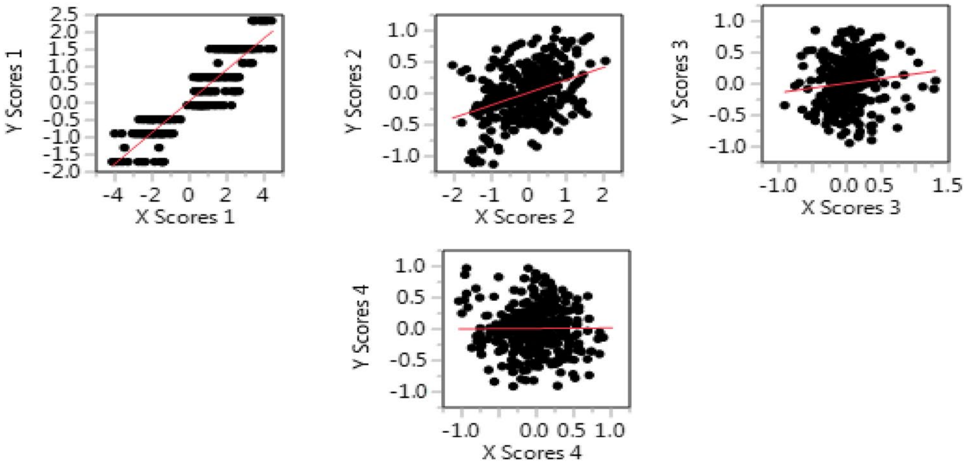
Figure 17 shows the VIP versus the coefficients of the model for each key drilling parameter. A threshold of 0.8 is utilized to refine the model. Any key drilling parameter that has a VIP less than 0.8 was ignored. Figure 18 shows the VIP versus the coefficients of the key drilling parameter after applying the 0.8 VIP threshold.

Figure 19 shows the loading plot of the ECD model. Again, the closer the variables to each other indicate a strong relationship and vice versa. The cumulative R-squared for this model is 0.88. Equation 10 can be used to estimate ECD prior to drilling the Dammam formation:

$$\begin{aligned}\mathrm{ECD} = {} & 0.76 + 0.28 \times \mathrm{MW}\left(\frac{\mathrm{g}}{\mathrm{cc}}\right) - 0.00084 \\ & \times \text{Nozzels, TFA } (\mathrm{in.}^2) + 0.0013 \times \mathrm{PV}\ (\mathrm{cp}) \\ & + 0.00053\ \mathrm{ROP}\left(\frac{\mathrm{m}}{\mathrm{h}}\right) + 0.00057 \times \mathrm{WOB}\ (\mathrm{ton}).\end{aligned} \tag{10}$$

Figures 20 and 21 show the percent variation explained by *x* and *y*, respectively. Going back to the argument of adding factor 4, using Fig. 20, it is easy to see that factor 4 is contributing to the *x* variations especially for the variable nozzles TFA and WOB. Thus, it is necessary to add factor 4 to the model. Figure 21 shows that only factors 1 and 2 are contributing to the variation of *y*. Figure 22 shows the residual plot of the ECD model, and no trend is observed on the residual plot. Thus, the model is valid.

