APPLICATION OF MACHINE LEARNING IN WELL PERFORMANCE PREDICTION, DESIGN OPTIMIZATION AND HISTORY MATCHING

A Dissertation

by

ADITYA VYAS

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Akhil Datta-Gupta
Committee Members,	Michael J. King
	Bani K. Mallick
	Duane A. McVay
Head of Department,	A. Daniel Hill

August 2017

Major Subject: Petroleum Engineering

ABSTRACT

Finite difference based reservoir simulation is commonly used to predict well rates in these reservoirs. Such detailed simulation requires an accurate knowledge of reservoir geology. Also, these reservoir simulations may be very costly in terms of computational time. Recently, some studies have used the concept of machine learning to predict mean or maximum production rates for new wells by utilizing available well production and completion data in a given field. However, these studies cannot predict well rates as a function of time. This dissertation tries to fill this gap by successfully applying various machine learning algorithms to predict well decline rates as a function of time. This is achieved by utilizing available multiple well data (well production, completion and location data) to build machine learning models for making rate decline predictions for the new wells. It is concluded from this study that well completion and location variables can be successfully correlated to decline curve model parameters and Estimated Ultimate Recovery (EUR) with a reasonable accuracy. Among the various machine learning models studied, the Support Vector Machine (SVM) algorithm in conjunction with the Stretched Exponential Decline Model (SEDM) was concluded to be the best predictor for well rate decline. This machine learning method is very fast compared to reservoir simulation and does not require a detailed reservoir information. Also, this method can be used to fast predict rate declines for more than one well at the same time.

This dissertation also investigates the problem of hydraulic fracture design optimization in unconventional reservoirs. Previous studies have concentrated mainly on

optimizing hydraulic fractures in a given permeability field which may not be accurately known. Also, these studies do not take into account the trade-off between the revenue generated from a given fracture design and the cost involved in having that design. This dissertation study fills these gaps by utilizing a Genetic Algorithm (GA) based workflow which can find the most suitable fracturing design (fracture locations, half-lengths and widths) for a given unconventional reservoir by maximizing the Net Present Value (NPV). It is concluded that this method can optimize hydraulic fracture placement in the presence of natural fracture/permeability uncertainty. It is also concluded that this method results in a much higher NPV compared to an equally spaced hydraulic fractures with uniform fracture dimensions.

Another problem under investigation in this dissertation is that of field scale history matching in unconventional shale oil reservoirs. Stochastic optimization methods are commonly used in history matching problems requiring a large number of forward simulations due to the presence of a number of uncertain variables with unrefined variable ranges. Previous studies commonly used a single stage history matching. This study presents a method utilizing multiple stages of GA. Most significant variables are separated out from the rest of the variables in the first GA stage. Next, best models with refined variable ranges are utilized with previously eliminated variables to conduct GA for next stage. This method results in faster convergence of the problem.

DEDICATION

I dedicate this dissertation to my parents, my wife, my brother and my friends for their support during my studies at Texas A&M University.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Akhil Datta-Gupta for his continued guidance during entire period of my PhD study. His expanse of knowledge and readiness to listen to my problems made it possible for me to study in this department of petroleum engineering without any bottlenecks. I would also like to thank him for continued financial support during my entire PhD studies.

I would like to thank Dr. Michael King and Dr. Bani K. Mallick for their continued interest in my research studies. Their immense knowledge and invaluable comments during my presentations guided me to the right direction and also helped me to continue my PhD without any bottlenecks. I would like to thank Dr. Srikanta Mishra from Battelle for his invaluable suggestions regarding Machine Learning study included in this dissertation. His immense knowledge and guidance always helped me when I needed them. I would also like to thank Dr. Duane A. McVay for being a member in my committee.

I would also like to thank Phaedra Hopcus, Barbi Miller and Eleanor Schuler for their help during various occasions particularly with the paperwork involved during this graduate program.

I would also like to thank my colleagues in my research group at the department of Petroleum Engineering, Texas A&M University – Jixiang Huang, Kenta Nakajuma, Hyunmin Kim, Hye Young Jung, Changdong Yang, Atsushi Iino, Tsubasa Onishi,

Hongquan Chen, Feyisayo Olalotiti-Lawal, Xue Xu, Rongqiang Chen and Gill Hetz - for their invaluable suggestions.

I would also like to alumni of this research group – Xia Xiaoyang, Yanbin Zhang, Peerapong Ekkawong, Jichao Han, Kam Dongjae, Muhammed Al-Rukabi, Jeongmin Kim, Neha Bansal, Shingo Watanabe, Shusei Tanaka and Zheng Zhang - for their invaluable suggestions.

Finally, I would like to thank my professors in University of Oklahoma (where I did my Masters studies) who recommended me to this PhD program – Dr. Deepak Devegowda, Dr. Ramadan Ahmed and Dr. Jeffrey G. Callard.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This PhD dissertation work was supervised by Dr. Akhil Datta-Gupta (Committee Chair) and other three committee members - Dr. Michael J. King, Dr. Bani K. Mallick and Dr. Duane A. McVay.

Chapter II of this dissertation study involving Machine Learning based study includes various suggestions made by Dr. Srikanta Mishra from Battelle. This work has been accepted for presentation in one of the SPE conferences before the end of year 2017.

Chapter III of this dissertation involving hydraulic fracture optimization study was done in collaboration with Changdong Yang. Changdong Yang provided the upscaling code (Oda Method) and Fast Marching Method (FMM) based forward simulator for this study. This work has been published in Journal of Petroleum Science and Engineering (2017) with a modified workflow.

Chapter IV of this dissertation involving History Matching in Shale Oil reservoirs has been done in collaboration with Atsushi Iino. Atsushi Iino provided the Fast Marching Method (FMM) based reservoir simulator used for this study. This work has been presented in SPE conference (SPE 185719-MS) with a modified workflow and would also be presented in an upcoming URTeC conference (URTeC: 2693139) with a modified workflow.

All remaining work in this dissertation has been done independently by Aditya Vyas.

Funding Sources

This work was made possible by the financial support of the member companies of Model Calibration and Efficient Reservoir Imaging (MCERI) consortium.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND OBJECTIVES

1.1 Introduction

Reservoir Simulations in large and complex reservoirs can be very costly. Specifically, in unconventional reservoirs, where reservoir models are usually represented by millions of grid cells, oil and gas production forecasts can take a lot of time. Many times, an engineer wants to get a quick idea about how a given well will deplete in future so as to calculate the revenues that will be generated later on. Also, this may be needed even before a detailed geologic information about a new well is provided. Previously, studies have been done to predict maximum/mean oil production in a field using machine learning approaches (LaFollette et. al, 2012 and 2013; Zhong et al., 2015). However, these studies could not predict rate decline with time. The method presented in this chapter can predict decline curve model parameters and predict rate decline for a new well based on data collected from the field. This method is very fast after the needed data has been gathered and properly cleaned/tabulated. In this chapter, this method has been applied to calculate rate decline parameters of four commonly used decline models and also to predict Estimated Ultimate Recovery (EUR) for a new well. This may provide an early estimate of well production for a new well. Also, previous studies involved utilizing a single model based predictions which is not a robust method since it would bias the model towards the training data/machine learning tuning parameters. This chapter takes

advantage of a model averaging technique to make predictions based on weighted average of multiple models built using more than one set of data/tuning parameters.

Another problem under investigation is of finding an optimum hydraulic fracturing design in unconventional reservoirs. Previous studies in the literature involved application of analytical models (e.g., PKN model) to predict well production. However, these models are built for conventional reservoirs and are not suitable to be used in unconventional reservoirs. Also optimization of hydraulic fractures in a given permeability field has been presented earlier (Ma et. al, 2013). However, their study did not take into account the uncertainty in the permeability field. The workflow presented in this chapter can be used to optimize hydraulic fracture design for a given reservoir provided with some uncertainty in the geologic data. This study also discusses uncertainty in the natural fracture distribution and its effects on the Net Present Value (NPV). A synthetic reservoir model has been used for this study and optimization problem is solved for maximizing the NPV.

This study also deals with a field-scale case history matching problem in which a base model and parameters with their uncertainty are provided and a genetic algorithm based history matching approach is utilized. Previous studies related to this work involved history matching using a single set of uncertain parameters with a wide range of uncertainty ranges. This chapter study utilizes a multi-stage GA approach that can be used to identify key parameters (heavy-hitters) before proceeding to history matching. First stage of this workflow involves using only the key parameters and matching observed data. In subsequent stages, the refined variables achieved from the first stage are utilized with reduced uncertainty ranges in them. The variables not included in the first stage are

also included in the subsequent stages. This method accelerates the convergence of a stochastic history matching parameter which in this study is Genetic Algorithm (GA). This study also integrates GA with a Fast Marching Method (FMM) based reservoir simulator which is a faster alternative to commonly used commercial simulators. In this study, simulated cumulative oil, water and gas production have been matched with their corresponding observed/history data provided by the field operator. A production forecast has also been made and corresponding production has been compared to test the accuracy of history matching algorithm.

1.2 Dissertation Outline

This dissertation document contains several chapters each containing a different case study. In Chapter II, Eagle Ford well data has been gathered from a publicly available website and used with several machine learning algorithms in order to build models that can predict rate declines for a new well. This method is very fast after the needed data has been gathered and properly cleaned/tabulated. It can be used to calculate rate decline parameters of commonly used decline models and also to predict Estimated Ultimate Recovery (EUR) for a new well. This may provide an early estimate of well production for a new well.

In Chapter III, a detailed workflow for hydraulic fracture design optimization has been presented. This workflow based on genetic algorithm can be used to optimize hydraulic fracture design for a given reservoir provided the geologic data including permeability and porosity is known. This study also briefly discusses about the uncertainty

in the natural fracture distribution and its effects on the optimization of Net Present Value (NPV). A synthetic reservoir model has been used for this study and optimization problem is solved for maximizing the NPV.

In Chapter IV, a field case study has been presented in which a set of uncertain parameters/variables with production history data are provided and objective is to match history data by applying genetic algorithm based workflow. A multi-stage GA approach has been used in this study to accelerate the convergence of GA. The multi-stage GA approach utilizes heavy hitter variables in the first stage to fine tune the variables making most impact. Subsequent stages, however include all variables with updated uncertainty ranges. Simulated cumulative oil, water and gas production have been matched with their corresponding observed/history data provided by the field operator. A production forecast has also been made and corresponding actual production has been compared to test the accuracy of history matching algorithm.

Finally, in Chapter V, conclusions from this dissertation study have been presented and recommendations for possible extension/improvement to current work are suggested.

CHAPTER II

MACHINE LEARNING BASED INSIGHTS ON WELL PERFORMANCE IN EAGLE FORD WELLS

2.1 Introduction and Literature Review

Oil and gas wells have been in existence for a long time but it was only in recent times when importance of large sets of well data are realized by the petroleum industry. A large set of well data which includes well location data and well completion data are becoming available in a format that can be easily used by data scientists. Since shale oil and gas revolution started in USA, a large number of wells have been drilled and their data collected. Many of these data are available in publically accessible websites on internet. This chapter deals with a study done using well data collected from more than 100 wells in the Eagle Ford reservoir. Well data used for this study include well location/depth parameters including latitude, longitude and total vertical depth and well completion parameters including number of hydraulic fractures, volume of fracturing fluid used, amount of proppant used, and completed length. Well data has been collected from the online database DrillingInfo. Only oil wells have been selected for this study.

Lee et al. (2002) applied classification and non-parametric regression algorithms for electrofacies characterization and permeability prediction in complex reservoirs. Model based clustering technique was used to identify clusters from well log responses. For each cluster, non-parametric regression technique was utilized to build model and predict corresponding permeability. The non-parametric regression algorithms include

ACE (Alternating Conditional Expectation), GAM (Generalized Additive Model) and NNET (Neural Networks). ACE based regression algorithm outperformed the other two regression methods in this study.

Perez et al. (2005) applied classification trees with well log response to predict electrofacies, lithofacies and hydraulic flow units in uncored wells. This study also reported the predictor variables that have most influence in classification tree based prediction. It was also reported that larger trees may be too sensitive to the statistical noise present in the data and therefore smaller (pruned) trees should be used for such kind of study.

Mishra (2012) reported a method to make predictions based on multiple models instead of single one. The final prediction is based on weighted average of predictions from all models. It was shown that more than one decline model can be fitted to a data with acceptable accuracy. However, their future predictions may vary a lot. To overcome this problem, the final predicted response variable, Estimated Ultimate Recovery (EUR) was predicted using multiple models aggregated together by Generalized Likelihood Uncertainty Estimation or GLUE (Beven and Binley, 1992; Neuman, 2003; Singh et al. 2010) methodology.

LaFollette and Holcomb (2011) presented data analytic results using Barnett shale horizontal wells. It was found that wells more than 3,500 – 4,500 ft of lateral length were less efficient in terms of production per foot. Also, it was found that, most wells are drilled in approximately 140 and 320 degrees of azimuth. Also, the best wells were those that were drilled near horizontal.

LaFollette et al. (2012) reported results for Bakken formation of the Eastern Williston Basin. They found production efficiency (production per foot of completed lateral) decreases with increasing lateral length. It shows that increasing number of stages and completed length alone did not find positive correlation with maximum monthly oil production (calculated during first 12 month production period). However, proppant concentration seemed to have a positive correlation with maximum monthly oil production.

LaFollette et al. (2012) presented results of North Texas Barnett Shale wells with emphasis on well completion and fracture stimulation. It was concluded in this paper that traditional linear regression methods are not suitable for this kind of data: prone erroneous data, missing data, non-linear data and data containing subtle interrelationships among variables. It was concluded that boosted tree method is more suited for this kind of data for regression purposes. The study also found a good correlation between maximum monthly oil production and amount of fracturing fluid used for fracking in the wells studied.

LaFollette (2013) presented data analytics results from Barnett shale and Bakken Shale. In Barnett shale case, relative influence of various variables in predicting maximum monthly gas production during first 12 month period was studied. TVD is found to be the most influential factor in this study using boosted tree model. In Bakken shale case, relative influence of various variables in predicting maximum monthly oil production during first 12 month period was studied. In this case, well location coordinates were found to be most influential in the study done using boosted tree model.

LaFollette et al. (2013) reported results using well data gathered from Bakken Light Tight Oil Play. This study was carried out using multivariate analysis of production data. It was found that well location that can be used as a proxy for reservoir quality is one of the most influential predictor for production forecast. It was also concluded that longer lateral wells are less efficient in terms of production per feet of lateral length.

LaFollette et al. (2014) reported results using well data gathered from Eagle Ford Formation in South Texas. This study carried out multivariate analysis on Eagle Ford production data. Reservoir quality was proxied by X-Y surface location since petrophysical data was unavailable. The completion variables used for this study included proppant amount, volume of fracturing fluid used, number of fracturing stages, and completed length (measured as difference between measured depths of bottom perforation and top perforation). Other variables included dip, azimuth and GOR. The proxies for production efficiency include maximum oil rate, barrels of oil produced per unit completed length and barrels of oil produced per pound of proppant used. The paper also reported trends in reservoir fluid parameters.

The study reported that GOR and well location are among the most important variables influencing multivariate analysis. This study also reported that even though production rates increases with increase in completed lateral length, the production per unit completed length reduces as completed length increases. Increase in proppant amount used for completion jobs is found to increase productivity in terms of maximum monthly production.

Holcomb et al. (2015) studied the productivity effects from spatial placement and well architecture in Eagle Ford shale horizontal wells. This study found that wells drilled and completed in GOR less than 5000 scf/bbl have lower maximum monthly oil production (during first 12 month period) per foot of length but then appear to have a lower percentage decline rate than higher GOR wells. This study could not find direct correlation between increased proppant consumption and increased well productivity.

Zhong et al. (2015) reported their results with Wolfcamp shale. They applied several machine learning algorithms to build models that can predict first 12 months of cumulative oil for oil wells. Machine learning algorithms used included Ordinary Least Squares (OLS), Support Vector Machines (SVM), Random Forests (RF) and Gradient Boosting Model (GBM). In their results, RF modeled the data most accurately. Also, they reported the predictor relative importance based on R^2 loss. In this method, each of the predictor variable was removed from predictor set one at a time while keeping rest of the predictors intact and checking the change in R^2, i.e., R^2 loss. The predictor having more R^2 loss associated with it is considered more important. Different machine learning algorithms had different ranking/predictor importance order in this study. In case of RF, fracturing fluid amount used for completion job turned out to be most influential factor.

Schuetter et al. (2015) reported their machine learning study using data set comprising wells in Wolfcamp Shale in West Texas (Delaware Basin and Central Basin). Response variable in this study was cumulative production in the first 12 months of oil production period. This study tried to predict first 12 month cumulative production for new test data wells based on machine learning models developed using training wells.

Machine learning algorithms used here were Ordinary Least Squares, Random Forest, Gradient Boosting Machine, Support Vector Regression (SVR) and Kriging. K-fold cross-validation technique was utilized to avoid overfitting. It was found that although Kriging based models fits training data perfectly, they did not perform well for test data. Also, study includes relative importance study of various predictor variables. It was found that TVD is most influential predictor among all predictors.

Centurion et al. (2012) presented their data analytics results using Eagle Ford well data. It was pointed out that most of the top productive wells in Eagle Ford lie in the counties of Dewitt and Karnes. However, the worst performing wells are not located in a particular location. Also, the wells completed using delayed release production chemicals have higher productivity than those which didn't use those chemicals. In the multivariable statistical analysis, most dominant predictors were identified and they included proppant volume, injection rates, treatment pressure, measured depth of deepest perforation, production chemicals combined with stimulation fluids and porosity indicator.

Centurion et al. (2013) reported their multivariate analysis results using Eagle Ford well data. The most significant variables found in their study were proppant per ft, pressure, cluster spacing, thickness, average porosity and perforation length.

Centurion et al. (2014) reported their data analytic results using LaSalle County wells in Eagle Ford shale. Cumulative oil production during first 3 months was considered as a proxy for well productivity. Multivariate analysis results showed most influential variables in this region to be completed length and stage spacing. Proppant pumped showed positive correlation with well productivity. Also, increased shut-in time between

hydraulic fracture treatment and the first day of production also had a positive effect on well productivity. Reduction in well spacing led to lower initial productivity but increased overall productivity of the region in a longer term.

2.2 Methodology

Eagle Ford well data has been downloaded from drillinginfo (website: info.drillinginfo.com). More than 100 well data has been collected and analyzed using various machine learning techniques. First, well data has been analyzed using exploratory data analytic techniques such as scatterplot and boxplot. Next, machine learning techniques such as Random Forest (RF), Gradient Boosted Machine (GBM), Support Vector Machine (SVM) and Multivariate Adaptive Regression Splines (MARS) have been utilized in order to predict rate decline in Eagle Ford wells. Since the production rate data of these wells are mostly noisy, it is difficult to model them with smooth models. However, a novel approach explained in this section can handle this problem using machine learning algorithms in conjunction with decline rate models used in oil industry. The well rate data is first fitted with one of the commonly used decline models listed below.

2.2.1 Rate Decline Models

2.2.1.1 Arp's Decline Model

Arp's decline equation (Arps, 1945) can be represented as follows:

$$q(t) = \frac{q_i}{(1+bD_i t)^{\frac{1}{b}}} \tag{2.1}$$

where,

$q(t)$ = rate at time t (STB/D)

q_i = initial rate (STB/D)

D_i = initial decline rate (1/month)

b = hyperbolic decline coefficient (dimensionless)

t = time (months)

Exponent b in above equation shows type of decline in a well (**Table 2.1**).

Table 2.1: Exponent 'b' in Arp's decline curves

b value	Decline type
b = 0	Exponential
0 < b < 1	Hyperbolic
b = 1	Harmonic

Fig. 2.1 shows an example well's predictions made by Arp's decline model keeping Initial flow rate, D_i same but varying exponent, b. It may be seen that for higher b values, model predicts higher production rates.

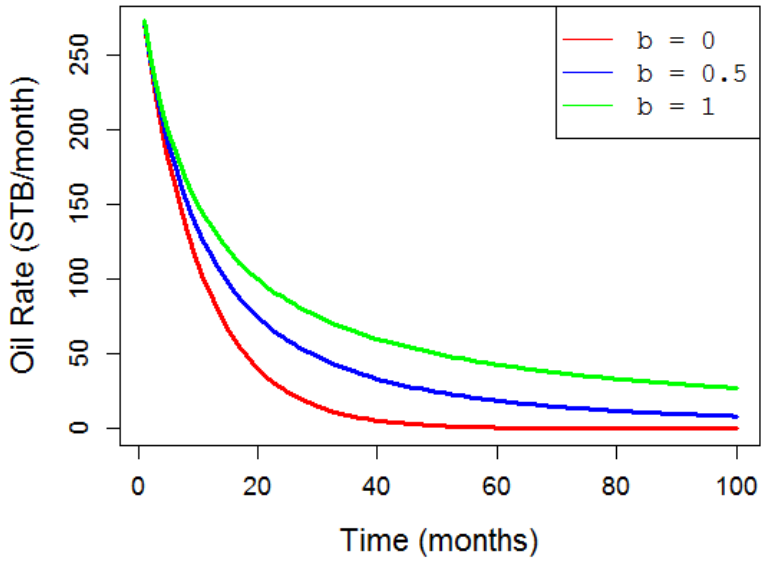


Figure 2.1 An example well prediction made by Arp's decline model

2.2.1.2 Stretched Exponential Decline Model (SEDM)

Valko and Lee (2010) presented Stretched Exponential Decline Model which is a specialized decline model for unconventional reservoirs and predicts rate decline in transient flow regime. Since unconventional wells produce in transient flow regimes, SEDM is more suitable for them compared to Arp's decline model. **Eq. 2.2** shows SEDM equation.

$$q(t) = q_i exp\left[-\left(\frac{t}{\tau}\right)^n\right] \quad (2.2)$$

where,

$q(t)$ = rate at time t (STB/D)

q_i = initial rate (STB/D)

τ = characteristic relaxation time (month)

n = exponent parameter (dimensionless)

t = time (months)

Johnston (2006) explained stretched exponential decay process as a sum of exponential decay with a "fat tailed" probability distribution of time constants. Valko and Lee (2010) explained SEDM to be a sum of large number of individual exponential decays. It was also reported by Valko and Lee (2010) that Arp's may predict physically unrealistic Estimated Ultimate Recovery (EUR) values for b ≥ 1 but SEDM will always give finite value of EUR. **Fig. 2.2** shows how Arp's can fit early rate data really well but would over predict production at long term period.

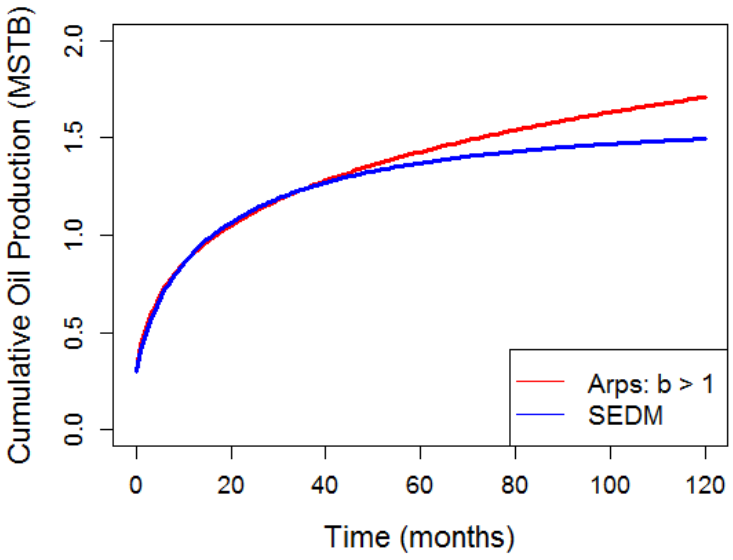


Figure 2.2 Comparison of Arp's and SEDM decline models

2.2.1.3 Duong Model

Duong (2011) presented following equation in the case of fracture dominated flow characteristics. This equation (**Eq. 2.3**) is derived empirically for shale gas and tight gas reservoirs.

$$q(t) = q_1 t^{-m} exp\left(\frac{a}{1-m}(t^{1-m} - 1)\right) \quad (2.3)$$

where,

$q(t)$ = rate at a time t (STB/D)

q_1 = flow rate on first day (STB/D)

a = intercept constant

m = slope parameter. Duong (2011) showed that for the unconventional reservoirs m > 1

t = time (months)

2.2.1.4 Weibull Model

Another way to model decline curve is through Weibull growth curve (Weibull, 1951; Mishra, 2012).This equation (**Eq. 2.4**) is generally used for modeling time-to-failure in applied engineering problems.

$$P(t) \equiv G_P = M\left\{1 - exp\left(-\left(\frac{t}{\alpha}\right)^{\gamma}\right)\right\} \quad (2.4)$$

where,

G_P = cumulative production at time t

M = carrying capacity (Max. cumulative production)

γ = shape parameter

α = scale parameter

t = time (months)

Differentiating **Eq. 2.4** gives (Weibull, 1951; Mishra, 2012):

$$\mathrm{q(t)} = M\frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1} exp\left(-\left(\frac{t}{\alpha}\right)^{\gamma}\right) \quad (2.5)$$

where,

$q(t)$ = rate at time t (STB/month)

M, the carrying capacity, is the maximum cumulative production set by this equation. This means that cumulative production cannot reach unrealistic values as in the Arp's model in some cases. Since it is a fitting parameter like α and γ, a close approximate value of M is needed to fit Weibull curve on a well rate decline data. For this study, cumulative well oil production during the available well oil production period with ± 10 % margin has been assumed for best range within which M should lie. α , the scale parameter, is that value of time at which (1-1/e) or 63.2% of the resources have been produced (Mishra, 2012). γ, the shape factor, shows how rate of growth changes with time and is usually less than 1 for unconventional reservoirs (Mishra, 2012).

Once the well rate data is collected for all the wells included in this study, all of the above decline models are used to fit them with a best match and the parameters of corresponding decline models are stored for further study. Also, the Estimated Ultimate Recovery (EUR) for each well is calculated as a numeric integral of monthly oil production over 30 year period (360 months):

$$EUR = \sum_{i=1}^{360} q_i \quad (2.6)$$

where,

EUR = Estimated Ultimate Recovery

q_i = monthly oil rate (STB/month) of i^{th} month

2.2.2 Machine Learning Algorithms

Once well rate data is collected and fitted with the decline models discussed previously, the data is tabulated such that each row corresponds to a well and each column corresponds to one of the variables (predictors or responses). **Table 2.2** shows the response and predictor variables used for each of the decline curve models. As shown in **Table 2.2**, predictor variable are unchanged across each of the decline models but response variables change.

The data table is divided randomly into 80% - 20% partition so that 80% of the rows are utilized to train machine learning model (called as training data) and remaining 20% of the rows are used for testing (called as test data) the model accuracy. In this study, different machine learning algorithms have been applied to the data under investigation. Following subsections briefly presents the main idea behind some of these algorithms that provided better results than the remaining ones. The three machine learning algorithms that produced better prediction results than others are: Random Forests (RF), Gradient Boosted Machines (GBM) and Support Vector Machines (SVM). However, results for Multivariate Adaptive Regression Splines (MARS) are also shown in this chapter for comparison purposes only.

Table 2.2 Response variables of decline models for Machine Learning

	Arp's	SEDM	Duong	Weibull
Response Variables	D_i, b, EUR	tau, n, EUR	a, m, EUR	γ, α, M, EUR
Predictor Variables	Well Latitude and Longitude, TVD, Difference between TVDs of Heel and Toe, Completed Length, Number of Fracture stages, Amount of fracturing fluid and Proppant used for fracking			

Once a model has been trained, it can then predict the decline curve parameters of new wells which in this case are test data wells. Oil rate decline with respect to time can then be predicted by using decline curve parameters and corresponding decline equation.

This study also deals with finding the relative influence of various predictor variables for building a model. This can be regarded as a variable importance or sensitivity study in which it is possible to identify most important and least important predictor variables to build a model.

A short description of four of the machine learning algorithms applied to Eagle Ford study is presented in the following sections.

2.2.2.1 Random Forests (RF)

Breiman (2001) reported an ensemble based learning method based on Classification and Regression Trees (CART) concept. A single Classification Tree consists of a series of partition such that each partition divides data points into two

dissimilar groups as shown in **Fig. 2.3 (a)**. However, in reality, a partition by linear boundaries may not be able to partition data into pure classes. This is shown in **Fig. 2.3 (b)** by impurities of whites among black colored circles and impurities of blacks among white circles. These impurities can be minimized by further partitioning the variable space. The mathematical quantity to be minimized here is called the Gini Impurity Index (Breiman, 1996) which is a measure of impurities present in a given partition/compartment.

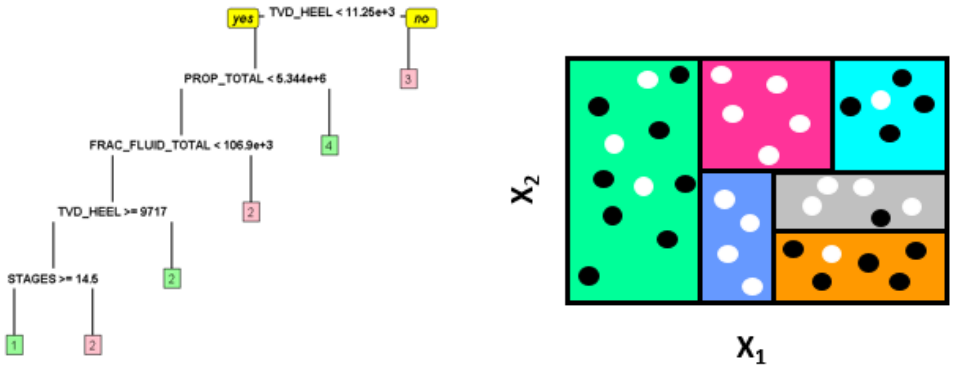
$$Gini\ Impurity\ Index = \sum_{i=1}^{k} p_i(1 - p_i) = 1 - \sum_{i=1}^{k} p_i^2 \quad (2.7)$$

where,

p_i = probability of training dataset belonging to i^{th} class

k = number of classes (or categorical variables)

In a pure node (consisting only of one type of class), this Gini Index should be equal to 0. In order to partition a variable space, different possibilities are tested including different variables and different point of partition in a given variable's range. This is repeated at each node until Gini's Index is minimized or number of terminal nodes exceed the specified set limit. The final prediction value at a terminal node is governed by majority vote.



(a) (b)

Figure 2.3 (a) Classification Tree example (b) Equivalent partition for a two variable case

Regression Trees are similar to a Classification Trees but in their case prediction is made for a continuous variable (real number) instead of a categorical variable (class) as shown in **Fig. 2.4**.

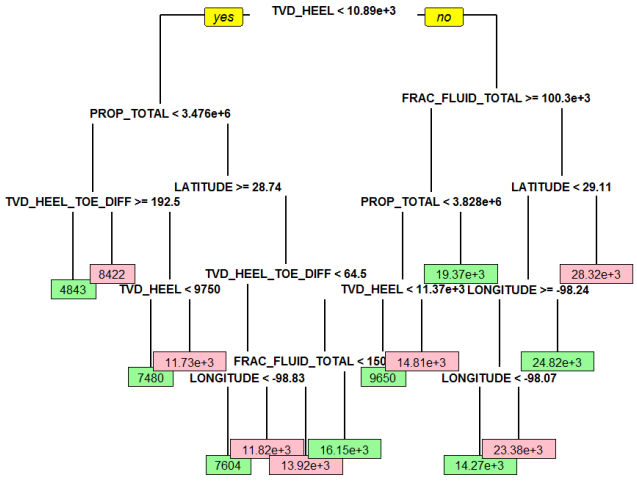


Figure 2.4 An example Regression Tree from Eagle Ford data predicting maximum oil production

The values at each node is calculated by minimizing Residual Sum of Squares (RSS) using **Eqs. 2.8** and **2.9** (Shalizi, 2006):

$$RSS = \sum_{c=1}^{n} \sum_{i=1}^{n_c} (y_i - m_c)^2 \quad (2.8)$$

$$m_c = \frac{1}{n_c} \sum_{i=1}^{n_c} y_i \quad (2.9)$$

where,

c = number of nodes

n_c = number of data points in a node

y_i = observed or actual response value

In order to partition a variable space, different possibilities are tested including different variables and different point of partition in a given variable's range. This is repeated at each node until RSS is minimized or number of terminal nodes exceed the specified set limit. The final prediction value at a terminal node is governed by mean prediction value. Cost Complexity (Cp) in a regression tree (Perez et. al, 2003) is given by:

Cp = Training Error + k × No. of terminal nodes (2.10)

where,

k = cost complexity factor. If k = 0, tree will not control no. of terminal nodes and only error rates are involved making tree larger than needed. If k is very large, tree will be very short with high training error and biased model

Fig. 2.5 shows C_p vs cross validation error/misfit error in Eagle Ford data. As can be seen in this figure tree size of 2 gives minimum Cp. However, it must be noted that a very small size of tree can bias the model for the training data. In the Random Forest

package in R, tree sizes are controlled by providing a range within which total number of terminal nodes should lie. This is an indirect way of controlling Cp. The *default minimum number of nodes is 5 for regression trees* in Random Forest package used in this study. Therefore in the example shown below, the tree size of 5 would be appropriate.

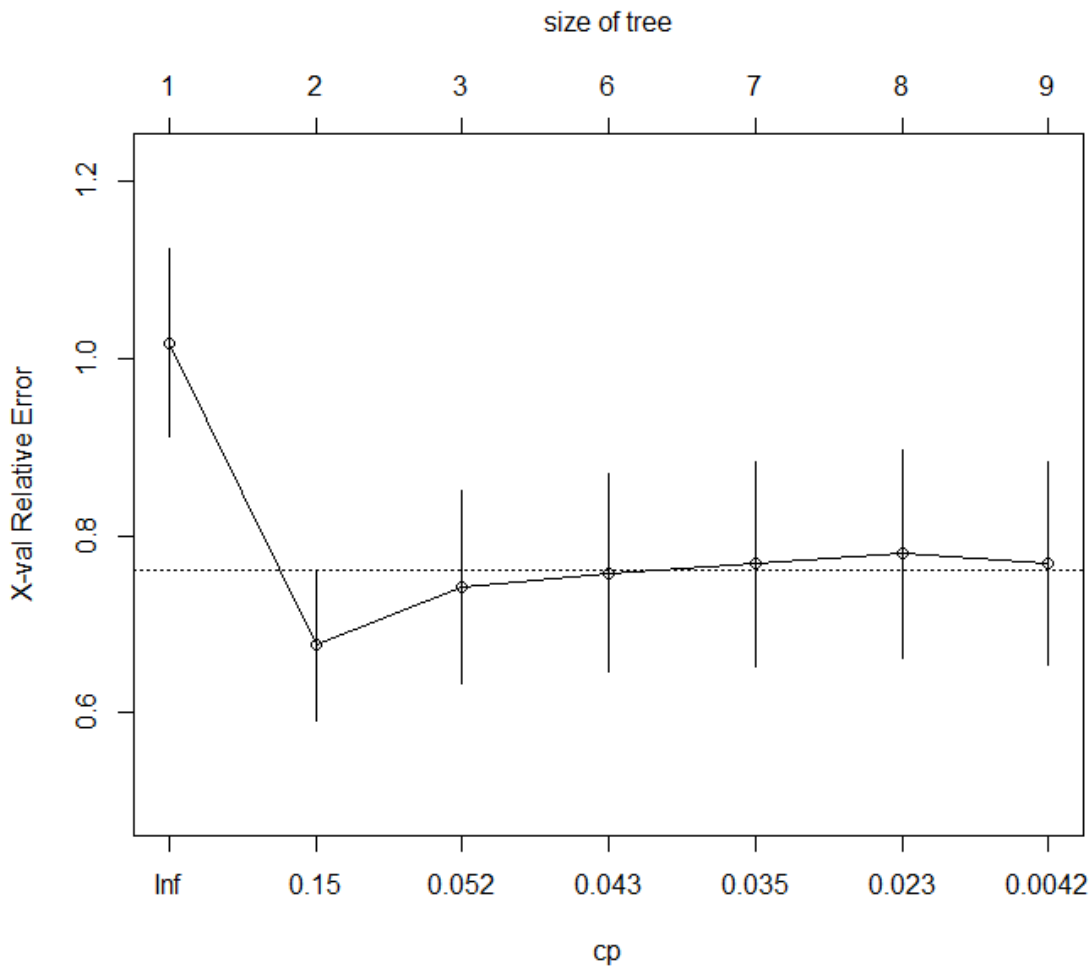


Figure 2.5 Cost complexity and size of a regression tree against misfit error using Eagle Ford data

A Random Forest (Breiman, 2001) is an ensemble based machine learning algorithm which is comprised of a large number of uncorrelated trees (Classification or Regression Trees). Instead of fitting data with a single Classification or Regression Tree, a random forest of multiple uncorrelated trees is constructed. Each tree is derived from a bootstrap subsample of given data as well as a bootstrap subsample of variables from predictor variable set leading to a different order of partitioning. During prediction process for a new dataset (not used for training the Random Forest), final prediction is based on majority vote (Random Forest of Classification Trees) or averaged response (Random Forest of Regression Trees).

2.2.2.2 Gradient Boosted Machine (GBM) Regression

Gradient Boosted Machine (Friedman, 2001 and 2002) is an ensemble tree based machine learning algorithm in which a true model is represented by a series of trees such that each subsequent tree is fitting the error residual of the previous tree (**Fig. 2.6**). Friedman (2001 and 2002) reported that "Gradient Boosting of the regression trees produces competitive, highly robust, interpretable procedures for both regression and classification, especially mining less than clean data".

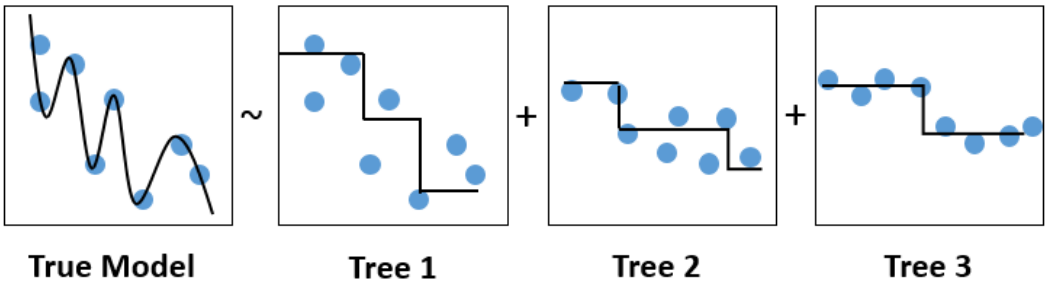


Figure 2.6 Approximate representation of a Gradient Boosted Tree Model (Modified from Gradient Boosted Regression Trees in scikit-learn, https://www.slideshare.net/DataRobot/gradient-boosted-regression-trees-in-scikitlearn)

A simple mathematical formulation of gradient boosted trees is presented below (source: scikit-learn.org website (http://scikit-learn.org/stable/modules/ensemble.html).

A general form of additive model is given by:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{2.11}$$

γ_m = step length

$h_m(x)$ = basis functions

The gradient boosting additive model can be represented as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2.12}$$

where,

$h_m(x)$ = regression/classification tree used as a basis functions/weak learners

For each stage, $h_m(x)$ is chosen to minimize the loss function L for the given model F_{m-1} and its fit $F_{m-1}(x_i)$

$$F_m(x) = F_{m-1}(x) + arg \min_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x)) \tag{2.13}$$

This minimization problem is solved numerically via steepest descent method.

$$F_m(x) = F_{m-1}(x) + \gamma_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i)) \quad (2.14)$$

where,

$$\gamma_m = arg \min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right) \quad (2.15)$$

The initial model, $F_0(x)$ is usually chosen to be the mean of target values in case of regression problems.

2.2.2.3 Support Vector Machines (SVM) Regression or Support Vector Regression (SVR)

Smola and Schölkopf (2004) presented Support Vector Regression (SVR) or Support Vector Machine (SVM) Regression which has become quite successful among machine learning algorithms. This algorithm tries to fit function, f(x), on a given training dataset such that the maximum deviation of a data point from this function is equal to **ε**. However, complexity of f(x) is controlled so that f(x) is kept as flat as possible.

Eq. 2.16 shows the term that is needed to be minimized and **Eq. 2.17** shows that constraints used while minimizing **Eq. 2.16**.

Objective is to find: $f(\vec{x}) = \vec{w}.\vec{x} + b$, by:

minimizing: $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$ (2.16)

subjected to constraints: $\begin{cases} y_i - (\vec{w}.\vec{x} + b) \leq \varepsilon + \xi_i \\ y_i - (\vec{w}.\vec{x} + b) \leq -(\varepsilon + \xi_i^*) \\ \xi_i, \xi_i^* > 0 \end{cases}$ (2.17)

Eq. 2.16 also shows the slack term variables (Cortes and Vapnik, 1995, Smola and Schölkopf, 2004) in order to avoid overfitting in the model. The second term in **Eq. 2.16**

shows the cost term containing slack variables, ξ_i, ξ_i^* which include points with deviations more than ε . By controlling the constant C (where C > 0), the contribution of the second term in **Eq. 2.16** can be controlled. This is also a way to control the trade-off between the flatness of f(x) and the limit up to which data points having deviations larger than ε are tolerated in the machine learning model. Using Lagrange multipliers (α_i, α_i^*) to solve above minimization problem, the above equations become:

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\, x_i \tag{2.18}$$

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) < x_i, x > + b \tag{2.19}$$

where,

α_i, α_i^* = Lagrange multiplier

$<.,.>$ = dot product

Aizerman et al. (1964) and Nilsson (1965) showed how to map a training data to some feature space $\mathcal{F}$ i.e., $\Phi: X \rightarrow \mathcal{F}$. This process simplifies the problem such that the optimization problem tries to find function f(x) in the feature space and not in actual input space.

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{2.20}$$

Once the data is in feature space, the function f(x) to be fitted can be more flat than fitting it in original data space.

2.2.2.4 Multivariate Adaptive Regression Splines (MARS)

Freidman (1991 and 1993) reported Multivariate Adaptive Regression Splines (MARS). **Eq. 2.21** shows the basic form of MARS:

$$\hat{f}(X) = a_0 + \sum_{m=1}^{M} a_m B_m(X) \quad (2.21)$$

where,

a_0 = constant

$\{a_m\}_1^M$ are the coefficients of expansion whose values are determined by least square fit of above equation:

$$\{a_m\}_1^M = \underset{\{a_m\}_1^M}{\operatorname{argmin}} \sum_{n=1}^{n} [y_n - a_m B_m(X)]^2 \quad (2.22)$$

X = $\{x_1, x_2, \ldots, x_p\}$ = variables in training data set

$B_m(X)$ = basis function

A basis function can be a constant, a hinge function or a product of any combination of one or more hinge functions. A hinge function is of following form:

$$[x - t]_+ = max(0, x - t) = \begin{cases} x - t, & if\ x > t \\ 0, & otherwise \end{cases} \quad (2.23)$$

$$[t - x]_+ = max(0, t - x) = \begin{cases} t - x, & if\ x < t \\ 0, & otherwise \end{cases} \quad (2.24)$$

In above equations, the constant *t* is called as a knot, which is a point at which model function f(X) changes direction. The final form of MARS equation becomes (Friedman 1991):

$$\hat{f}(X) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[\pm\left(x_{v(k,m)} - t_{km}\right)\right]_+ \quad (2.25)$$

where,

$\{v(k,m)\}_1^{K_m}$ = variable set associated with m^{th}basis function B_m

The training process in MARS algorithm consists of a Forward Pass and a Backward Pass. During Forward Pass, a pair of terms are added at each step until a pre-

specified limit of maximum number of terms is reached. On the contrary, during the Backward Pass, the least effective term is removed in each step (one term at a time). To decide which term needs to be discarded, Generalized Cross-Validation is used. **Eq. 2.25** gives the formula to calculate GCV. It is proportional to the data fitting error but inversely proportional to the number of terms in the model. GCV is a trade-off between the number of terms and the Mean Squared Error (MSE) and helps dealing with the problem of overfitting in MARS. Generalized Cross Validation (GCV) is calculated as:

$$GCV = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{f}_M(x_i)\right]^2}{\left[1-\frac{C(M)}{N}\right]^2} \tag{2.26}$$

y_i = observed values

$\hat{f}_M(x_i)$ = model predicted values

N = no. of observations/predictions

$C(M)$ = cost complexity function $\propto$ no. of basis functions used in model

At the end of the forward pass, an over fit MARS model larger than needed terms is trained. Backward pass or the pruning pass consists of removing terms from existing MARS equation in steps and checking GCV. GCV should first decrease to a minimum value before taking off again. At that point optimum number of terms are achieved. **Fig. 2.7** shows a GCV plot for a MARS model with Eagle Ford data. In this figure the removal of terms should be stopped at step number 5.