**Fig. 16** Scores plots for ECD model



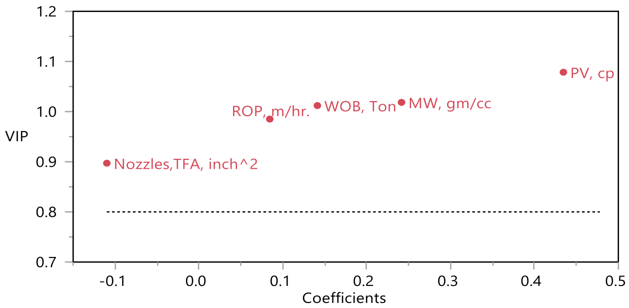


**Fig. 18** VIP versus coefficient for ECD model (after refining)

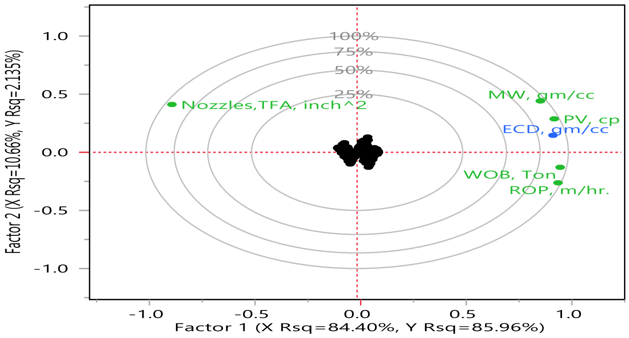


**Fig. 19** Correlation loading plot of ECD model

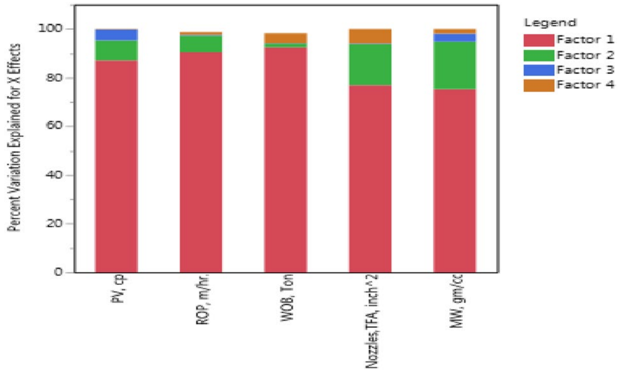


**Fig. 20** Percent of variation explained by each factor for *X*-variables (ECD model)

## Tornado chart of ECD model

Figure 23 shows a tornado chart for the significant parameters of ECD. A 10% sensitivity is used in this analysis. The base parameters are as follows: MW = 1.06 g/cc, PV = 9 cp, TFA = 4.42 in.$^2$, ROP = 7 m/h, and WOB = 8 ton. Figure 23 shows that ECD is highly influenced by MW, and least influenced by TFA.

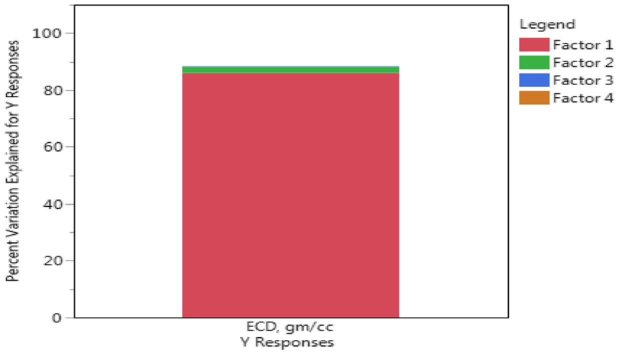


**Fig. 21** Percent of variation explained by each factor for *Y*-variable (ECD model)

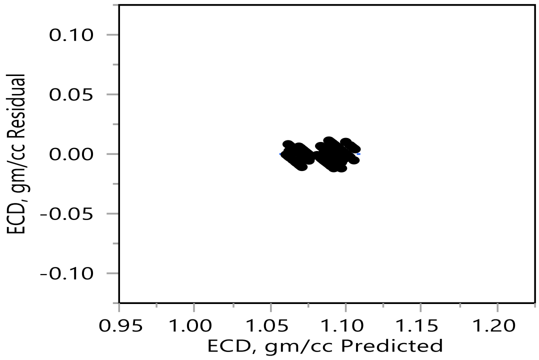


**Fig. 22** Residual plot of ECD model

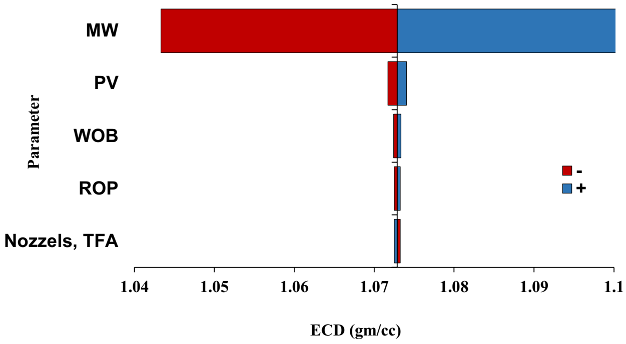


**Fig. 23** Tornado chart for ECD model

## ROP model

Once again, the same analysis and procedure explained previously are utilized to create a model to estimate ROP. Figure 24 shows the root mean of PRESS versus number of latent factors for the ROP model, 3 or 4 factors will minimize the root mean of PRESS, but it is still not clear how many latent factors should be used. Figure 25 shows the

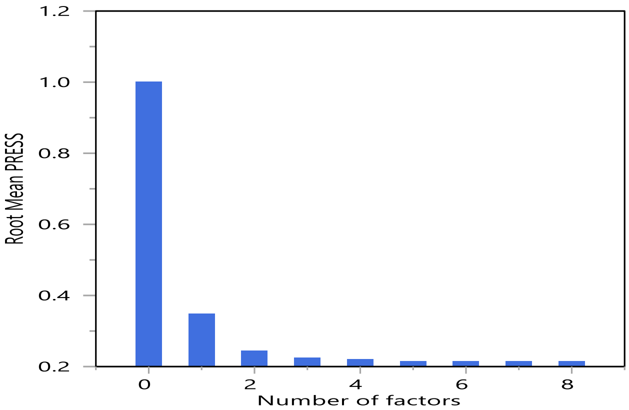


**Fig. 24** Number of latent factors versus root mean PRESS for ROP model

scores plots for 3 factors of the ROP model. From Fig. 25, it is easy to see that the three latent factors are doing a good job explaining the variation of the data. Thus, three factors are chosen to be the optimum number of factors.