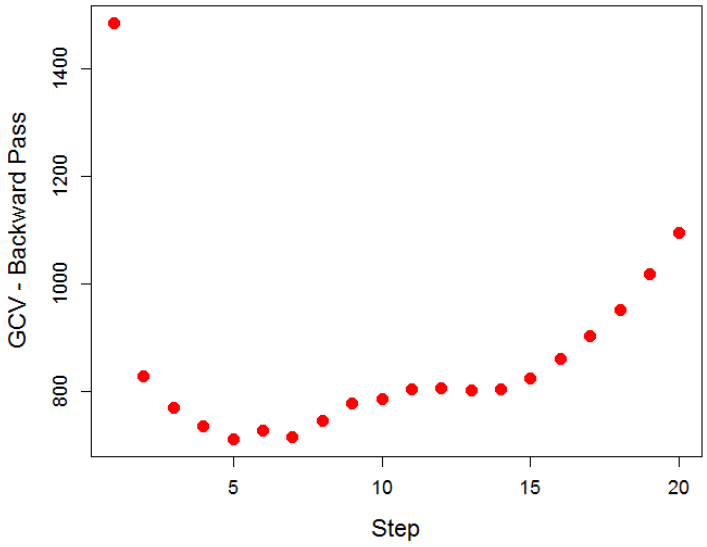


Figure 2.7 An example of GCV plot using Eagle Ford data

2.2.3 Model Averaging

One of the usual practice to train a machine learning model is to use an entire training dataset by minimizing the training data misfit. Another way is to use a k-fold cross validation approach. This dissertation section involves k-fold validation approach for calculation of misfit. **Fig. 2.8** shows steps for training a machine learning model using this approach. Once raw well data is collected which in current study is from Eagle Ford database, each oil well's rate decline is fitted with one of the four decline models – Arp's (Arp's 1945), SEDM (Valko and Lee, 2010), Duong (Duong, 2011) or Weibull (Weibull, 1951 and Mishra, 2012). The corresponding parameters of these decline models are then derived based on best fit **(Table 2.2)**. The dataset now contains both predictor variables and response variables. Outlier points are removed based on engineering judgement, e.g., wells having unrealistic proppant mass or fluid volumes are removed. This dataset is now

split into 80% training data and 20% test data. Test data is not used for training any of the Machine Leaning models in this study. Training data is further split into 10-folds (k = 10). As shown in **Fig. 2.8**, various combinations of training data subset and test data subset can be derived from main training data. This training data can be used to train a machine learning Model with different input values of tuning parameters provided in the grid form to the training data set. Therefore, each of the training data subset set with one of the tuning parameter combination results in a single machine learning model which is tested against corresponding test data subset resulting in an error calculated in terms of RMSE. A large number of such models with corresponding RMSE errors are then used to predict the main test data (not used for training purposes). However, since each model will predict a different value of a response variable, a model averaging technique known as Generalized Likelihood Uncertainty Estimation or GLUE is utilized here to combine the outputs of all trained machine learning models and result in single output prediction. Model averaging helps dealing with problem of overfitting.

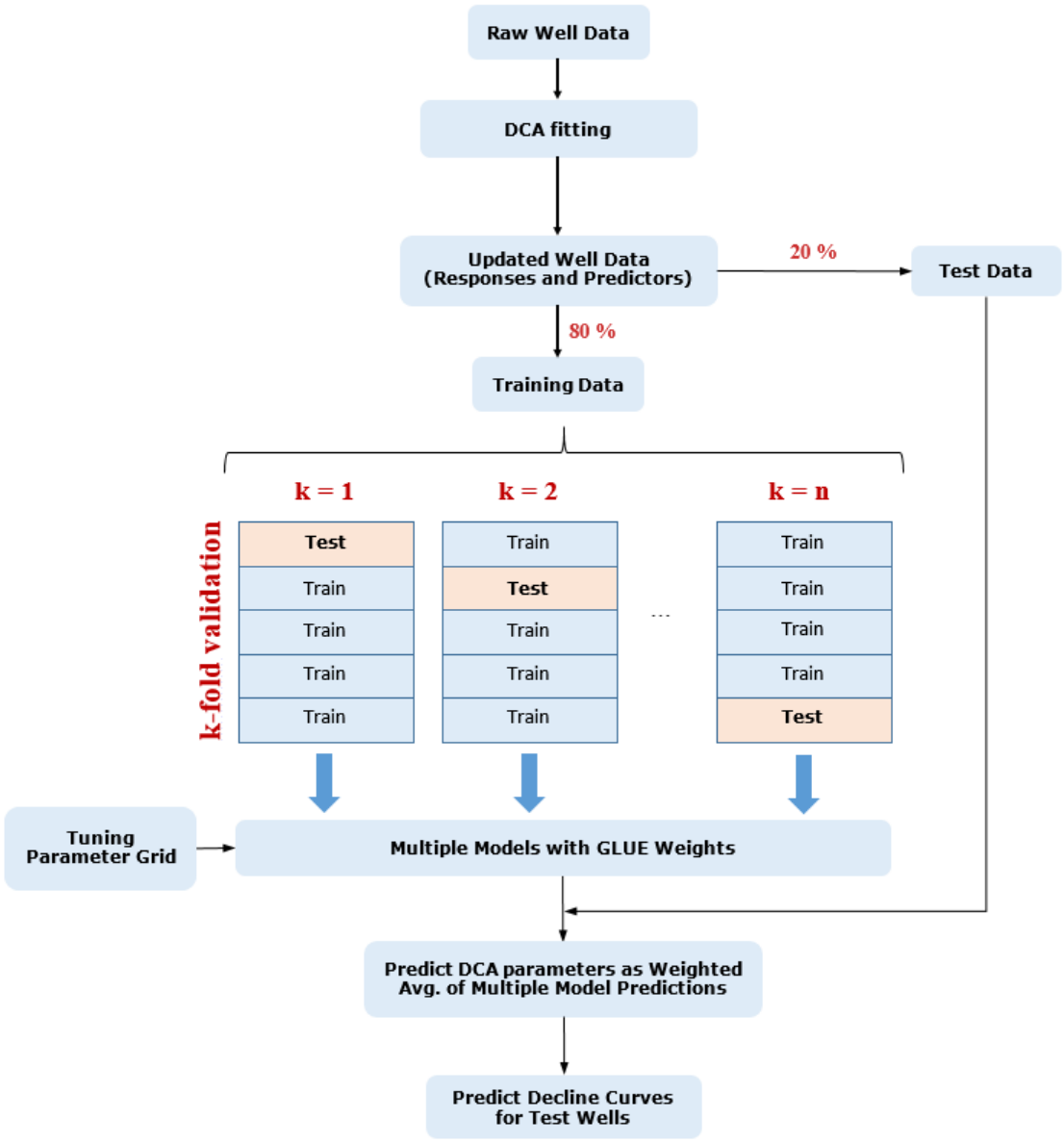


Figure 2.8 Workflow steps for model training and prediction

2.2.3.1 Generalized Likelihood Uncertainty Estimation (GLUE)

Generalized Likelihood Uncertainty Estimation or GLUE is derived from Bayesian Model Averaging. **Eq. 2.27** shows Bayesian Model Averaging method. This method calculates the weights for individual models and the final output prediction is

weighted average of all models. For a given model *j*, its weight is given by (Draper 1995, Kass and Raftery 1995 and Hoeting et al. 1999):

$$weights, w_j \propto p(M_j|D) = \frac{p(D|M_j)p(M_j)}{\sum_j p(D|M_j)p(M_j)} \quad (2.27)$$

where,

$p(M_j)$ = prior probability of Model *j*

$p(D|M_j)$ = model likelihood given by prediction error for data D

$$= \int P(d|\theta_j, M_j)p(\theta_j|M_j)d\theta_j$$

$P(d|\theta_j, M_j)$ = joint probability of a model *j* (function of prediction errors)

$p(\theta_j|M_j)$ = prior probabilities of parameters

Since it is difficult to calculate the likelihood integral, Beven and Binley (1992) and Beven (2000) proposed GLUE formula which simplified **Eq. 2.27** with **Eq. 2.28**.

$$p(D|M_j) \propto exp\left[-N\frac{{\sigma_{e,j}}^2}{{\sigma_o}^2}\right] \quad (2.28)$$

where,

N = shape factor

$\sigma_{e,j}$ = variance of the errors of model $j = \frac{Sum\ of\ squared\ errors}{no.of\ observations}$

σ_o = variance in the observed data

$N \gg 1$ tends to give higher weightage to models with less fitting error

$N \ll 1$ tends to give similar weights to all models

Therefore, model weights are given by:

$$weights, w_j \propto p(M_j|D) = \frac{exp\left[-N\frac{{\sigma_{e,j}}^2}{{\sigma_o}^2}\right]p(M_j)}{\sum_j exp\left[-N\frac{{\sigma_{e,j}}^2}{{\sigma_o}^2}\right]p(M_j)} \quad (2.29)$$

A modified GLUE formula has been proposed by Mishra (2012) which simplifies **Eq. 2.29** even further:

$$p(D|M_j) \propto \left(\frac{{\sigma_o}^2}{{\sigma_{e,j}}^2}\right)^N \quad (2.30)$$

$$weights, w_j \propto p(M_j|D) = \frac{\left(\frac{{\sigma_o}^2}{{\sigma_{e,j}}^2}\right)^N p(M_j)}{\sum_j\left(\frac{{\sigma_o}^2}{{\sigma_{e,j}}^2}\right)^N p(M_j)} \quad (2.31)$$

or,

$$p(D|M_j) \propto \frac{1}{{RMSE_j}^2} \quad (2.32)$$

where,

$RMSE_j$ = Root Mean Square Error of model j to observed data

$$weights, w_j \propto p(M_j|D) = \frac{\frac{1}{{RMSE_j}^2}p(M_j)}{\sum_j\frac{1}{{RMSE_j}^2}p(M_j)} \quad (2.33)$$

Finally, the final output response from multiple models can be derived from weighted sum of individual responses from all models as:

$$Response = \sum_{j=1}^{no.\,of\ models} w_j Response_j \quad (2.34)$$

2.2.4 Relative Influence of Predictor Variables

Relative influence of a predictor variable is calculated as the relative change in the RMSE (Root Mean Squared Error), AAE (Average Absolute Error) or R^2 (Coefficient of

Determination) if a given predictor is removed from the training data set and rest of the steps remain unchanged during model training process.

Eq. 2.35 shows the formula to calculate relative influence of p^{th} predictor using R^2. From **Eq. 2.35**, it can be seen that relative influence of a predictor variable is its proportion of variance that is predictable from a model. Relative Influence of a p^{th} predictor is given by:

$$RI_p = abs\left(\frac{{R^2}_p - {R^2}_{-p}}{{R^2}_p}\right) \tag{2.35}$$

where,

${R^2}_p = R^2$ of model with all predictors included

${R^2}_{-p} = R^2$ of model with all predictors except p^{th} predictor are included

Eq. 2.35 can be applied to other two metrics – RMSE and AAE by replacing R^2 by RMSE and AAE respectively. **Eqs. 2.36** and **2.37** shows formulas to calculate RMSE and AAE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.36}$$

$$AAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.37}$$

Eq. 2.38 (Schuetter et. al, 2015) shows how $pseudo\ R^2$ can be calculated. This version of R^2 indicates the proportion of variance in the response/dependent variable that is predictable from a model.

$$pseudo\ R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.38}$$

where,

y_i = observed value of i^{th} data point

$\hat{y}_i$ = predicted value of i^{th} data point

$\bar{y}$ = mean of observed values

Another metric that can be utilized here is normalized mean-standard deviation ratio **(Eq. 2.39)**. Instead of R^2, Median to Sigma ratio is utilized to create relative influence plots. However, this ratio has been normalized w.r.t corresponding ratio in observed data/actual data as in the case of R^2.

$$Normalized\ Median - Sigma\ Ratio = \frac{\left(\frac{Median}{\sigma}\right)_{predicted}}{\left(\frac{Median}{\sigma}\right)_{Observed}} \quad (2.39)$$

In this study, relative influence of a predictor variable is calculated by first calculating the quantity for model evaluation - RMSE, AAE or R^2 - including all the predictor variables in training data set ($R^2{}_p$) and then calculating it without including the predictor p in the training data set ($R^2{}_{-p}$). Finally, using **Eq. 2.35** will give relative influence of that predictor.

2.3 Eagle Ford Field Case Study

The Eagle Ford data is collected for about multiple wells from the commercial database Drillinginfo (https://info.drillinginfo.com/). The raw data is cleaned to remove outliers. Only the wells satisfying following criteria (about 100 wells) were used:

- Well Production Period > 12 months
- Initial flow rates < 40,000 STB/month

- STAGES > 4
- 50,000 bbl < Total Fracturing fluid < 200,000 bbl
- CLENGTH > 2000 ft
- Calculated EUR <= 300 MSTB
- Wells with too much noise in rate decline data.

Fig. 2.9 shows the pairwise scatter plots for various predictor variable data collected. It may be observed that a few pairs of variables shown in this figure have some correlation between them. For e.g., completed length, stages and total proppant amount seem to have some correlation among them. However, this study uses all these predictor variables in order to see the individual effects on regression and variable relative importance study.

The EUR value for each of the wells is calculated based on decline curve extrapolation to 30 years of production. Each of the four decline models would result in a different EUR for a given well. As an exploratory analysis, these EURs can be regressed by a regression tree to identify variables making more impact than others on EUR. **Figs. 2.10** through **2.13** show these regression trees. As is obvious from these figures, Initial Flow Rate, q_i, is clearly making the most impact on EUR among all decline models. Another way of doing this analysis is dividing the EUR range in Eagle Ford data into four groups or clusters based on quartiles. Cluster 1 contains wells with lowest EURs while cluster 4 contains the highest values of EURs. **Figs. 2.14** through **2.17** show results from the classification tree analysis for each of the decline models. Again, q_i comes out to be the most important variable.

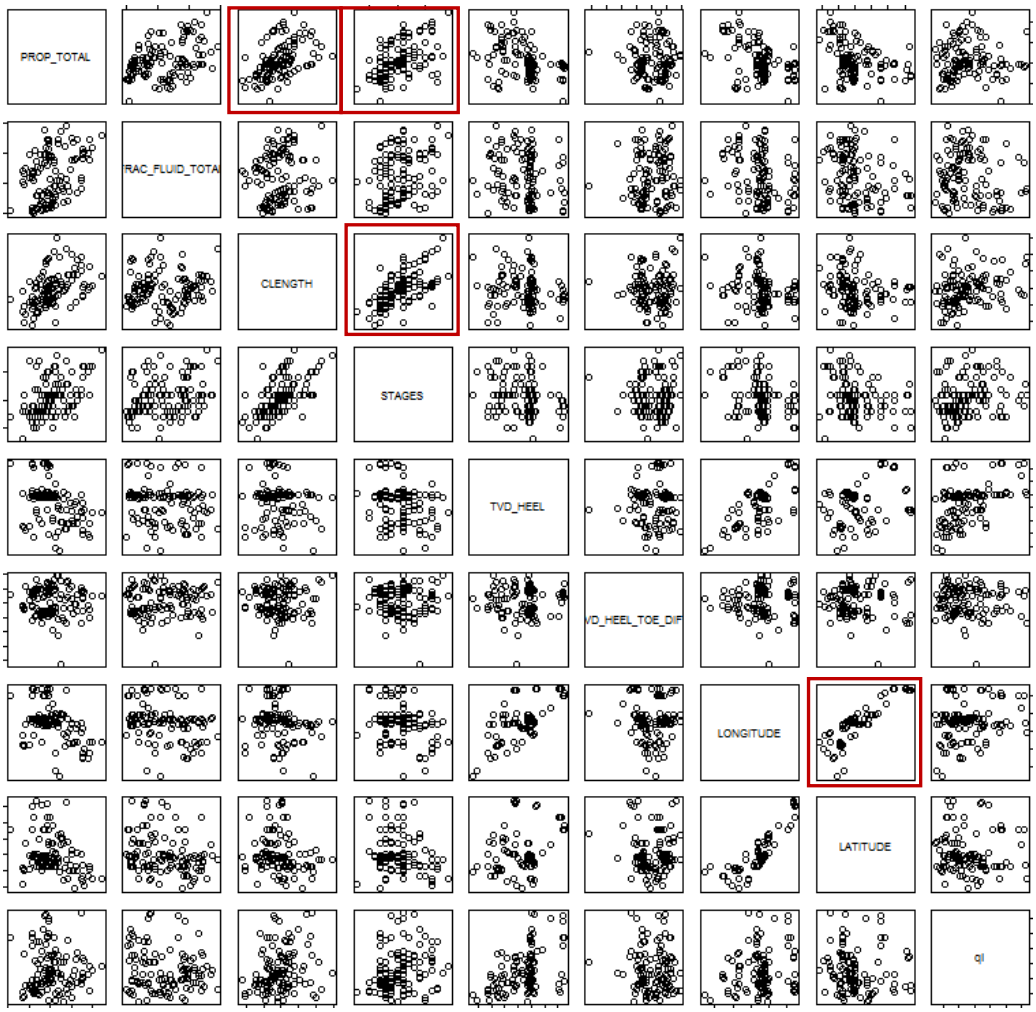


Figure 2.9 Pairwise scatterplots of various predictor variables in Eagle Ford data

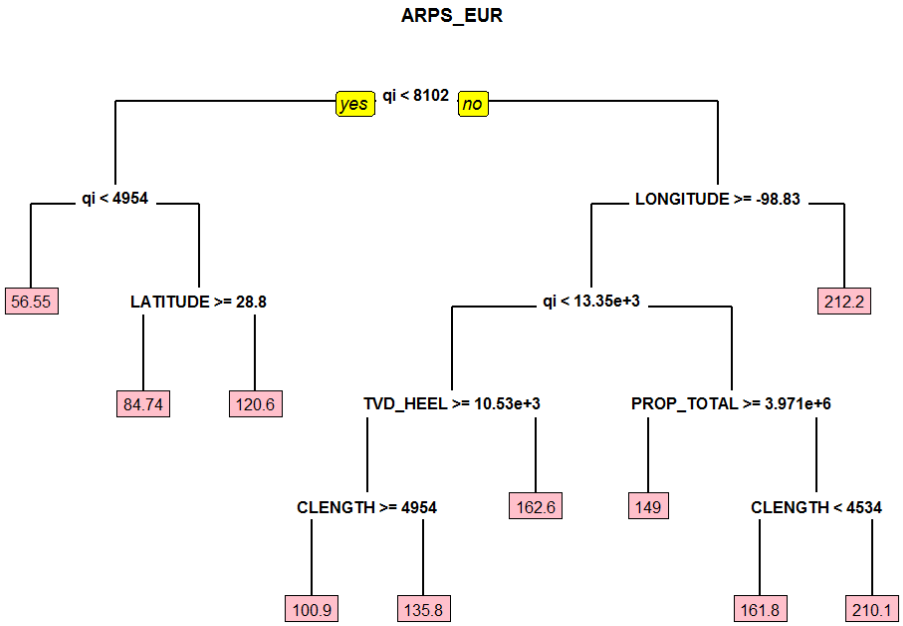


Figure 2.10 Regression Tree fitted on EUR calculated from Arp's Decline Model

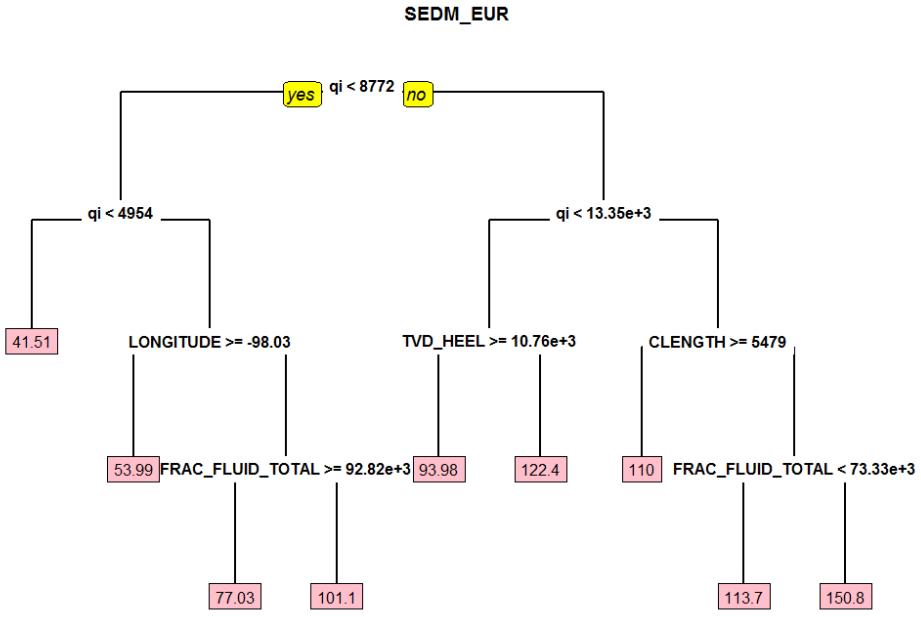


Figure 2.11 Regression Tree fitted on EUR calculated from SEDM Decline Model

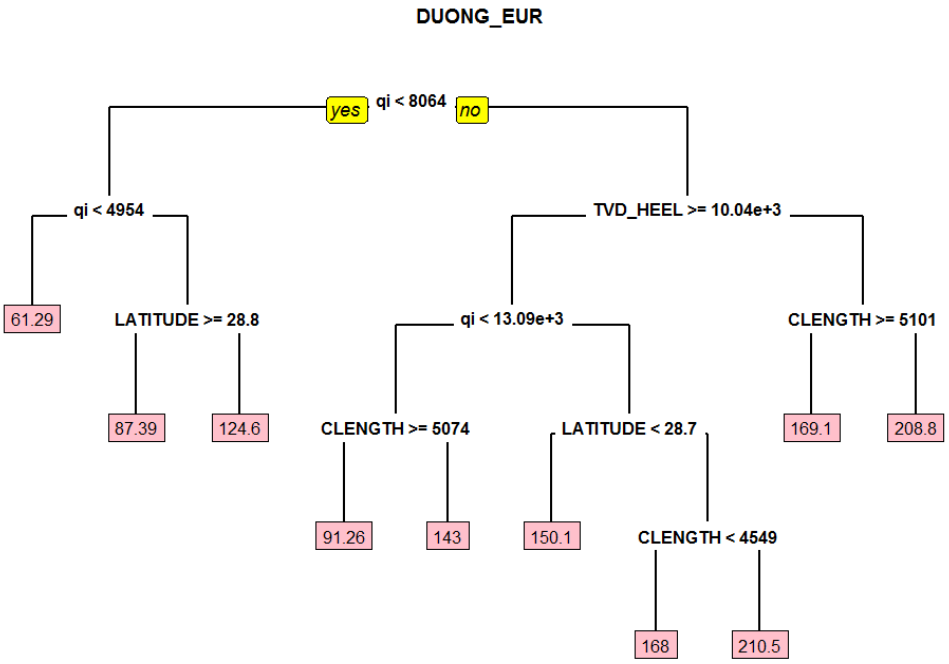


Figure 2.12 Regression Tree fitted on EUR calculated from Duong's Decline Model

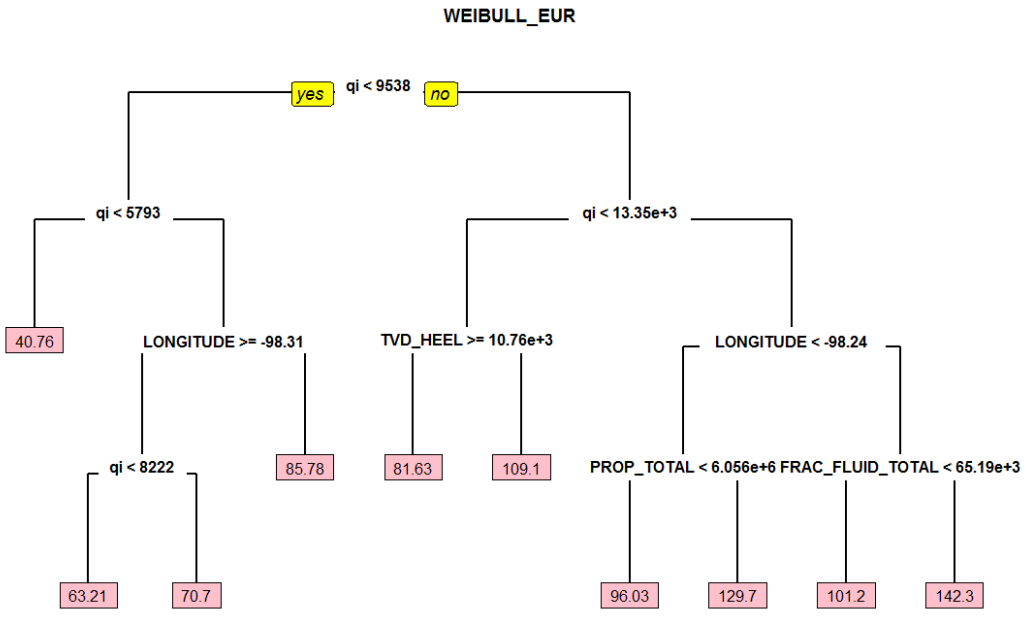


Figure 2.13 Regression Tree fitted on EUR calculated from Weibull's Decline Model

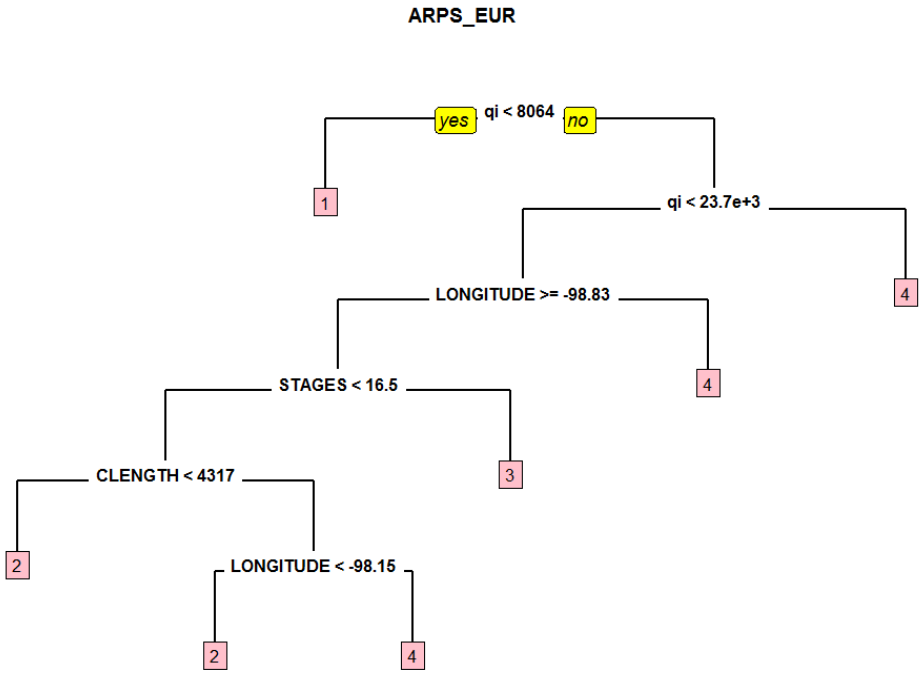


Figure 2.14 Classification Tree fitted on EUR clusters derived from Arp's Decline Model

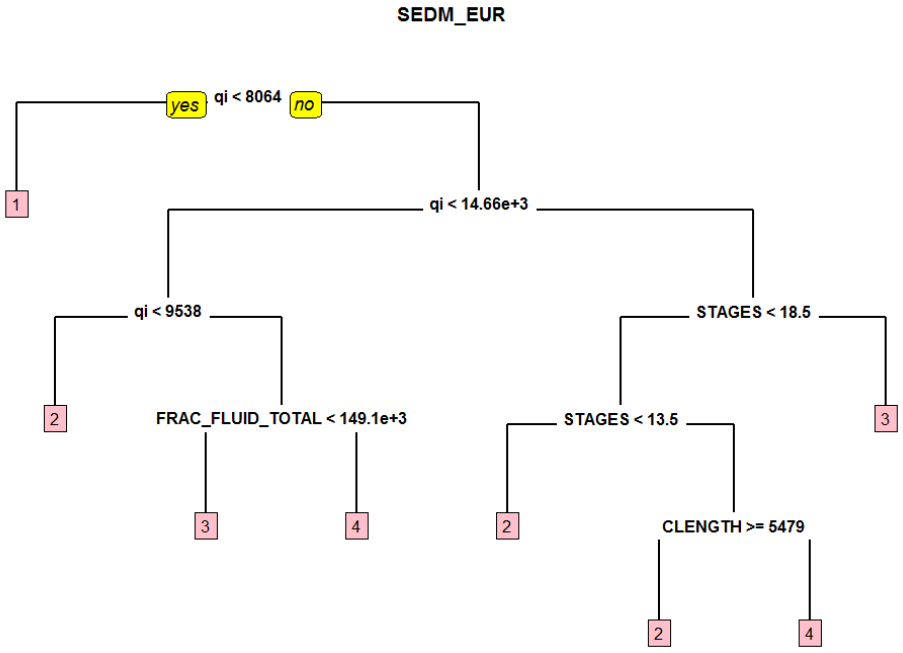


Figure 2.15 Classification Tree fitted on EUR clusters derived from SEDM Decline Model

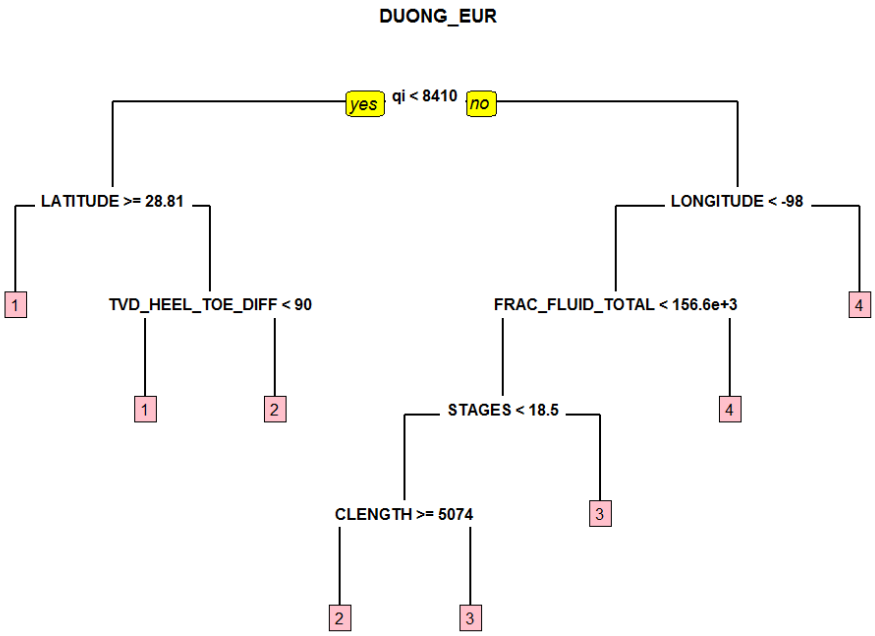


Figure 2.16 Classification Tree fitted on EUR clusters derived from Duong's Decline Model

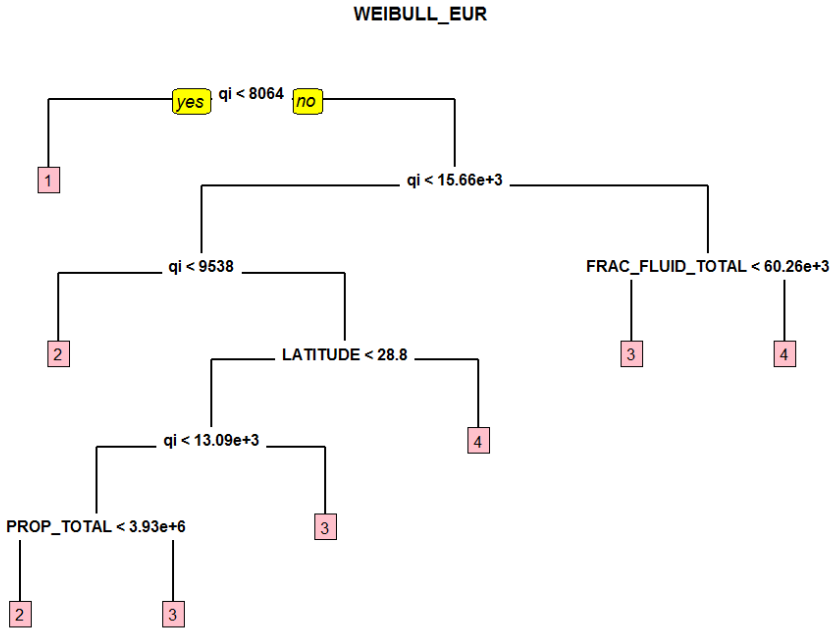


Figure 2.17 Classification Tree fitted on EUR clusters derived from Weibull's Decline Model

Based on previous results, qi has been identified to be the best candidate for clustering the well data for further analysis. As mentioned earlier, **Fig. 2.18** shows the 4 clusters created by dividing wells into four groups based on their Initial Flow Rates, q_i. **Fig. 2.19** shows the distribution of other predictors in these 4 clusters. It may be observed that cluster 4 which contains wells with highest Initial Flow Rates (q_i) also contains wells with highest Total Vertical Depths (TVD_HEEL) and Completed Lengths (CLENGTH) if median values of these boxplots are taken as the reference.

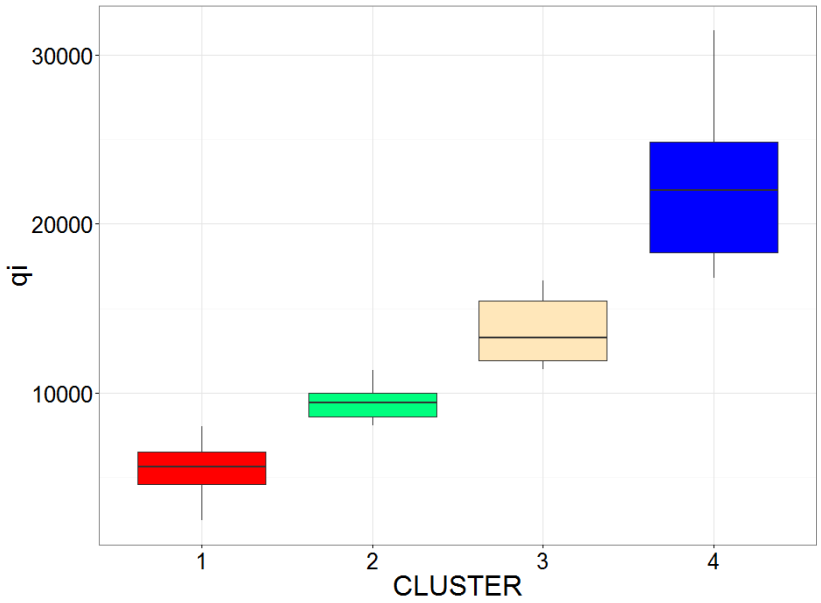


Figure 2.18 Well clusters based on Initial Flow Rate, q_i

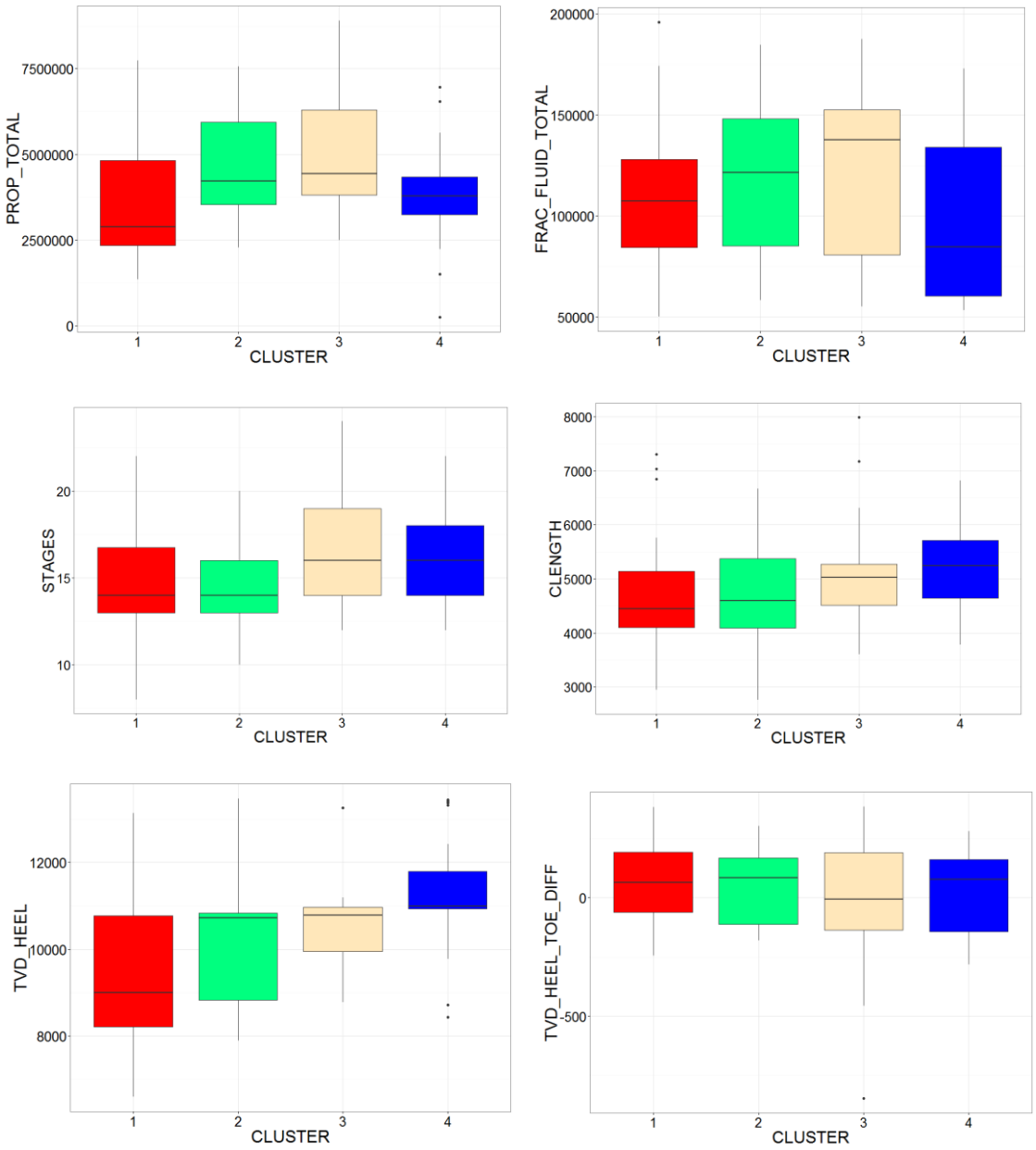


Figure 2.19 Predictor variable distribution in clusters derived from Initial Flow Rate, qi

Fig. 2.20 shows the location of the four clusters created based on Initial Flow Rate on the Texas map. **Fig. 2.21** shows wells in worst cluster 1 and best cluster 4 on map. Also shown in this figure is the spread of other study variables on the map. Only clusters 1 and

4 are included in these plots to view the difference between the highest Initial Flow Rate wells and Lowest Initial Flow Rate wells. It may be observed from these figures that most of the wells occurring in cluster number 4 are drilled in deepest depths. However, there are some exceptions to this observations shown in the map. This is because TVD is not be the only criteria to predict well production. However, only TVD_HEEL has some reasonable trend on the map.

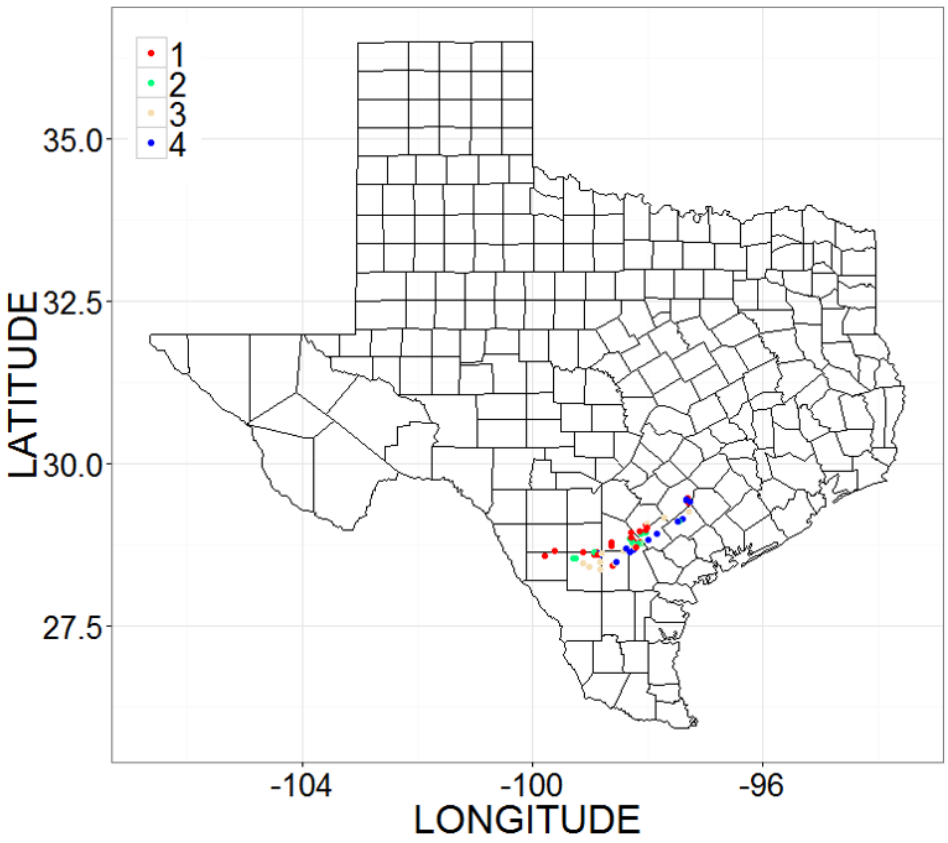


Figure 2.20 Study wells on Texas map color coded by cluster number

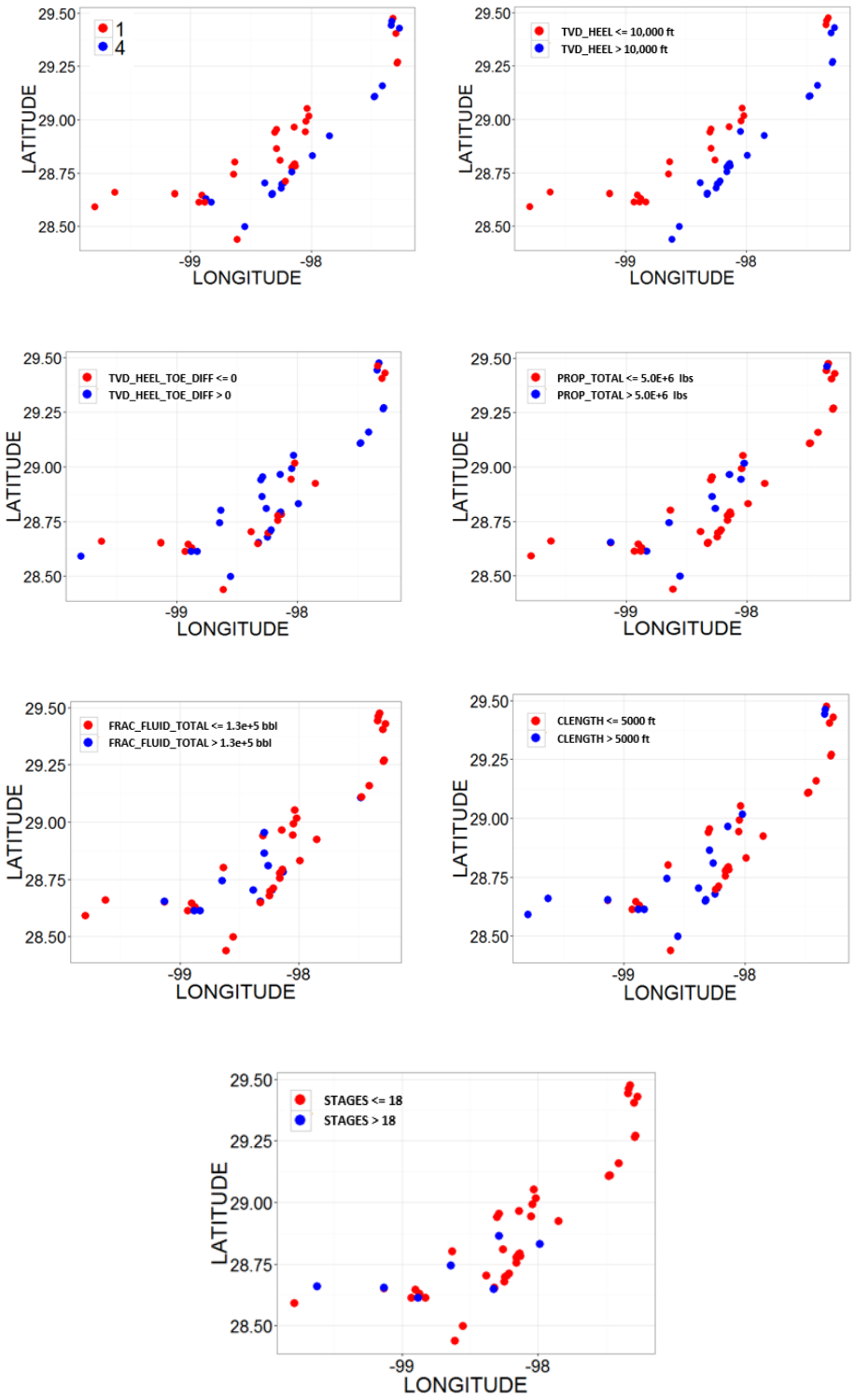


Figure 2.21 Correlation between cluster type and different variables

Fig. 2.22, 2.25, 2.28 and **2.31** show the comparison plots of different error metrics resulting from best fit of data using the 12 machine learning algorithms applied for this study. Best machine learning algorithm for each decline model is identified as the one which has lowest RMSE errors but R^2 to be close to unity. **Table 2.3** shows the best machine learning algorithms determined for each of the decline models.