Figure 26 shows the VIP versus the coefficients of the model for each key drilling parameter. A threshold of 0.8 is utilized to refine the model. Any key drilling parameter that has a VIP less than 0.8 was ignored. Figure 27 shows the loading plot of the ROP model. The cumulative R-squared for this model is 0.94. Equation 11 can be used to estimate ROP prior to drilling the Dammam formation:

$$\begin{aligned}\text{ROP} = {} & 6.94 - 1.00338 \times \text{MW}\left(\frac{\text{g}}{\text{cc}}\right) - 0.55 \\ & \times \text{Nozzles, TFA (in.}^2) + 0.027 \times \text{PV (cp)} \\ & + 0.012 \times \text{RPM} + 0.0021 \times \text{SPM} + 0.295 \\ & \times \text{WOB (ton).}\end{aligned} \tag{11}$$

Figure 28 shows the residual plot of the ROP model, and no trend is observed on the residual plot. Hence, the model is valid. Figures 29 and 30 show the variation explained by each latent factor for *x*- and *y*-variables, respectively. Looking at Fig. 29, it is easy to see that factor 3 is contributing to the variability of SPM and RPM. Thus, it was necessary to include the factor 3 on the model. Figure 30 shows that only factors 1 and 2 are contributing to the variability of *y*.

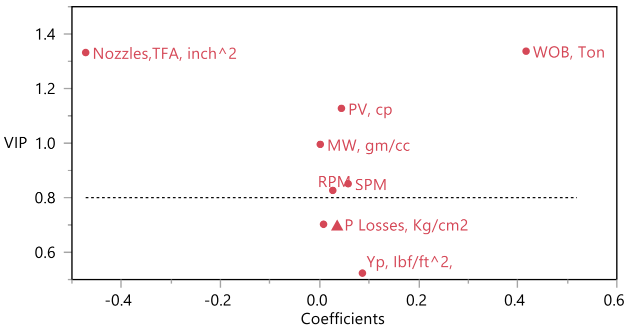


**Fig. 26** VIP versus coefficient for ROP model (prior to refining)

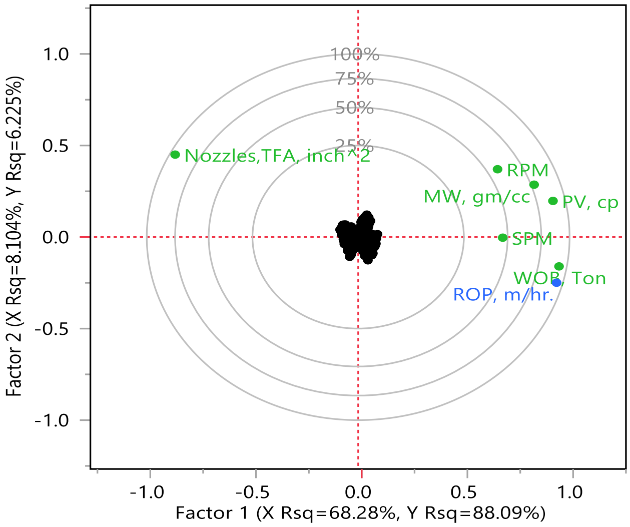
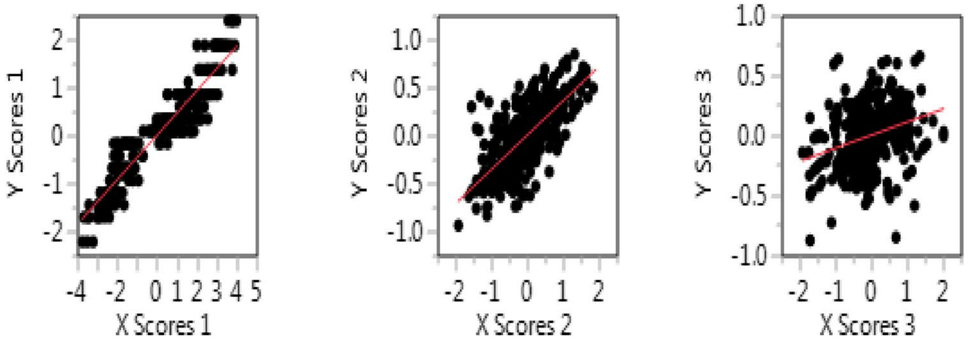