Table 2.3 Most suitable Machine Learning algorithm for each decline model

Decline Model	Best Machine Learning Algorithm
Arp's	GBM
SEDM	SVM
Duong	GBM
Weibull	SVM

Figs. 2.23, 2.26, 2.29 and **2.32** show the scatterplots showing predicted versus actual values of a decline curve parameter/EUR for RF, GBM, SVM and MARS algorithms. **Figs. 2.24, 2.27, 2.30** and **2.33** show the predicted decline curves for test data wells for each of the decline models applying the best machine learning algorithm. **Fig. 2.34** shows the comparison plots of predictions made in **Figs. 2.24, 2.27, 2.30** and **2.33.**

Since each of the four decline models under investigation have a different set of decline model parameters, comparing them together is not possible. However, if we compare EURs for these decline models together, it may be easier to identify the best combination of decline model and machine learning algorithm to predict well performance

in Eagle Ford wells. **Fig. 2.35** shows such comparison between EURs predicted from the four decline models. It may be recalled here that EURs are estimated based on extrapolation of a decline curve for 30 year period. Therefore Actual EURs mentioned in these figures are calculated by extrapolating best fit decline curves using actual rate data. This means that a well can have a different EUR for each of the four decline models for the same well rate data. From **Fig. 2.35** it may be seen that SEDM and Weibull have better prediction results compared to other two decline models. It may also be noted that Arps and Duong's models are predicting higher range of EUR for the wells compared to SEDM and Weibull models. This may be the likely reason for inaccurate prediction of EUR at higher values in case of Arp's and Duong's models. It should also be recalled here that Weibull model would require an initial estimate of the carrying capacity to fit decline model curve on a well data. This is however not required in case of SEDM model. Therefore, this may be regarded as an advantage of SEDM model over Weibull model.

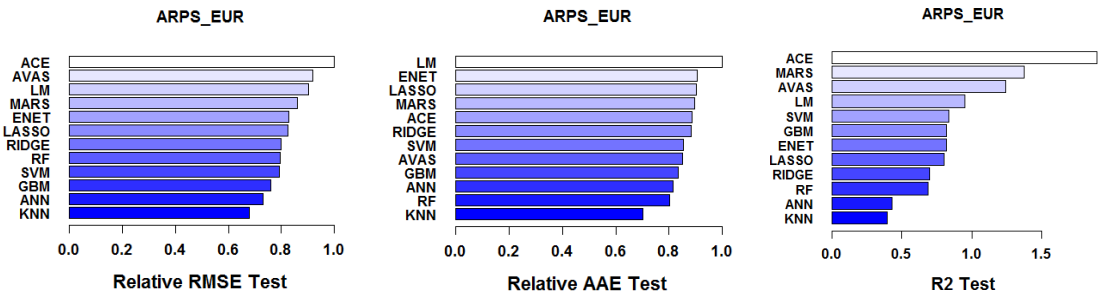


Figure 2.22 Error metric comparison for different machine learning algorithms taken into consideration for Arp's model

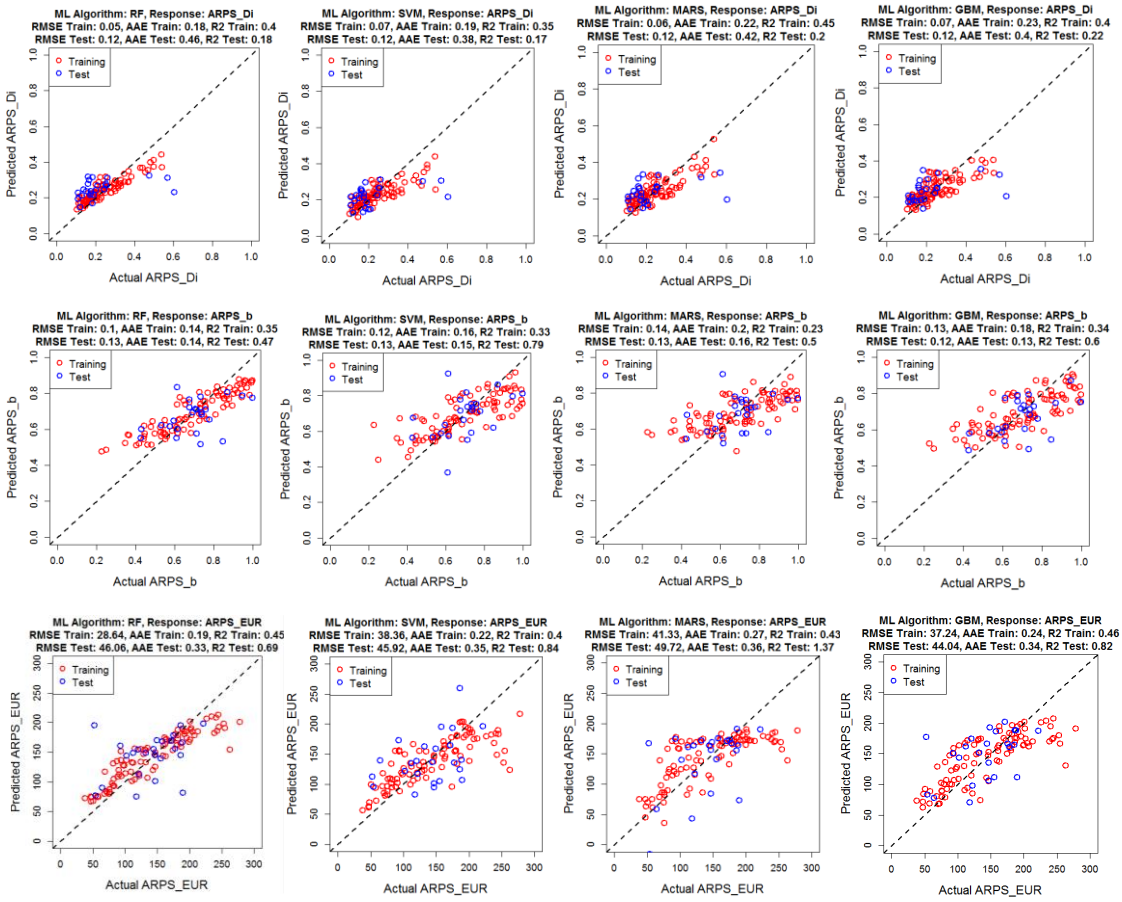


Figure 2.23 Scatterplots showing predicted vs actual values of Arp's decline model parameters and EUR

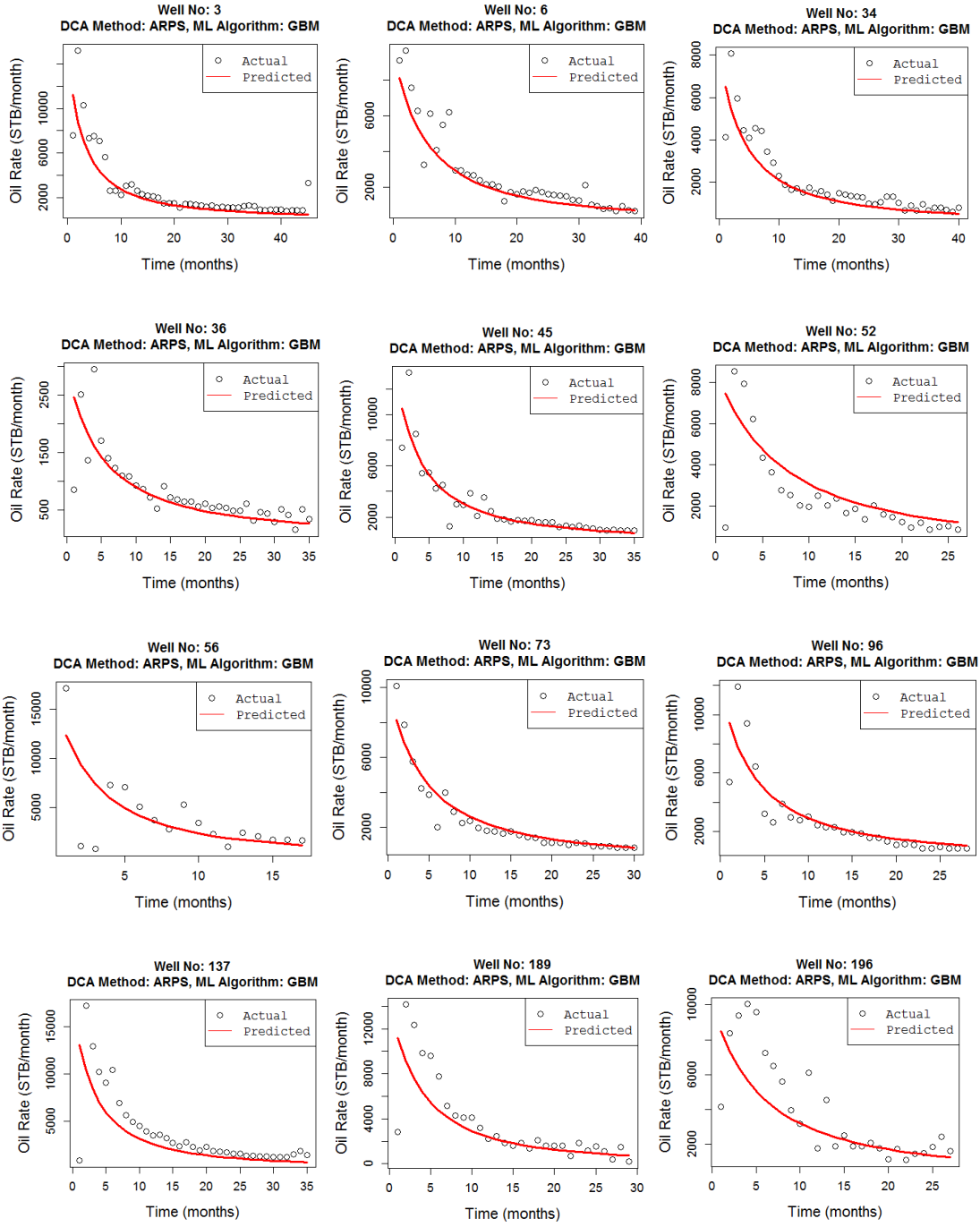


Figure 2.24 Prediction of Arp's decline curves using GBM

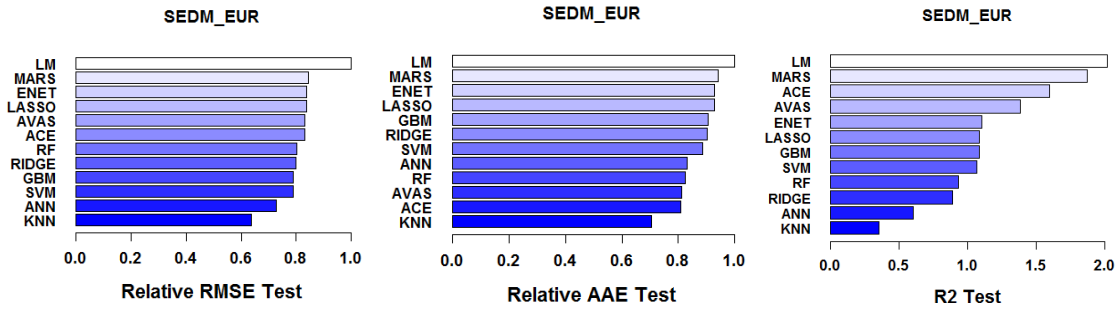


Figure 2.25 Error metric comparison for different machine learning algorithms taken into consideration for SEDM model

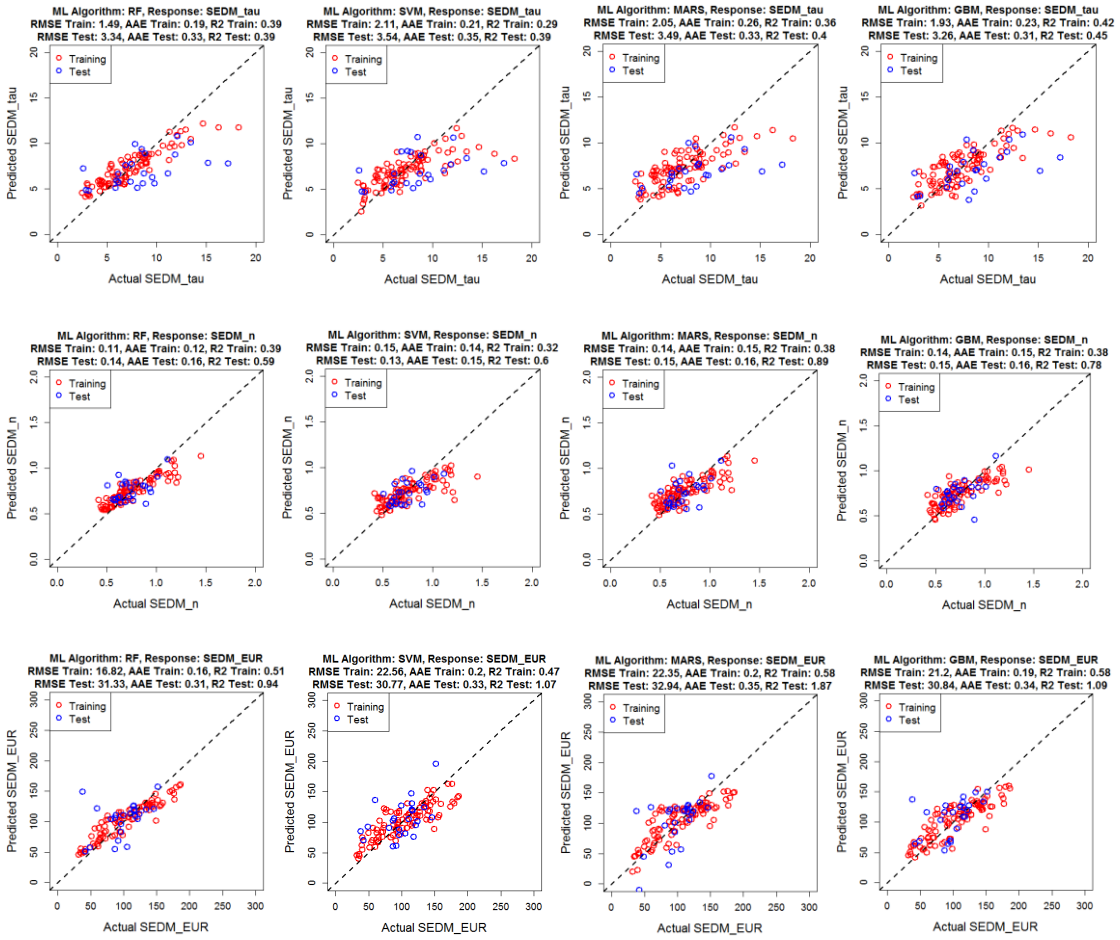


Figure 2.26 Scatterplots showing predicted vs actual values of SEDM decline model parameters and EUR

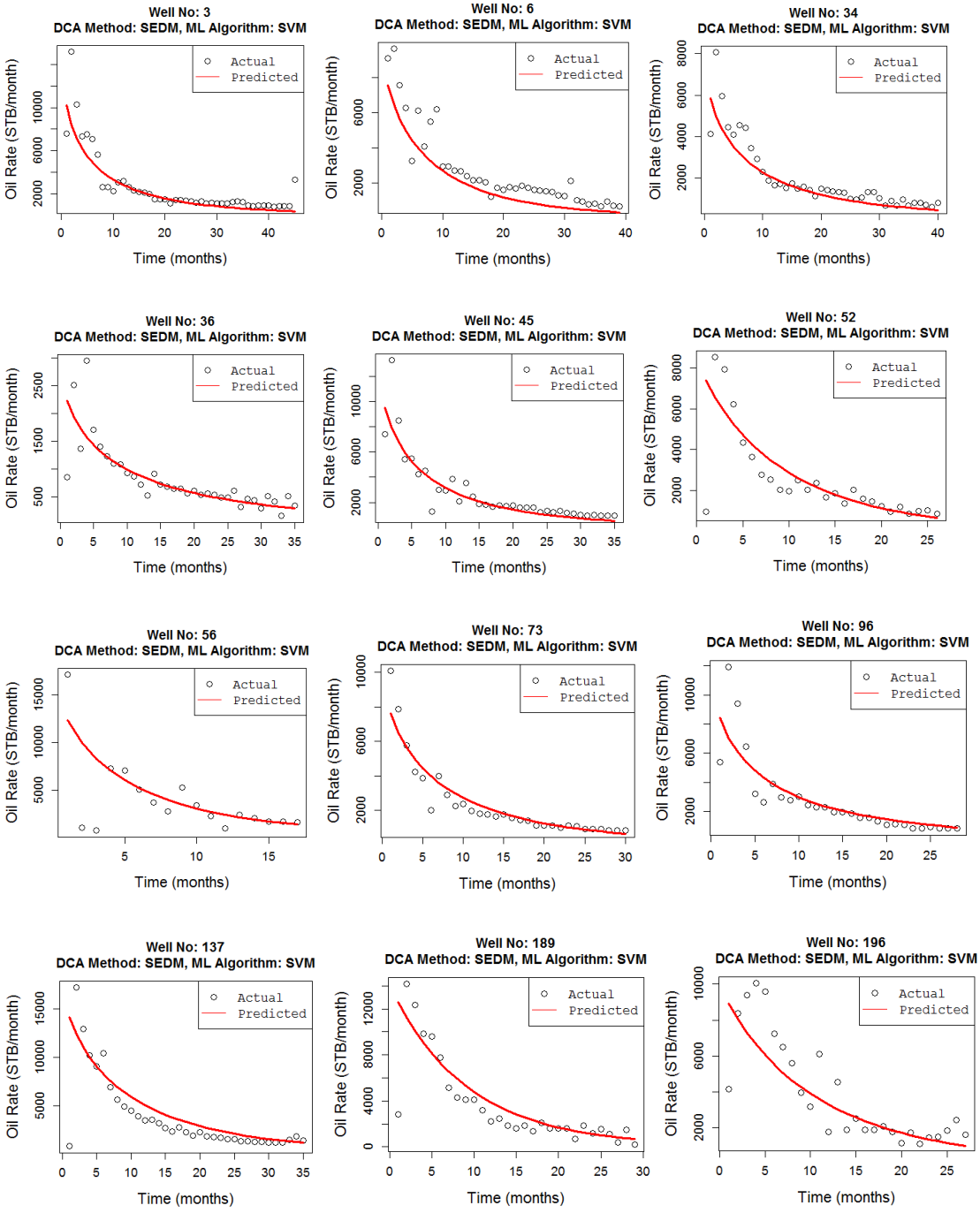


Figure 2.27 Prediction of SEDM decline curves using SVM

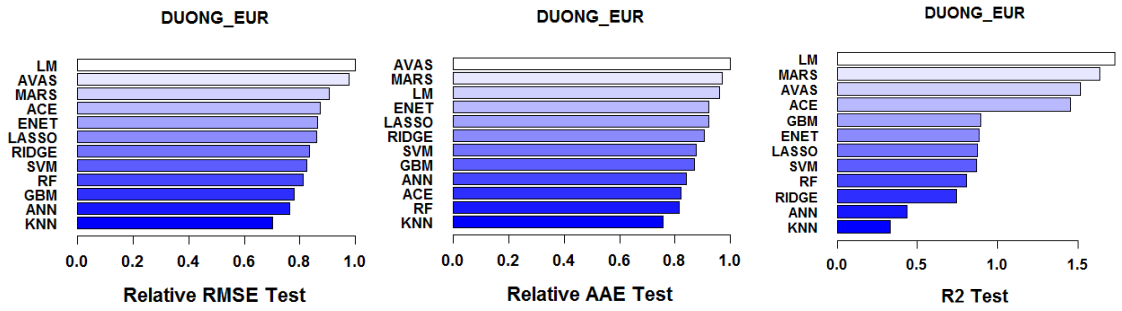


Figure 2.28 Error metric comparison for different machine learning algorithms taken into consideration for Duong's model

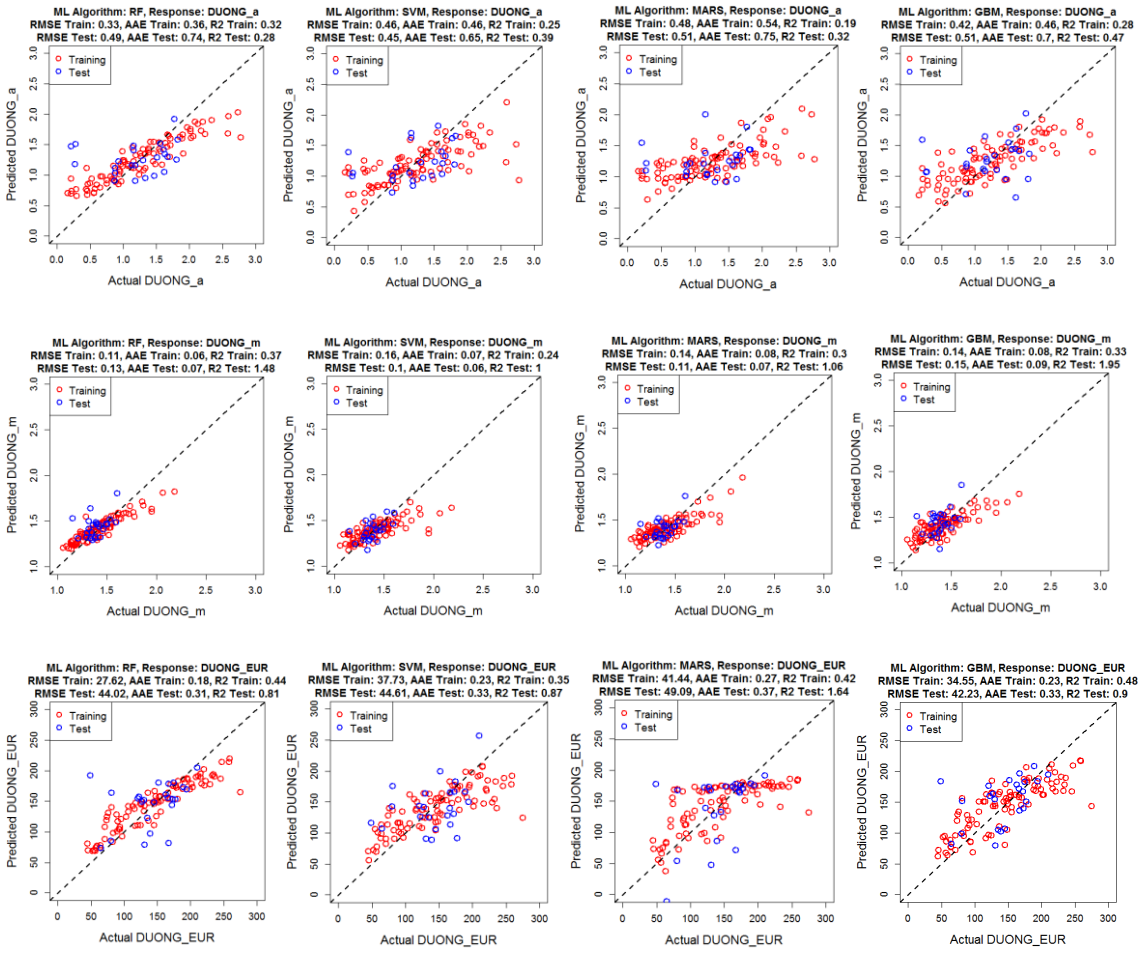


Figure 2.29 Scatterplots showing predicted vs actual values of Duong's decline model parameters and EUR

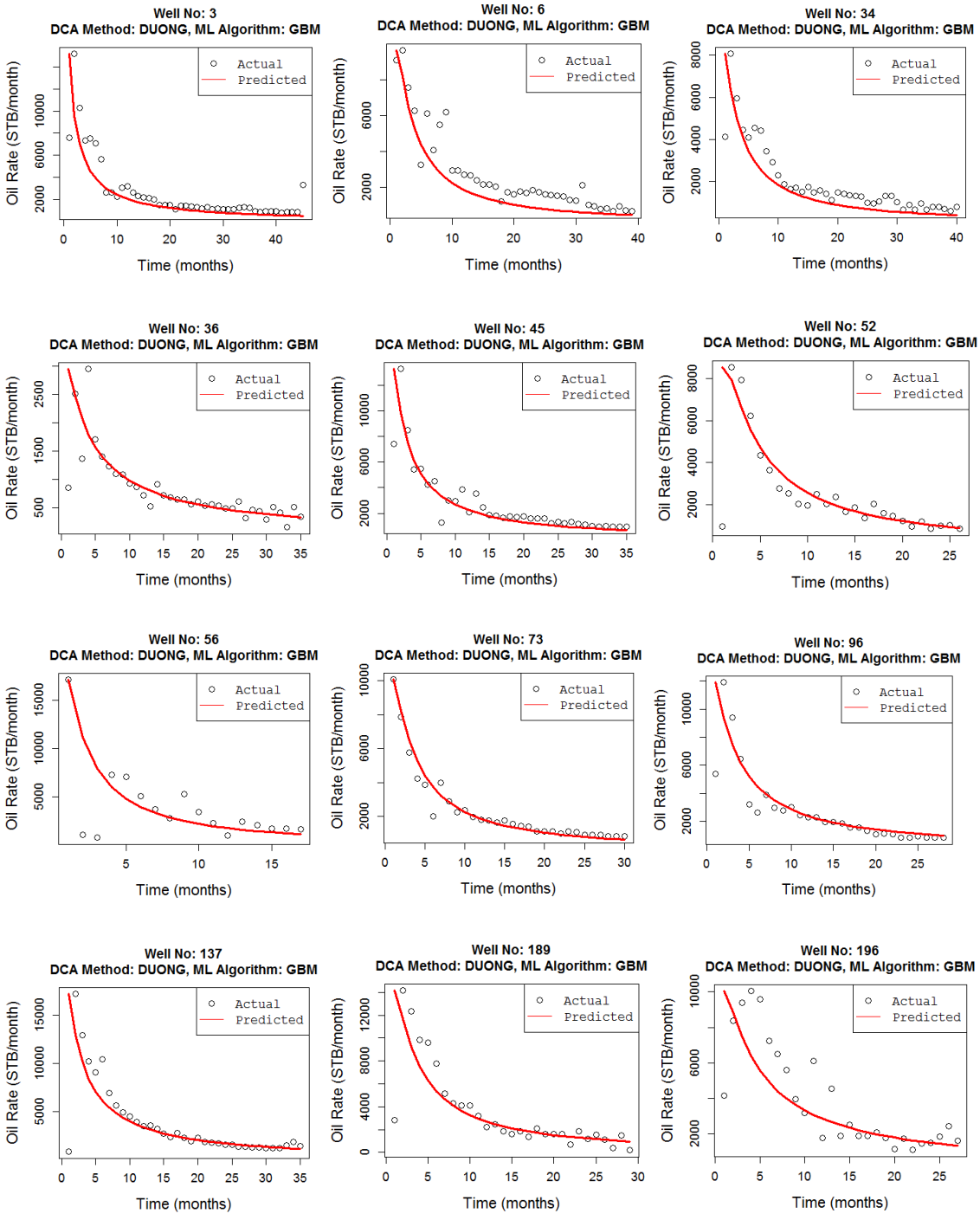


Figure 2.30 Prediction of Duong's decline curves using GBM

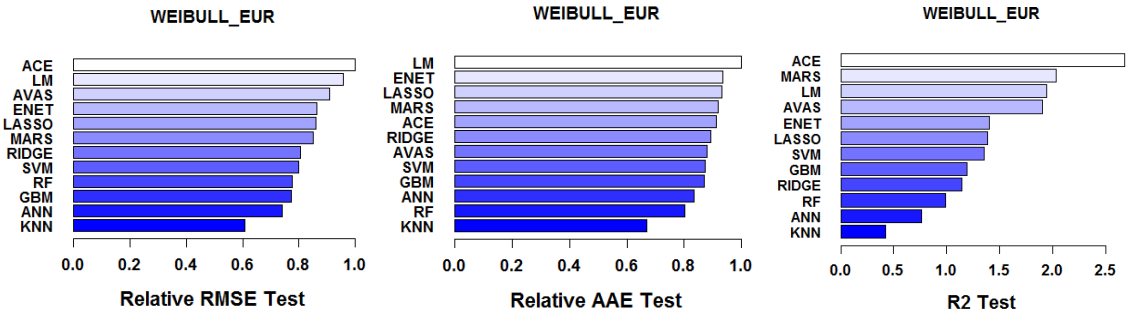


Figure 2.31 Error metric comparison for different machine learning algorithms taken into consideration for Weibull model

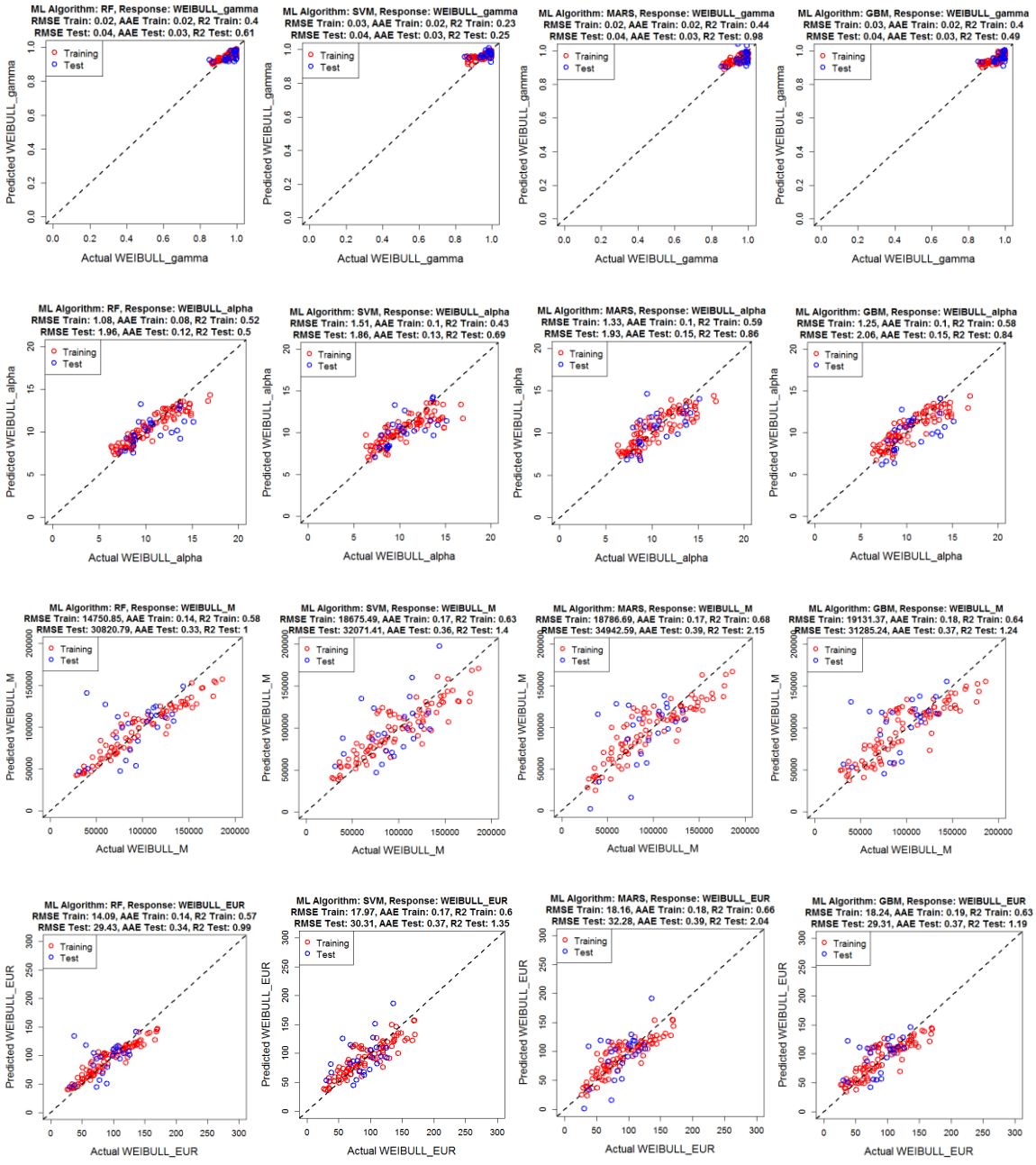


Figure 2.32 Scatterplots showing predicted vs actual values of Weibull's decline model parameters and EUR

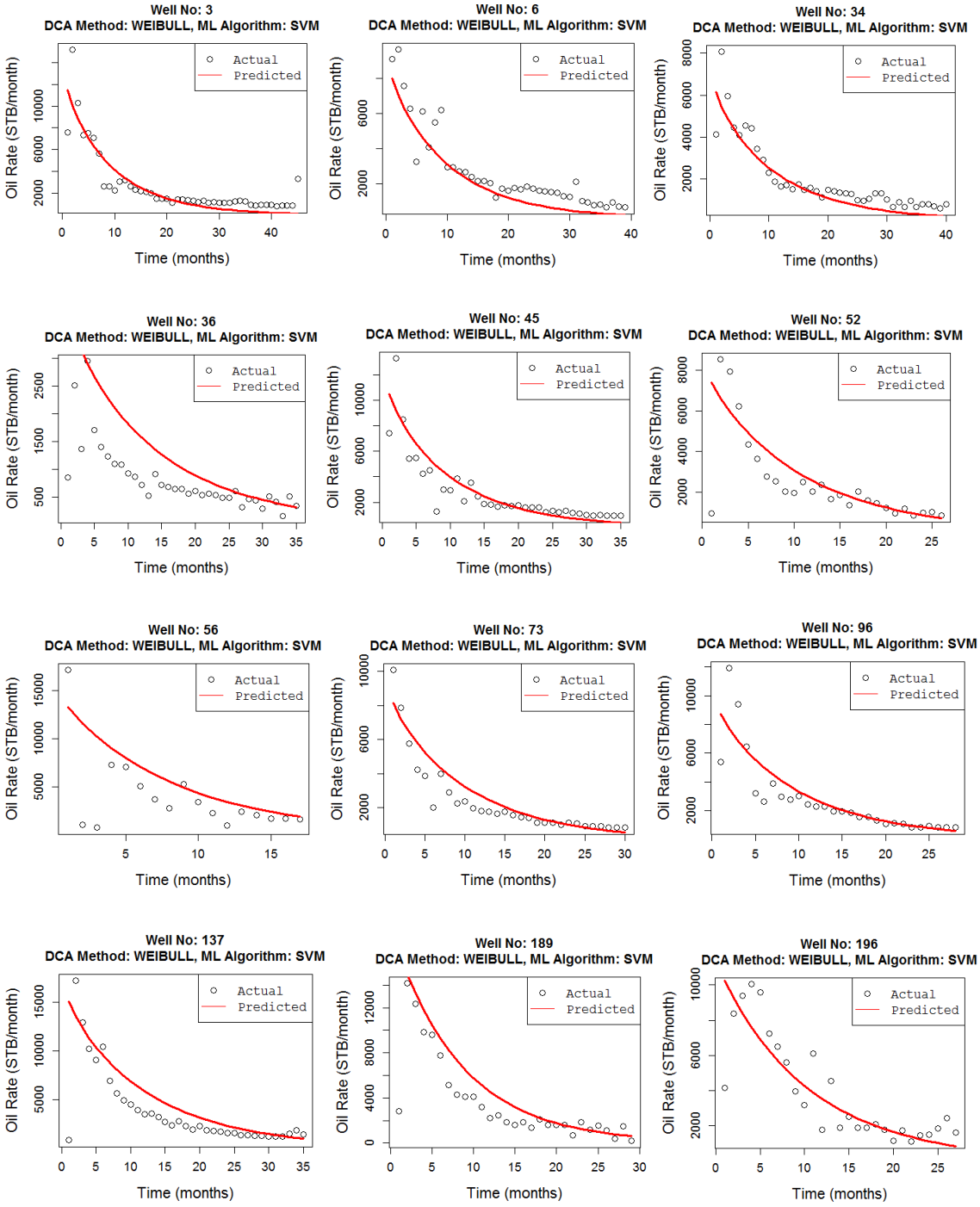


Figure 2.33 Prediction of Weibull's decline curves using SVM

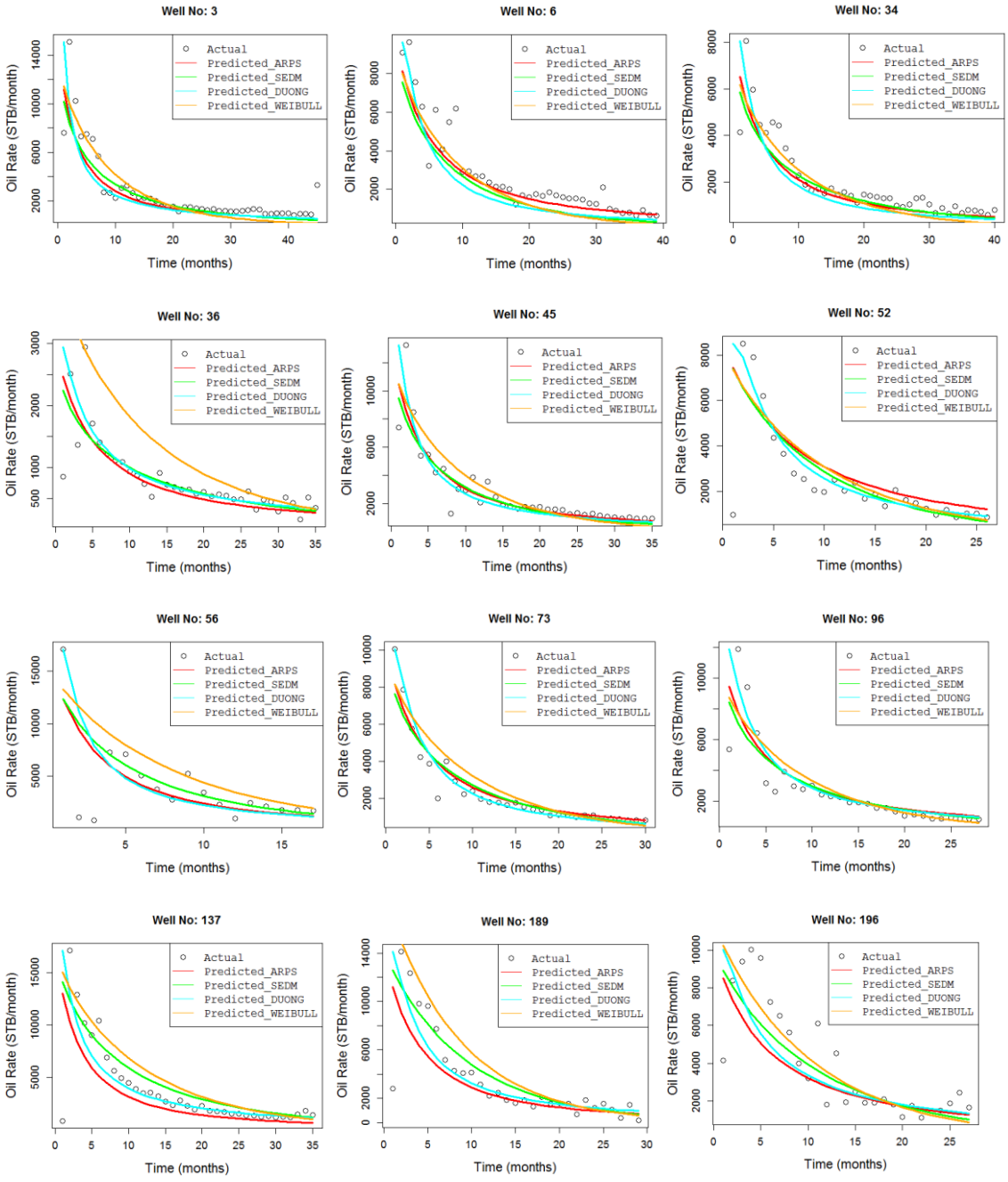


Figure 2.34 Comparison of predictions made by ARP'S - GBM, SEDM - SVM, DUONG – GBM and WEIBULL - SVM

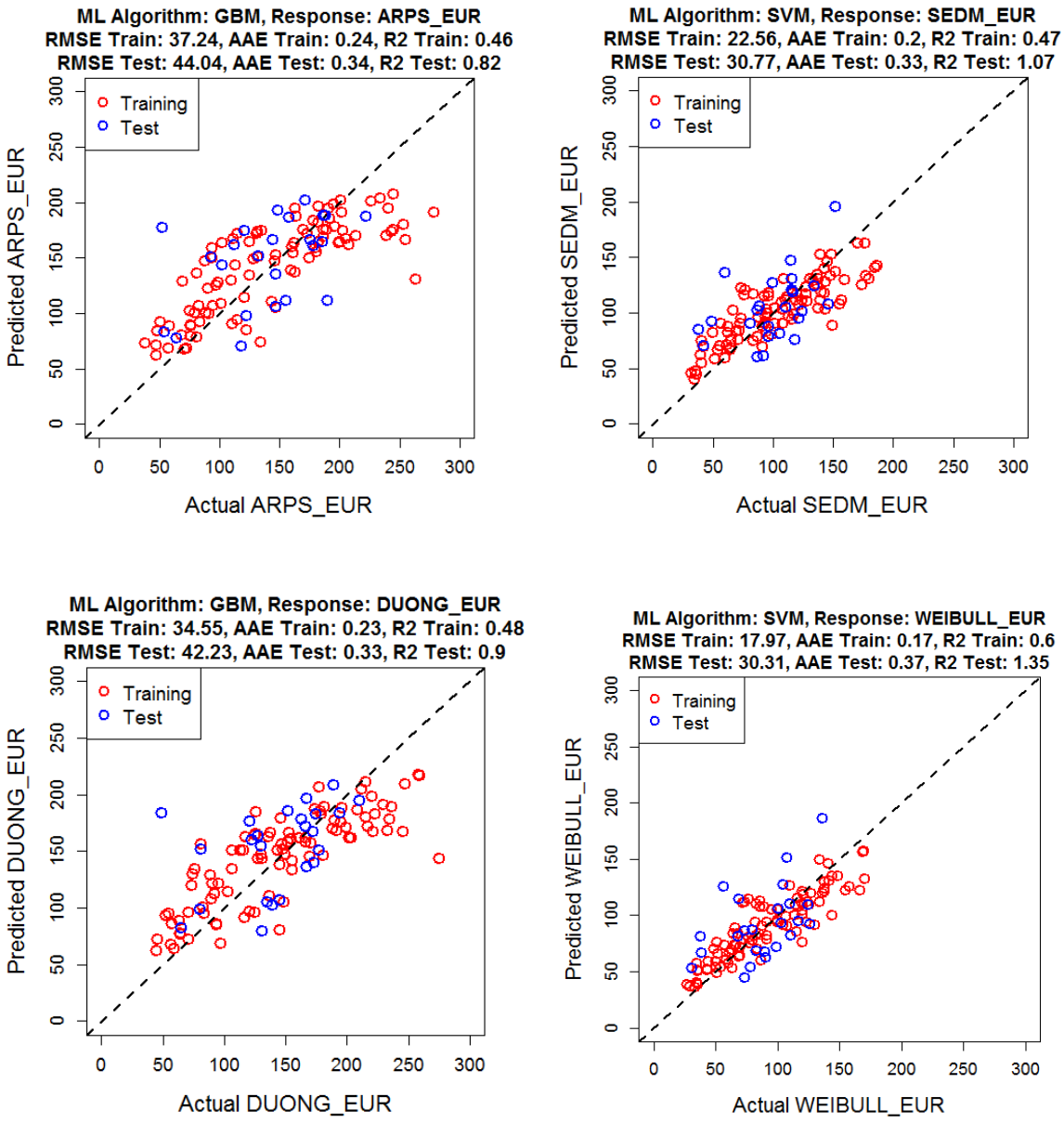


Figure 2.35 EUR prediction comparison among best candidates for each decline model

Fig. 2.36 shows the distribution of variable rankings based on RMSE errors. As described previously, variable rank is calculated based on relative change in test data error metric if the predictor variable is removed from machine learning model. **Fig. 2.36** shows variable ranking based on change in RMSE metric. A predictor variable can have a different rank in different decline model – machine learning combination. This relative influence/ranking plots are generated considering 4 decline models (Arp's, SEDM, Duong and Weibull) and 10 machine learning algorithms (RF, SVM, GBM, MARS, ANN, KNN, LM, RIDGE, LASSO and ENET) not including ACE and AVAS due to instability issues. Therefore, each predictor variable has 40 possible rank values across all these combinations. **Fig. 2.37** shows frequency histograms of predictor variable rank distributions and **Fig. 2.38** shows the Average Rank versus Rank Variance corresponding to each of the predictor variable. A variable with rank close to unity and with low rank variance is considered to be more important that others. As can be observed from these figures, initial flow rate, qi, is ranked at the top in all cases.