**Fig. 27** Correlation loading plot of ROP model

**Fig. 25** Scores plots for ROP model



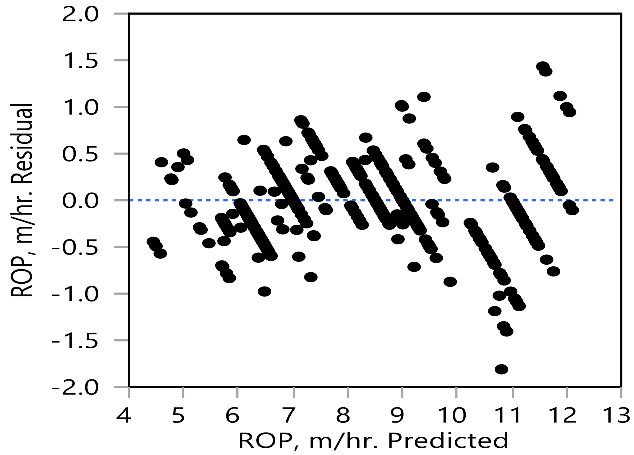


**Fig. 28** Residual plot of ROP model

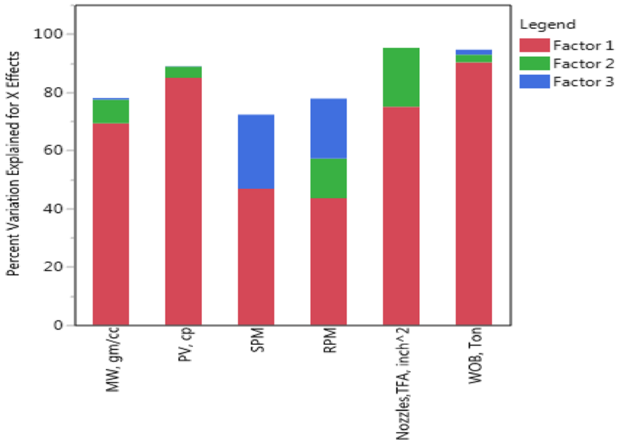


**Fig. 29** Percent of variation explained by each factor for *X*-variables (ROP model)

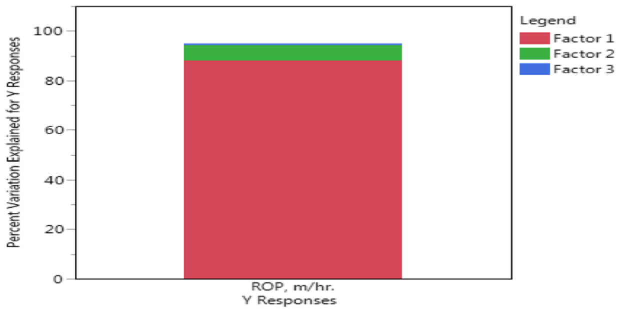


**Fig. 30** Percent of variation explained by each factor for *Y*-variable (ROP model)

## Tornado chart of ROP model

Figure 31 shows a tornado chart for the significant parameters of ROP. A 10% sensitivity is used in this analysis.

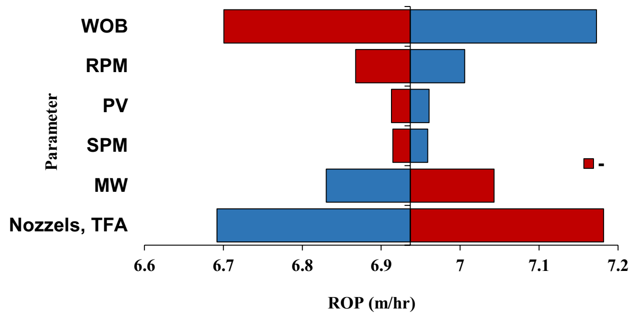


**Fig. 31** Tornado chart for ROP model

The base parameters are as follows: MW = 1.06 g/cc, PV = 9 cp, TFA = 4.42 in.$^2$, SPM = 105, RPM = 60, and WOB = 8 ton. Figure 31 shows that ROP is highly influenced by WOB and TFA, and least influenced by SPM.

## Models verifications and comparisons

An essential step that should be done is testing the models on a new data and see if they work or not. The new models are tested and compared to the old models presented by Al-Hameedi et al. (2017a). The new models are tested using new data sets of over 200 wells for the Dammam formation. The new data that were used to test the models were not used to create the models. Figures 32, 33, 34, 35, 36, 37 show a comparison between the old and the new models for partial, severe, and complete losses. Looking at these figures, it is easy to conclude that the new models are doing much better of estimating the actual mud loss, ECD, and ROP. Figures 32, 33, 34, 35, 36, 37 show the predicted versus the actual of the new and the old models for partial, severe, and complete losses, and it is easy to see the black line (45° line) overlaps with the data which indicates that there is a very strong correlation between the actual and the predicted data.

## Conclusions

This paper presents a deep statistical analysis of more than 500 wells in the Rumaila field. This work includes the application of advanced techniques to develop mathematical models to estimate volume losses in the Dammam formation, as well as the ECD and ROP associated with the losses model.