Figs. 2.39 to **2.41** show similar analysis as describe above based AAE metric and **Figs. 2.42** to **2.44** show the analysis based on R2 metric. **Figs. 2.45** to **2.47** show the analysis based on Median-Sigma ratio based metric. As can be observed here different error metric can provide different variable ranking analysis plots. However, it may be observed here that initial flow rate is always highly ranked among all cases. Also, since TVD has been obseved to be a critical predictor during exploratory analysis conducted previously and since R2 metric gives TVD high importance after initial flow rate, it may

be logical here to assume R2 base variable importance plots to be more accurate compared to other metrics.

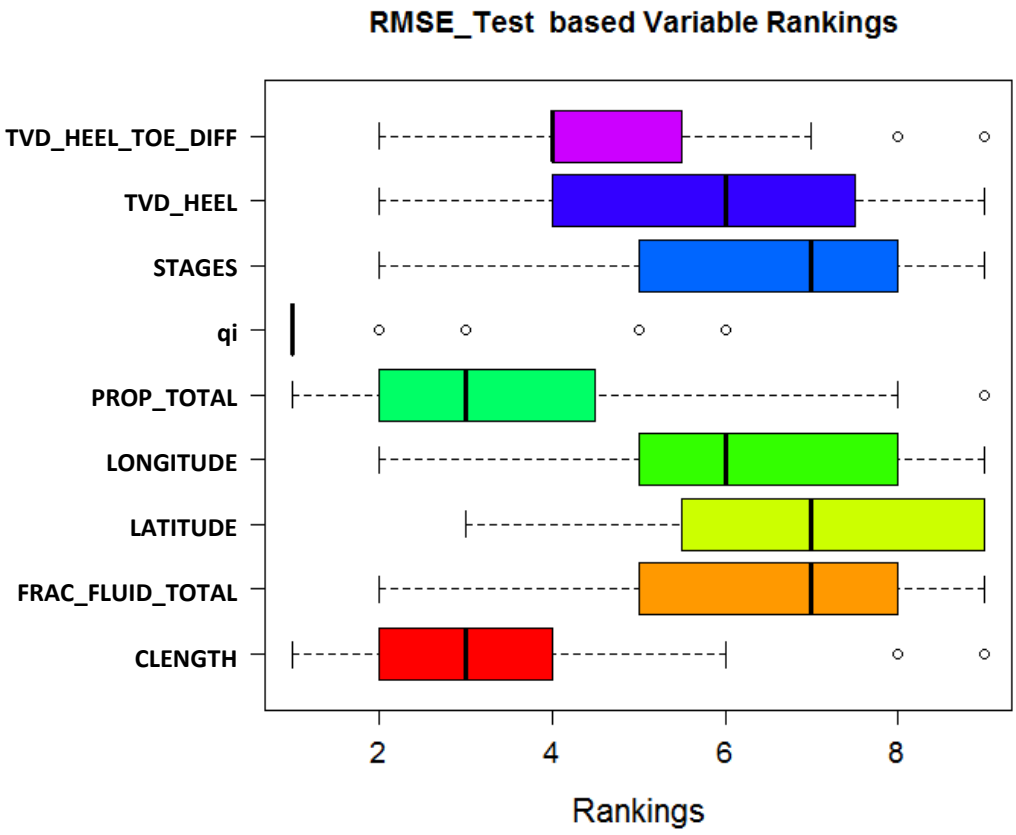


Figure 2.36 RMSE based variable ranking distribution

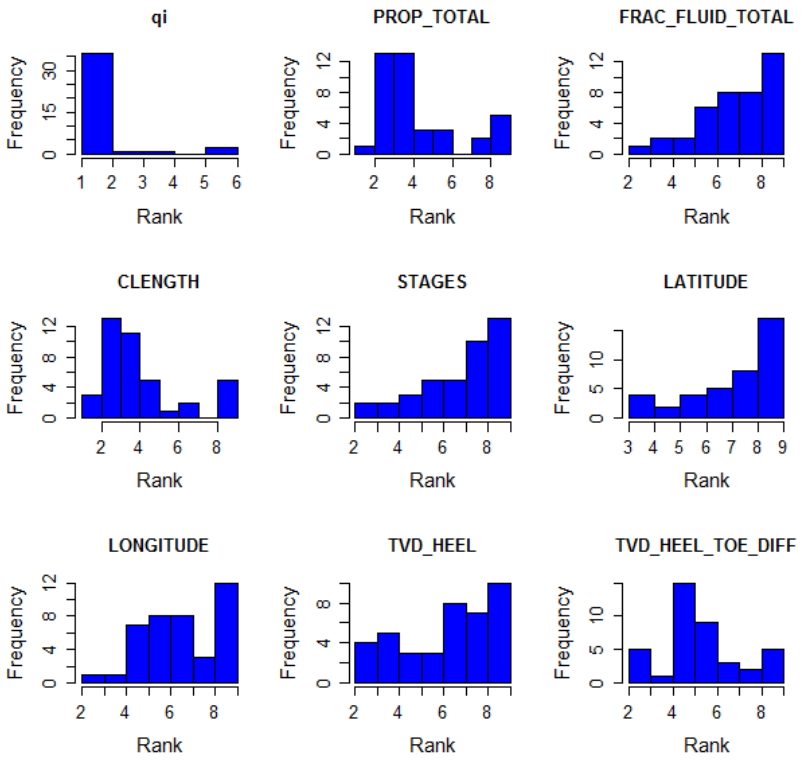


Figure 2.37 RMSE based variable ranking frequency distribution

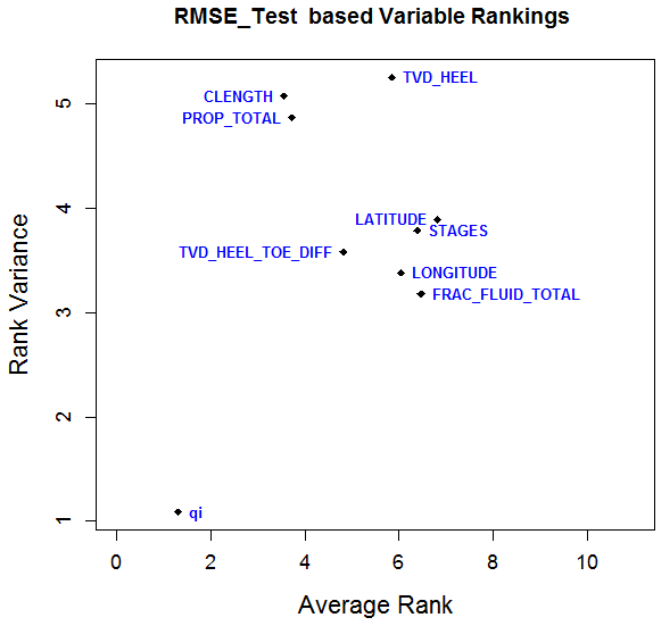


Figure 2.38 RMSE based variable average rank vs rank variance

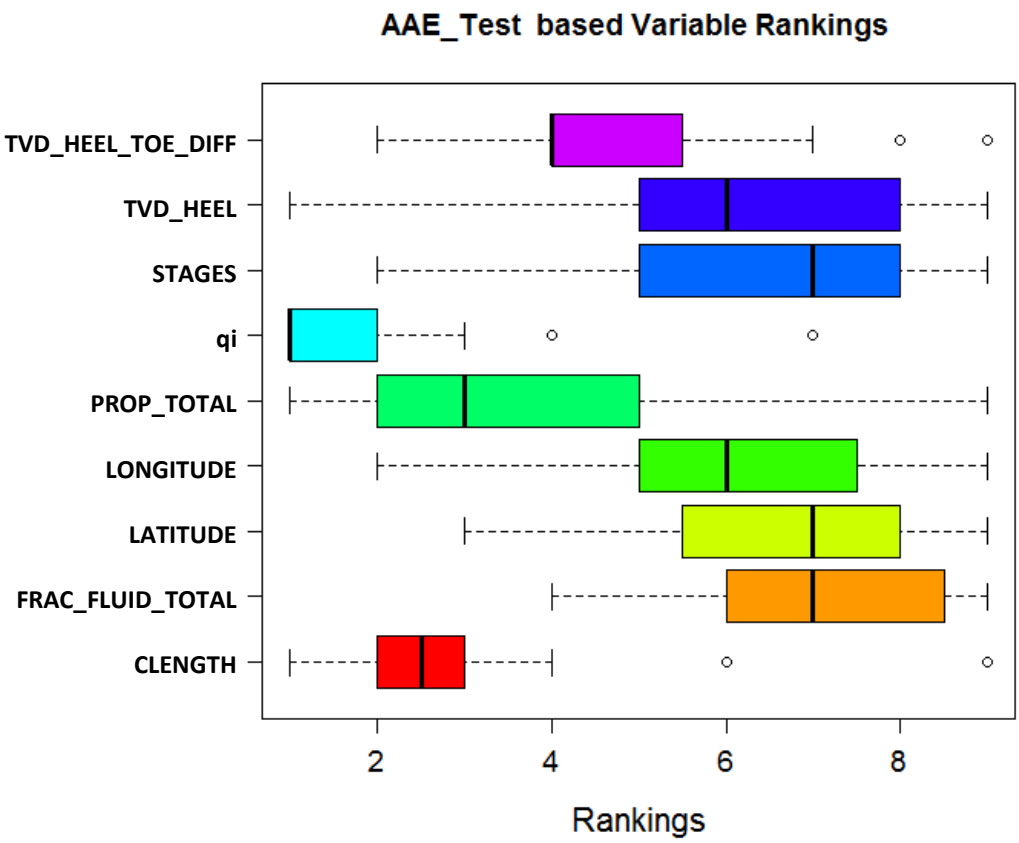


Figure 2.39 AAE based Variable Ranking distribution

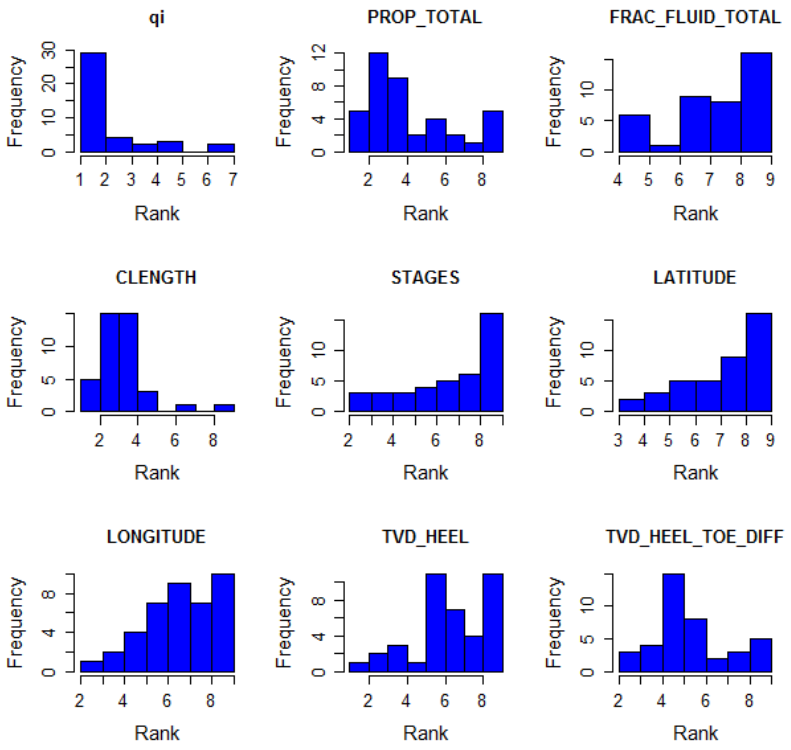


Figure 2.40 AAE based variable ranking frequency distribution

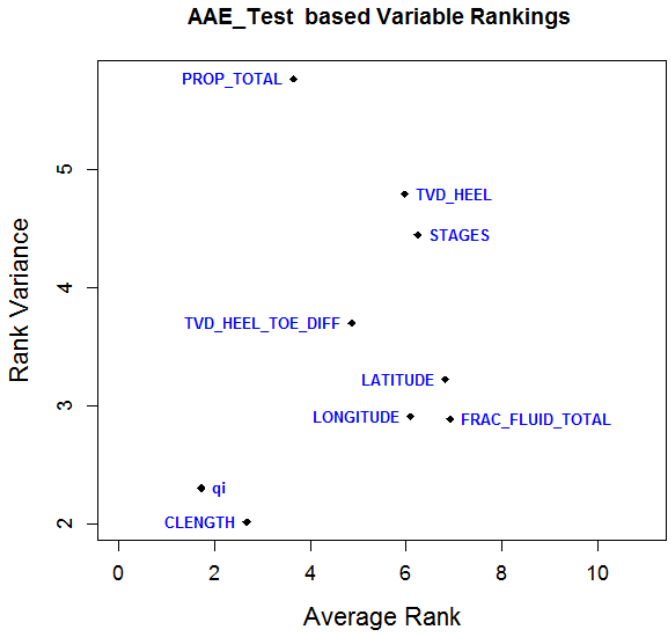


Figure 2.41 AAE based variable average rank vs rank variance

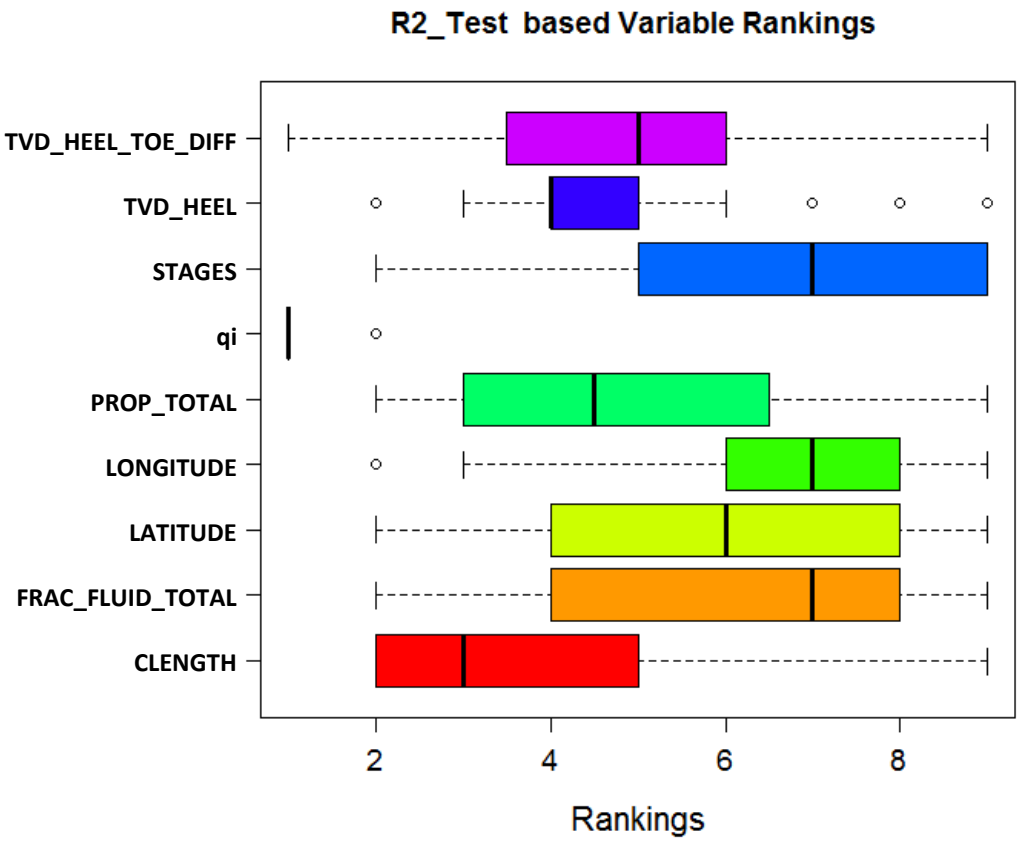


Figure 2.42 R^2 based variable ranking distribution

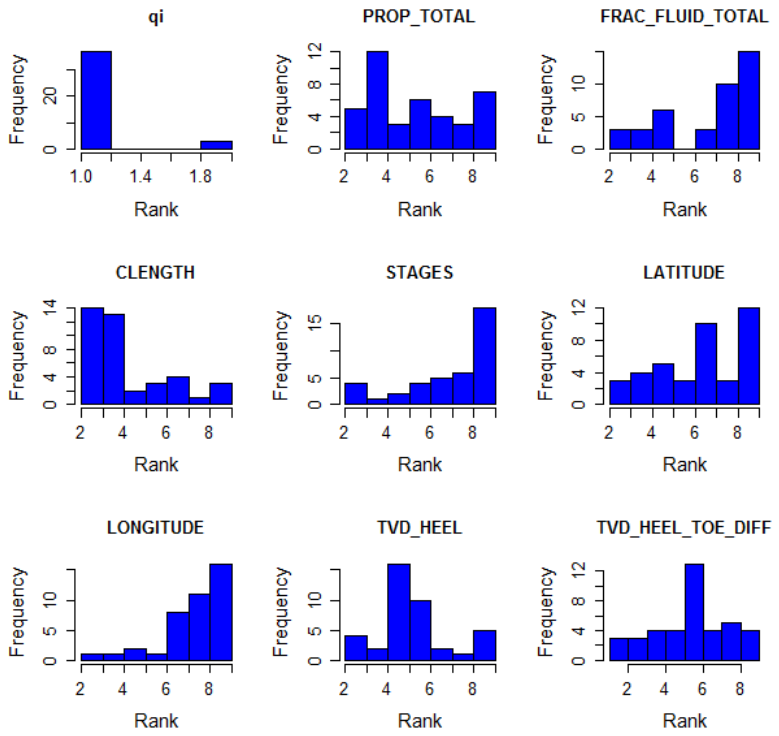


Figure 2.43 R^2 based variable ranking frequency distribution

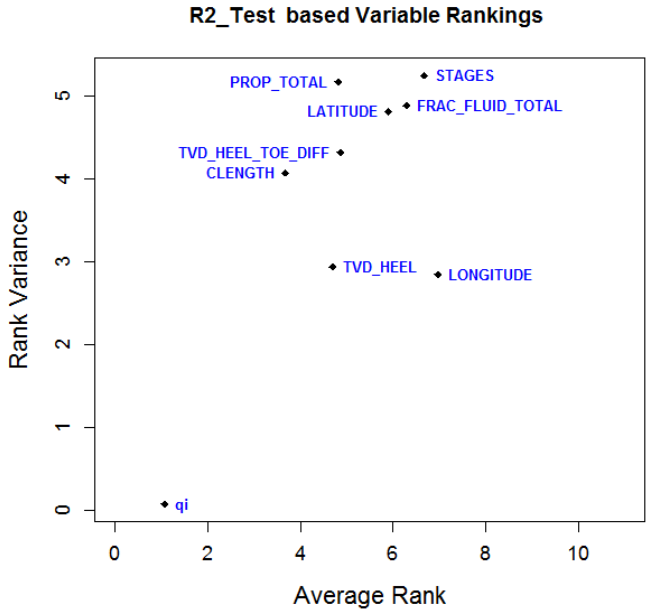


Figure 2.44 R^2 based variable average rank vs rank variance

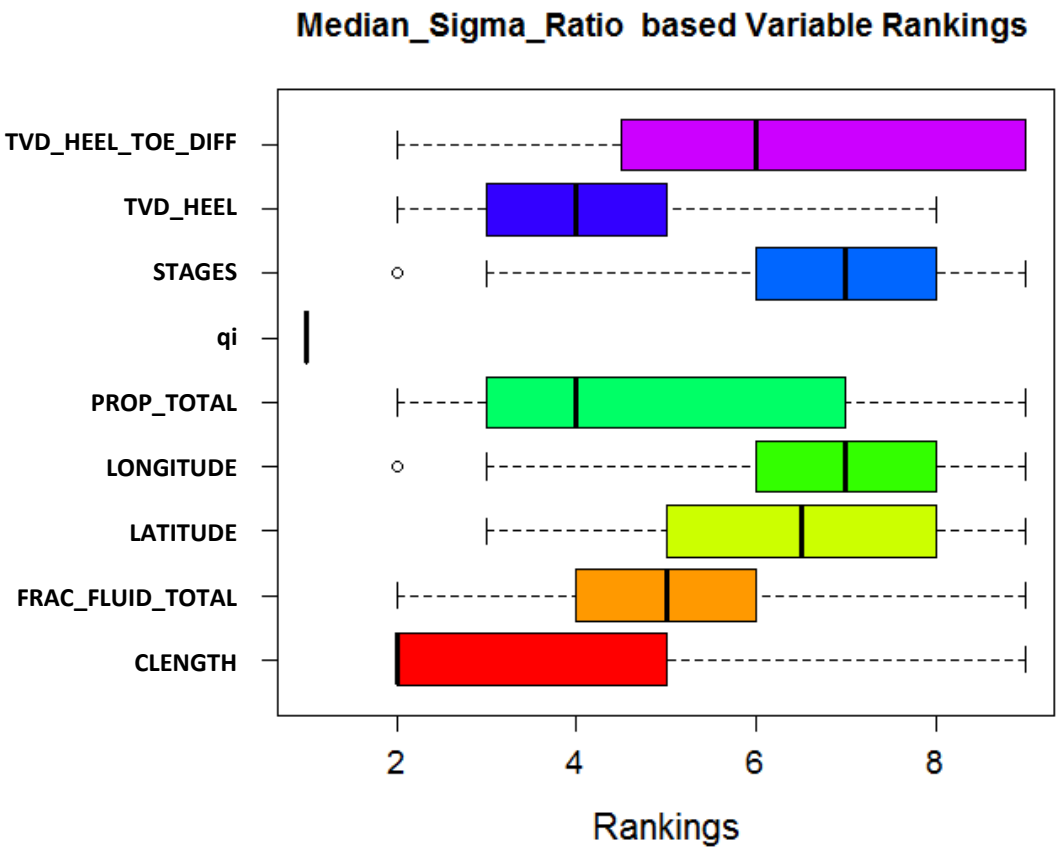


Figure 2.45 Median-Sigma ratio based variable ranking distribution

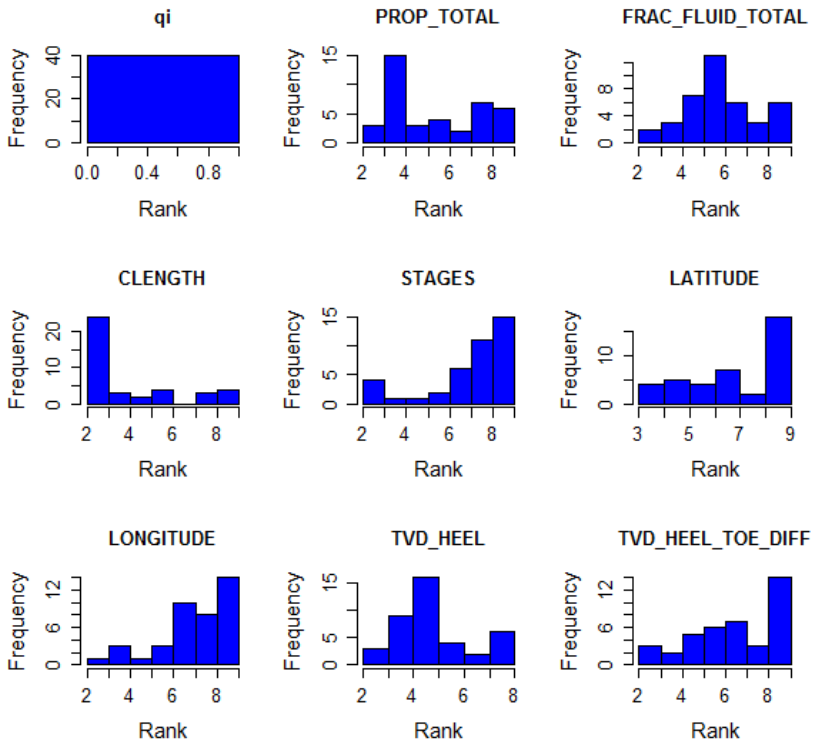


Figure 2.46 Median-Sigma ratio based variable ranking frequency distribution

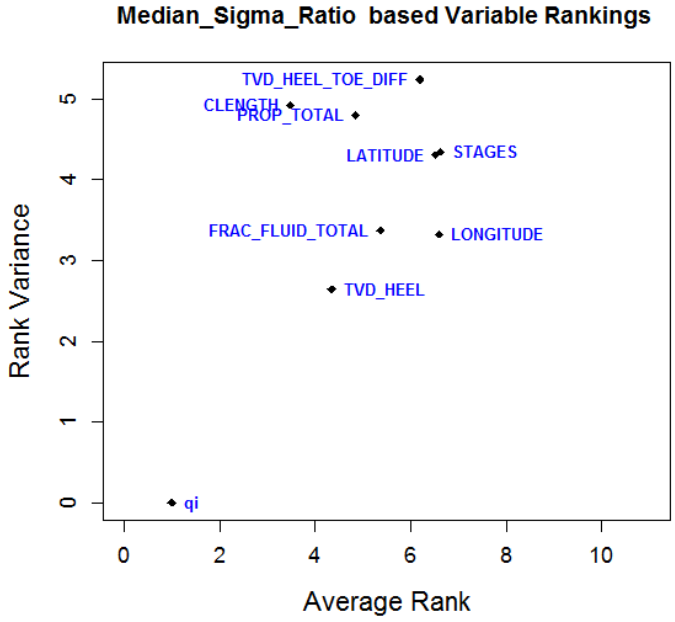


Figure 2.47 Median-Sigma ratio based variable average rank vs rank variance

2.4 Summary

1. Rate decline model parameters for Arps, SEDM, Duong and Weibull decline models can be linked to well completion and location variables using Machine Learning.
2. Rate decline curves are predicted for each of the four decline models and compared with observed data of test wells.
3. Most suitable Machine Learning algorithms for predicting decline curve parameters for each of decline models have been identified in this study.
4. SEDM with SVM is found to be the most suitable combination to predict EUR.
5. Relative Variable Importance study shows that initial flow rate to be most influential predictor followed by total vertical depth.

CHAPTER III

HYDRAULIC FRACTURE DESIGN AND OPTIMIZATION IN UNCONVENTIONAL SINGLE PHASE GAS RESERVOIR USING GENETIC ALGORITHM*

3.1 Introduction and Literature Review

In USA, shale oil and gas production has been on the rise particularly during the last decade. However, due to very low permeability in these reservoirs, hydraulic fracturing becomes an essential requirement for economical production. These hydraulic fractures are created after pumping large amount of fracturing fluid and proppant to support fractures thus created. Once created, this process increases conductivity and surface area for fluid flow in the reservoir which increases the well production. Well production can be increased by increasing the number of hydraulic fractures. However, it may not be economical to increase investments in this process beyond a certain point. This study focuses on getting close to this 'most economical' point by applying a class of evolutionary algorithm know as genetic algorithm in a synthetic unconventional reservoir. This chapter will use this reservoir case to optimize various parameters associated with hydraulic fracturing design.

* Parts of the text and data reported in this chapter is reprinted with permission from Yang, C., Vyas, A., Datta-Gupta, A., Ley, S.B. and Biswas, P., 2017. Rapid multistage hydraulic fracture design and optimization in unconventional reservoirs using a novel Fast Marching Method. Journal of Petroleum Science and Engineering. Copyright 2017 Elsevier

Holditch (1992) reported that there is plenty of oil and gas reserves as long as it is possible to exploit them economically. It was also reported that horizontal wells with multiple hydraulic fractures using waterfrac technology is key for hydrocarbon production from shales. It was also reported that going forward the biggest technological benefits will be found in cost cutting improvements.

Saldungaray et al. (2013) emphasized the role of fracture conductivity on well productivity. Fracture conductivity is dependent on the type of proppant/fracturing fluid used and type of technique used for fracturing job. It was also reported that the number of hydraulic fractures and spacing between them are dependent on rock fabric and formation permeability. The three parameters – the rock fabric, natural fracture distribution and the reservoir permeability – are noted as most important while optimizing the number of hydraulic fractures used in a well.

Rankin et al. (2010) noted that since transverse fractures in horizontal wells provide small intersection area, multiple stages with higher conductivity proppants are needed to improve the flow capacity of the connection between fractures and wellbore. Superior productivity is reported using more than 10 hydraulic fractures in the Bakken study area reported.

Morales et al. (2010) presented a modified genetic algorithm to optimize well placement in a reservoir. It was identified that in a complex heterogeneous reservoir, optimum location of a well based on intuition is difficult to achieve.

Kennedy et al. (2012) presented well placement optimization process and identified required combination of petrophysical, geochemical, and geomechanical

properties of a reservoir. It was reported that resource development simply based on uniformly spaced hydraulic fractures may not be ideal for a heterogeneous reservoir. It was reported that a naturally fractured reservoir can be drained better if a complex network of fractures can be created during hydraulic fracturing process. However, in order for optimization of well placement and hydraulic fracture design, a good amount of knowledge about reservoir is needed. This study reports that tools such as include electrical resistivity imaging LWD logs can be utilized in order to maximize the knowledge about a reservoir. Also, techniques such as micro seismic monitoring can be used to determine the details of hydraulic fractures created after fracking. A high definition resistivity log can be used to identify natural fractures, induced fractures (from nearby offset wells), faults and bedding planes.

Helgesen et al. (2005) presented a novel resistivity tool for accurate wellbore placement. This tool is reported to have depth of investigation nearly 5 times the conventional multiple propagation resistivity tools.

Biswas and Ley (2015) introduced a novel approach for natural fracture interpretation using log data. This paper makes use of compressional waveforms instead of shear waveforms allowing faster and accurate determination of natural fractures. At least 4 (one in each sector) raw waveforms are used in this method as an input out of which the first one is muted in time domain and filtered in frequency domain. This process is repeated in each sector and RMS energy is calculated. A modified stacking algorithm is used to amplify the finer perturbations in the data and to stabilize the waveforms. Since compressional waveform data is not collected by every cross-dipole/wireline tool, this

paper suggests to make use of first arrival waveforms or “leaky mode waveforms” since these waves have compressional velocities.

Sierra et al. (2013) concluded from their paper that reservoir permeability is the main driver during decision making regarding hydraulic fracture spacing along horizontal well. It was also concluded that fracture complexity is important only in reservoirs having permeability lower than 100 nd. In reservoirs having permeability more than that, optimally placed planar fractures should be sufficient to maximize gas recovery factor. Also, proppant settling effects which are frequently observed in waterfracs, influence the fracture spacing. It was also concluded that in case of stress dependent permeability and/or porosity, smaller fracture spacing should be used. However, if the hydraulic fractures are not properly propped, smaller fracture spacing cannot compensate. It was concluded that knowledge of stress dependency of reservoir permeability and porosity is needed in deciding fracture spacing. Also, type of proppant used can alter the fracture conductivity and therefore put an effect on optimal hydraulic fracture spacing.

Ma et al. (2013) reported their hydraulic fracture placement optimization results. This study uses both derivative free genetic algorithm based optimization and finite difference based optimization of NPV. However, this study is not optimizing the fracture half-length and proppant/fracturing fluid amount. It was found in their study that in a heterogeneous reservoir, fracture spacing in high permeability region is lower than in low permeability region. Also, in case of the finite difference based optimization method, the optimum model showed near uniform spacing in low permeability region of the reservoir.

Yang et al. (2012) reported a hydraulic fracture optimization method using a pseudo 3D hydraulic fracturing model for a multilayered formation. Their approach integrated Linear Elastic Fracture Mechanics (LEFM), Unified Fracture Design (UFD) and 2D PKN model. This paper presented an algorithm that can help in determining what treating pressure and other treatment parameters are needed to achieve optimum placement of a given amount of proppant of specified quality. This method also informs about the layers which act as containment barriers for vertical fracture propagation at a specified treating pressure level.

Pitakbunkate et al. (2011) reported that fracture optimization based on Unified Fracture Design (UFD) results in optimum fracture geometry. It was also reported that fracture height growth depends on inter layer stress differential and not on individual stress values. In low permeability reservoirs, large fracture height is accompanies by larger fracture half lengths. It was also reported that there is a need to study fracture height migration to prevent fracture migrations into water zones.

Warpinski et al. (1998 and 2005) reported how hydraulic fracture growth and geometry can be detected using microseismic data. During hydraulic fracturing treatment, changes in pore pressure affect planes of weakness (natural fractures and bedding planes) adjacent to the hydraulic fracture and allow them to undergo shear slippage. These shear slippages are like small earthquakes (and hence called “microseisms” or micro earthquakes). These microseisms emit elastic wave signals that can be detected by transducers located for analysis.

Maxwell et al. (2002) concluded from Barnett shale studies that real time microseismic images can be utilized to fracture geometry in currently uneconomic regions of Barnett shale in order to make them economic. Fisher et al. (2005) also reported results from Barnett shale. The paper reports that there can be three types of fractures – simple, complex and very complex. In a shale reservoir having a presence of natural fractures, the fracture complex is more likely to be "very complex". A very complex network of fractures allows a fracture fairway to be created with many fractures in multiple orientations resulting in large contact area between well and reservoir. This paper reported various technologies available that can be utilized to gather information regarding fracture parameters such as height, length and azimuth. These technologies include Surface Tiltmapping, Downhole Tiltmapping and Microseismic Mapping. This paper reports that in Barnett shale wells, it is the cumulative fracture-network length (combining both hydraulic fractures and natural fractures) that controls the reservoir connectivity and not the conventional fracture half lengths. The paper reports ways to estimate fracture growth by history matching recorded fracture data.

Cipolla et al. (2009) used dual permeability based reservoir model to simulate creation of Stimulated Reservoir Volume (SRV). The paper concludes that with the availability of reservoir geologic data such as core data and microseismic data can be used to history match the simulated data. Gas recovery can be increased by increasing the complexity of fracture network. The paper also reports that in low Young's modulus formations, effect of stress dependent network fracture conductivity becomes dominant resulting in lower recovery. This effect is usually observed after 1-2 years of production.