The three models developed in this study can be used to estimate mud losses prior to drilling the Dammam formation. Alternatively, given a target loss volume, the models can be used in reverse, to set key drilling parameters to limit losses while drilling. The volume loss models provide

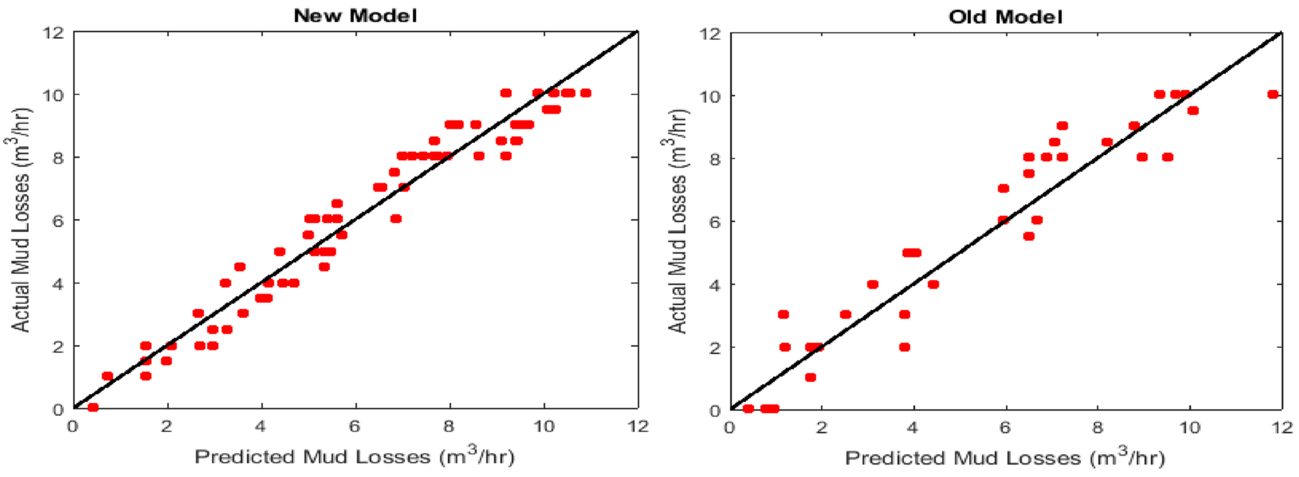


Fig. 32 Predicted versus actual mud loss for partial losses

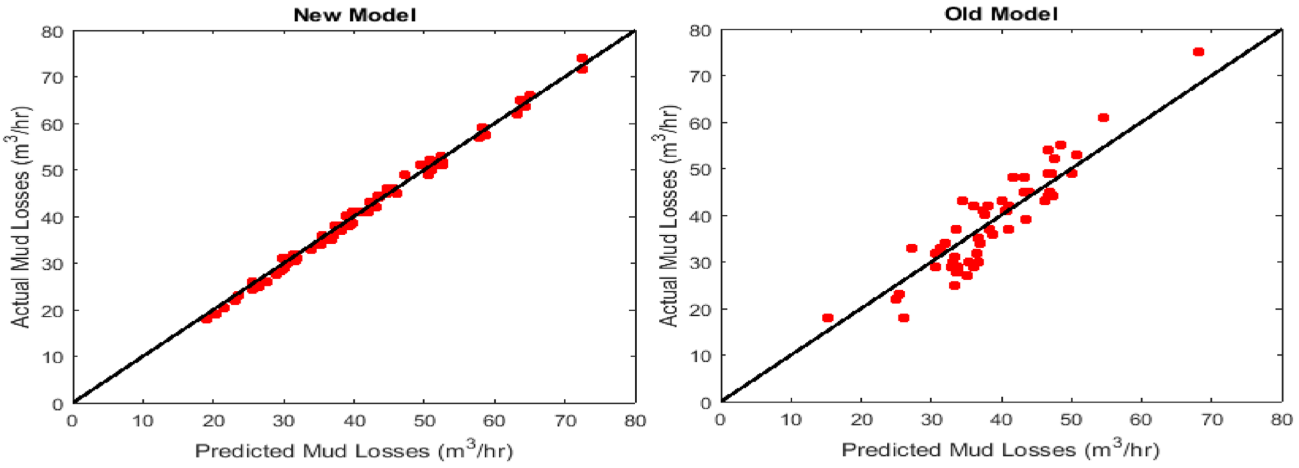


Fig. 33 Predicted versus actual mud loss for severe and complete losses

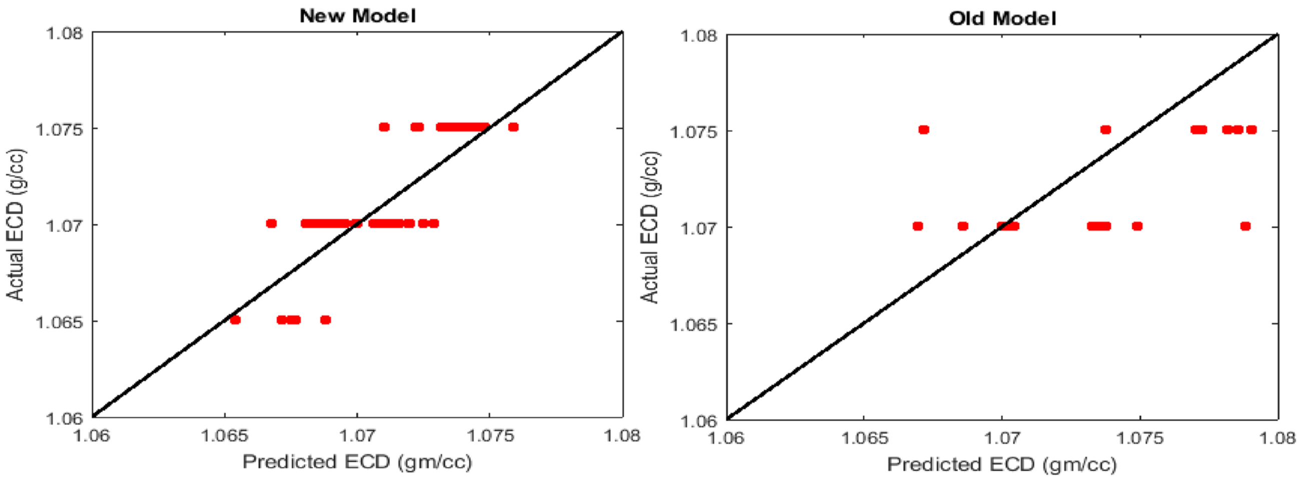


Fig. 34 Predicted versus actual ECD for partial losses

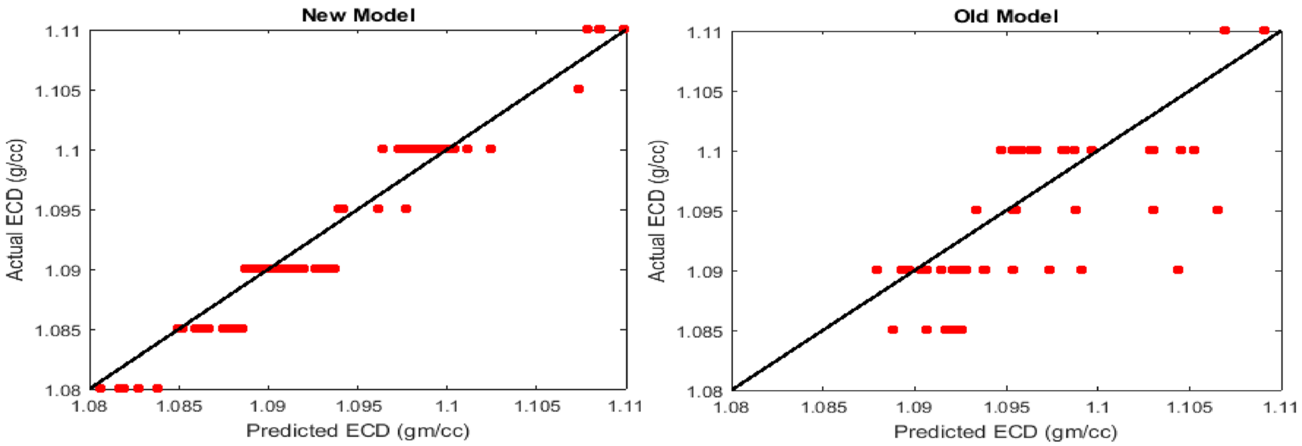


**Fig. 35** Predicted versus actual ECD for severe and complete losses

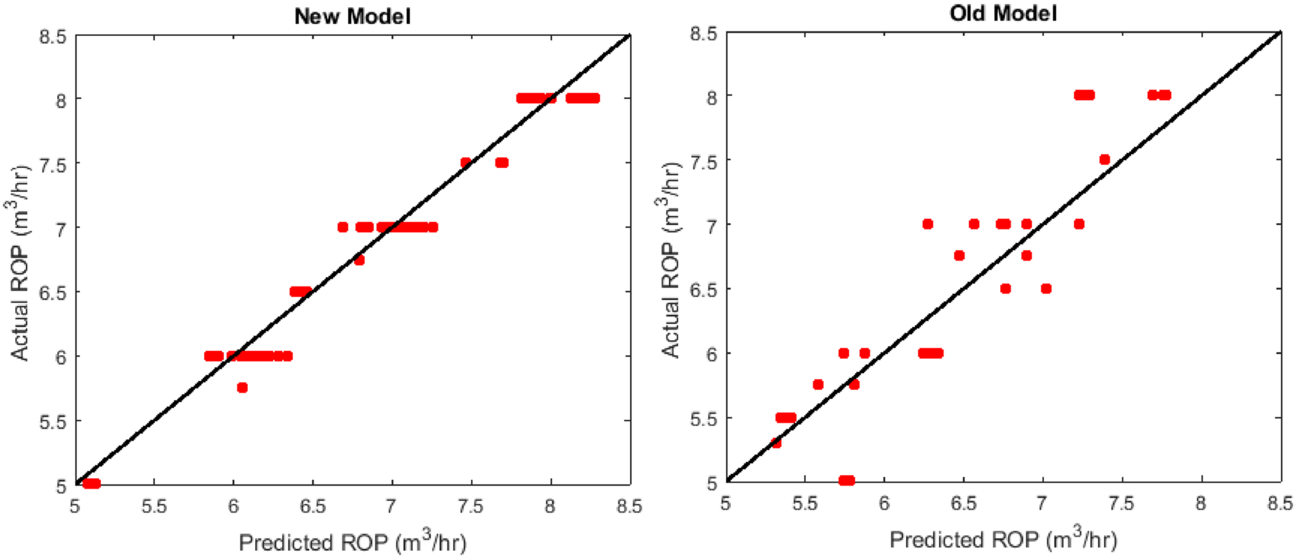


**Fig. 36** Predicted versus actual ROP for partial losses

greater consistency in the approach to handling mud losses for wells drilled in the Rumaila field. The models provide a formalized methodology for responding to losses and provide a means of assisting drilling personnel to work through the mud loss problems in a more systematic way. Based on this study, the following conclusions are made:

- Three advanced statistical models are developed that will help to estimate volume loss, ECD, and ROP prior to drilling the Dammam formation.
- The new models are doing a much better job than the old models in estimating mud loss, ECD, and ROP.
- TFA is a very important parameter for the ROP model. It has a negative influence on the ROP model. Thus, the selection of the nozzle size should be done very carefully.
- Using advanced statistical methods—such as PLS—will enhance the prediction of volume loss, ECD, and ROP. One challenge of using the PLS method is the selection of the optimum number of latent factor. Thus, carefully inspecting the root mean of PRESS plot and the scores plots as well as the percent variations of *y* and *x* plots will help to select the optimum number of latent factors.
- The three equations that were developed in this study can be used globally if the characteristics of the formation are the same as the Dammam formation.

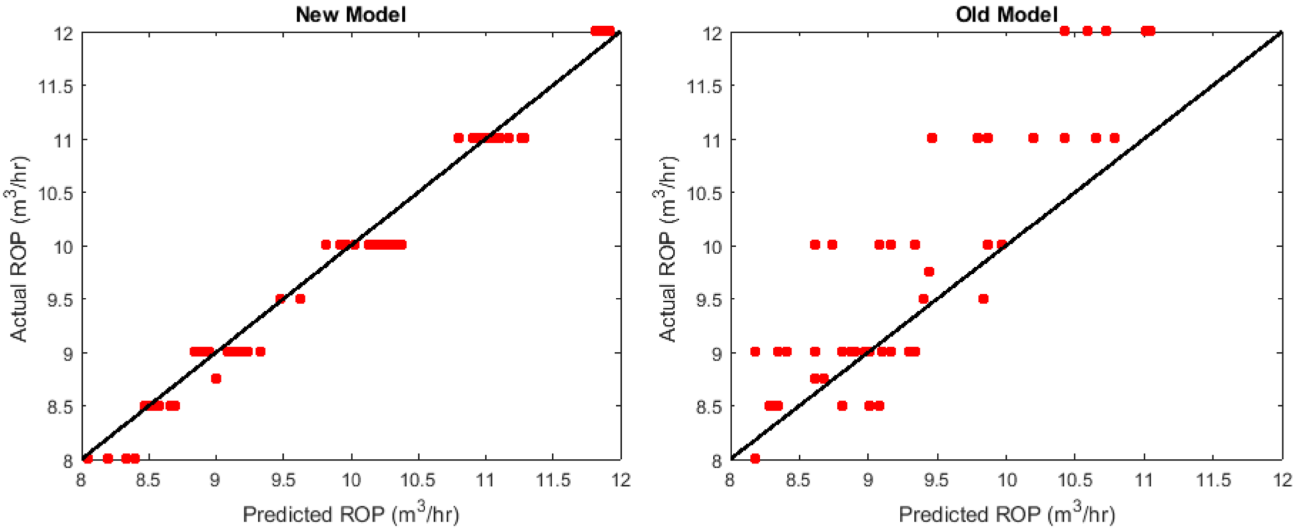


**Fig. 37** Predicted versus actual ROP for severe and complete losses

**Acknowledgements** The authors would like to thank Basra Oil Company from Iraq and British Petroleum Company for providing us various real field data.

## References

Al-Ameri KT, Jafar MSA, Pitman J (2011) Modeling hydrocarbon generations of the Basrah Oil Fields, Southern Iraq, based on Petromod with palynofacies evidences. AAPG search and discovery article #90124 © 2011 AAPG Annual Convention and Exhibition, April 10–13, 2011, Houston, Texas

Al-Dhafeeri AM, Nasr-El-Din HA, Seright RS, Sydansk RD (2005) High-permeability carbonate zones (Super-K) in Ghawar Field (Saudi Arabia): identified, characterized, and evaluated for gel treatments. Soc Pet Eng. https://doi.org/10.2118/97542-MS

Aldhaheri MN, Wei M, Bai B (2016) Comprehensive guidelines for the application of in-situ polymer gels for injection well conformance improvement based on field projects. Soc Pet Eng. https://doi.org/10.2118/179575-MS

Al-Hameedi AT, Dunn-Norman S, Alkinani HH, Flori RE, Hilgedick SA (2017a) Limiting drilling parameters to control mud losses in the Dammam formation, South Rumaila Field, Iraq. American rock mechanics association, August 28. http://www.onepetro.org/

Al-Hameedi AT, Dunn-Norman S, Alkinani HH, Flori RE, Hilgedick SA (2017b) Limiting drilling parameters to control mud losses in the Shuaiba formation, South Rumaila Field, Iraq. Paper AADE-17-NTCE-45, 2017 AADE national technical conference, Houston, Texas, April 11–12, 2017. http://www.AADE.org

Alvarado V, Thyne G, Murrell G (2008) Screening strategy for chemical enhanced oil recovery in Wyoming basins. Soc Pet Eng. https://doi.org/10.2118/115940-MS

Arshad U, Jain B, Ramzan M, Alward W, Diaz L, Hasan I, Aliyev A, Riji C (2015) Engineered solutions to reduce the impact of lost circulation during drilling and cementing in Rumaila Field, Iraq. This paper was prepared for presentation at the international petroleum technology conference held in Doha, Qatar, 6–9 December

Baker RO, Stephenson T, Lok C, Radovic P, Jobling R, McBurney C (2012) Analysis of flow and the presence of fractures and hot streaks in waterflood field cases. Soc Pet Eng. https://doi.org/10.2118/161177-MS

Basra Oil Company (2012) Various daily reports, final reports, and tests for 2007, 2008, 2009, 2010, 2011 and 2012. Several Drilled Wells, Basra's Oil Fields, Basra

Darley HC, Gray GR (1988) Composition and properties of drilling and completion fluids, vol 720, 6th edn. Gulf Professional Publishing, Oxford

Eriksson L, Johansson E, Kettaneh-Wold S, Trygg J, Wikstr¨om C, Wold S (2006) Multi- and megavariate data analysis. Part I. Basic principles and applications. Umetrics Academy, Montpellier

Hegde CM, Wallace SP, Gray KE (2015a) Use of regression and bootstrapping in drilling inference and prediction. Soc Pet Eng. https://doi.org/10.2118/176791-MS

Hegde C, Wallace S, Gray K (2015b) Using trees, bagging, and random forests to predict rate of penetration during drilling. Soc Pet Eng. https://doi.org/10.2118/176792

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning: with applications in R. Springer, New York. https://doi.org/10.1007/978-1-4614-7138-7

Leite Cristofaro RA, Longhin GA, Waldmann AA, de Sá CHM, Vadinal RB, Gonzaga KA, Martins AL (2017) Artificial intelligence strategy minimizes lost circulation non-productive time in Brazilian deep water pre-salt. Offshore Technol Conf. https://doi.org/10.4043/28034-MS

Messenger JU (1981) Lost circulation. PennWell Corp, Tulsa

Moore PL (1986) Drilling practices manual, 2nd edn. Penn Well Publishing Company, Tulsa

Nayberg TM, Petty BR (1986) Laboratory study of lost circulation materials for use in oil-base drilling muds. Paper SPE 14995 presented at the deep drilling and production symposium of the society of petroleum engineers held in Amarillo, TX, 6–8 April

Osisanya S (2002) Course notes on drilling and production laboratory. Mewbourne School of Petroleum and Geological Engineering, University of Oklahoma, Oklahoma **(Spring)**

Parks B (2010) Southern Iraq's Rumaila field kicks into high gear. http://www.drillingcontractor.org/southern-iraq%92s-rumaila-field-kicks-into-high-gear-7691. Accessed June 2016

SAS Institute Inc (2008) SAS/STAT® 9.2 user's guide. SAS Institute Inc, Cary

Tran TN, Afanador NL, Buydens LMC, Blanchet L (2014) Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). Chemom Intell Laboratory Syst 138:153–160. https://doi.org/10.1016/j.chemolab.2014.08.005 **(ISSN 0169-7439)**

Transparency Market Research (2013) Drilling fluids market (oil-based fluids, synthetic-based fluids and water-based fluids) for oil and gas (offshore & onshore)—global industry analysis, size share, growth, trends and forecast, 2012–2018, p 79. http://www.transparencymarketresearch.com/drillingfluid-market.html. Accessed July 2017

Tufféry S (2011) Data mining and statistics for decision making. Wiley, Chichester, West Sussex

Wallace SP, Hegde CM, Gray KE (2015) A system for real-time drilling performance optimization and automation based on statistical learning methods. Soc Pet Eng. https://doi.org/10.2118/176804-MS

Wold S, Johansson E, Cocchi M (1993) PLS—partial least squares projections to latent structures. In: Kubinyi H (ed) 3D QSAR in drug design, theory, methods, and applications. ESCOM Science Publishers, Leiden, pp 523–550

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