Savitski et al. (2013) reported from their studies that even though the aperture of a hydraulic fracture is greater than natural fractures, the total area of activated (pressurized) natural fractures can be significant which makes them relevant to production. Another conclusion made by this study is that DFN connectivity does not cause a characteristic response that would allow one to determine DFN connectivity from stimulation data. It was also concluded that stress perturbation is not sufficient to stimulated non-conductive natural fractures and that initial natural fracture conductivity is critically important. It was also concluded from their study that lower injection rate will results in larger stimulated reservoir volume in the presence of conductive natural fracture, though it will also result in hydraulic fractures of lower width that may be susceptible to premature screen-out.

Riahi and Damjanac (2013) conducted numerical simulations to study interaction between hydraulic fractures and natural fractures. This study concluded that for a given injected volume, lower injection rates result in greater proportion of DFN being affected during hydraulic fracturing propagation. It was also concluded that DFN properties such as density, length distribution and fracture orientation are critical to the overall response of the formation during hydraulic fracturing.

Dershowitz et al. (2000) integrated DFN methods with conventional dual porosity reservoir simulators. It was reported that permeability of the natural fracture system depends on the fracture intensity, the connectivity of the natural fracture system and the distribution of the natural fracture transmissivities. This study made use of the tensor approach of Oda (1985). Using this approach, equivalent permeability of each grid block containing natural fractures can be generated and then further simulations can be carried

out. However, Oda method is suitable only in well-connected natural fractures only since it does not take fracture connectivity into account.

Various authors have reported their methods for long term reservoir performance forecasting. Arps (1945), Fetkovich (1980) and Valko and Lee (2010) proposed decline curve based production predictions. Ilk et al. (2010) and Song and Ehlig-Economides (2011) proposed their methods for reserve estimation and production forecast using pressure/rate transient analysis. These analytic methods are fast but not as accurate as numerical simulator available commercially due to their inadequacies to incorporate complex heterogeneities in field. Fan et al. (2010) used a numerical simulator to predict shale gas production in Haynesville shale. Shale gas log data is used to gather information about reservoir porosity, permeability, TOC, saturations, etc. History matching the early production data is then done to calibrate the reservoir properties. Microseismic data can give idea of fractures created during hydraulic fracturing process. It was reported in this paper that difference in stress contrast can lead to different complexities of fracture network created during hydraulic fracturing treatment. This study shows two types of complexities due to difference in stress anisotropies. Other factors affecting fracture network include rock fabric, preexisting natural fractures and layering. Once a model is calibrated using available production data, microseismic data, core data, etc., a reasonable forecast can be made for future.

The use of commercial reservoir simulators can give a very accurate production forecast but this method is costly and time consuming process. Lee (1982) proposed the concept of radius of investigation in homogeneous reservoirs. It is defined as the

propagation distance of a “peak pressure” disturbance for an impulse source or sink (Lee 1982). Datta-Gupta et al. (2011) extended this concept to heterogeneous reservoirs with arbitrary well conditions and the diffusive equation then turns out to be the Eikonal equation which can be solved very efficiently by a class of front tracking methods known as Fast Marching Methods (FMM) presented earlier by Sethian (1996 and 1999).

Sehbi et al. (2011) used the concept of drainage volume for optimizing hydraulic fracture stages in Tight Gas Reservoirs. Their study used a high frequency asymptotic solution of the diffusivity equation to generalize the concept of radius of drainage (Lee, 1982) to horizontal wells. In this study done in cotton valley formation well, ten hydraulic fractures with 500 ft of half-length came out to be most optimum. Increasing number of stages beyond that would yield diminishing returns. Besides application in optimization problem, drainage volume calculations gave an additional advantage of flow visualization with no additional simulations.

Xie et al. (2015a) revisited FMM and proposed a geometric pressure solution based on depth of investigation to estimate transient pressure behavior in unconventional wells with multistage hydraulic fractures. Well diagnostic plot was generated from pressure depletion behavior that could be used to identify various flow regimes. The advantage of using this technique is that transient pressure response for a multimillion grid cell based reservoir model can be obtained within in seconds. Xie et al. (2015b) integrated shale gas production data and microseismic data using FMM to obtain reservoir and hydraulic fracture properties. Fracture parameters such as fracture half lengths and fracture permeability and reservoir parameters such as matrix permeability and SRV permeability

were determined using a history matching process based on Genetic Algorithm (GA). Since FMM combined with geometric approximation is computationally very efficient compared to commercially available forward simulators, this history matching problem could be completed very fast.

Zhang et al. (2013) extended the concept of FMM based reservoir simulation to complex flow geometry and anisotropic properties. This study derived the FMM formulation in corner point grids. Zhang et al. (2014 and 2016) derived a new formulation of the diffusivity equation using diffusive time of flight as a spatial variable transforming three dimensional simulation problem to a one dimensional one. The diffusive time of flight (DTOF) embeds the information regarding reservoir heterogeneity. A one dimensional problem is then solved using finite difference method rapidly.

Fujita et al. (2016) extended the DTOF formulation to triple-continuum modeling for modeling shale gas reservoirs. Physical mechanisms like Knudsen diffusion and slippage effects, adsorption/diffusion in nanopore surfaces, rock compaction in fractures due to geomechanical effects and gas diffusion due to Kerogen content we included in the FMM based unconventional shale gas simulator.

3.2 Methodology

3.2.1 Fast Marching Method

This study uses a dual porosity unconventional shale gas model for optimizing hydraulic fracture design. The forward model to calculate gas rate production is based on a Fast Marching Method based reservoir simulator (Zhang et. al, 2014 and 2016). A short

description of this method with various equations is provided in this part of dissertation. However, a more detailed explanation can be found in the reference provided in this section.

Description of FMM method starts with the concept of radius of investigation proposed by Lee (1982). Radius of investigation can be defined radius of investigation in homogeneous reservoirs as the propagation distance of a "peak pressure" disturbance for an impulse source or sink (Lee 1982). Datta-Gupta et al. (2011) extended this concept to heterogeneous unconventional reservoirs with horizontal wells with multistage hydraulic fracturing. Propagation equation of peak pressure front can be derived by using asymptotic ray theory widely used in electromagnetic and seismic wave propagation (Virieux et. al, 1994). Vasco et al. (2000), Kulkarni et al. (2000) and Datta-Gupta and King (2007) used a high frequency asymptotic solution of the diffusivity equation to derive Eikonal equation (**Eq. 3.2)** for propagating pressure front for impulse source. The general diffusivity equation is given by:

$$\nabla.\left(\frac{k}{\mu}\vec{\nabla}\,p\right) = \phi c_t \frac{\partial p}{\partial t} \tag{3.1}$$

The Eikonal equation is given by:

$$\sqrt{\alpha}|\nabla\tau(\vec{x})| = 1 \tag{3.2}$$

where,

τ = diffusive time of flight (DTOF) or the propagation time of the pressure front

α = diffusivity $= k/(\varphi\mu c_t)$

k = permeability

ϕ = porosity

μ = fluid viscosity

c_t = total compressibility

The diffusive time of flight, DTOF, has a unit of square root of time and shows that pressure front propagates in the reservoir with a velocity given by the square root of diffusivity (Datta-Gupta et. al, 2011). It is dependent on reservoir properties but independent of flow rate (Datta-Gupta et. al, 2011). **Eq. 3.2** can be solved by a class of front tracking algorithm known as Fast Marching Method or FMM (Sethian, 1996 and 1999; Zhang et al., 2013, Xie et al., 2015a, 2015b). Using FMM, diffusive time of flight can be calculated for each grid block of a reservoir model. In a homogeneous reservoir, the contours of τ are related to the propagation time t of the pressure front through the following equation (Vasco et al., 2000; Kim et al., 2009):

$$\tau = \sqrt{\beta t} \tag{3.3}$$

where, β is 2, 4, and 6 for 1D linear, 2D radial, and 3D spherical flow patterns respectively. Due to irregular flow pattern, above values of β cannot be applied in heterogeneous reservoirs. However, diffusive time of flight can still help in visualizing pressure front in heterogeneous reservoirs.

The next step is to calculate well production rates based using diffusive time of flight. Once diffusive time of flight values for each grid block in reservoir model is calculated, different diffusive time of flight contours can be generated. The drainage pore

volume, V_p, inside a contour can be calculated by approximating it with the total drainage volume at cut-off. Therefore, FMM solver can generate the drainage pore volume as a function of the diffusive time of flight, $V_p(\tau)$. Zhang et al (2014 and 2016) derived a new formulation of the diffusivity equation using τ as a spatial variable. Instead of writing equation in physical coordinates, this paper presented a new equation in terms of diffusive time of flight (Zhang et al, 2014 and 2016):

$$\frac{1}{w(\tau)}\frac{\partial}{\partial\tau}\left(w(\tau)\frac{\partial p}{\partial\tau}\right)=\frac{\partial p}{\partial t} \tag{3.4}$$

where,

$$w(\tau)=\frac{dV_p(\tau)}{d\tau} \tag{3.5}$$

$w(\tau)$ gives the propagating speed of drainage surface.

Zhang et al (2016) showed the analogy between the diffusivity equation in radial coordinate and in τ coordinate. Therefore, solving the 1-D equation in τ coordinate will generate pressure w.r.t time. Here, τ is embedding all the heterogeneities in the reservoir. In case of dual porosity reservoir model, fluid flow occurs only between fracture to fracture or between matrix to fracture. Fluid flow within matrix is negligible and can be ignored.

In a dual porosity model, **Eqs. 3.6** and **3.7** are solved separately to model fluid flow. Mass balance equation in fracture-fracture flow (Yang et. al, 2017):

$$\frac{\partial(\rho\phi_f)}{\partial t}-\nabla.\left(\frac{\rho}{\mu}k_f\nabla p_f\right)=-\rho_{up}\sigma\frac{k_m}{\mu_{up}}(p_f-p_m) \tag{3.7}$$

Mass balance in matrix-fracture flow (Yang et. al, 2017):

$$\frac{\partial(\rho\phi_m)}{\partial t} = \rho_{up}\sigma\frac{k_m}{\mu_{up}}(p_f - p_m) \tag{3.8}$$

Since in the dual porosity model, FMM is used to solve pressure propagation in fracture system only. The generated diffusive time of flight contours are then used to calculate drainage pore volume. The mass balance fluid flow equations **Eqs. 3.7** and **3.8** are transformed to 1-D τ coordinate. During this transformation, the mass balance equation in matrix-fracture fluid flow keeps the same form as single porosity model but the mass balance equation in fracture-fracture fluid flow takes the following form (Zhang et al, 2014 and 2016):

$$\frac{p_f\tilde{c}_t}{Z}\frac{\partial p_f}{\partial t} - \frac{1}{w(\tau)}\frac{\partial}{\partial\tau}\left(w(\tau)\frac{p_f}{\tilde{\mu}Z}\frac{\partial p_f}{\partial\tau}\right) = -\frac{1}{\phi c_{ti}}\left(\frac{p}{\mu Z}\right)_{up}\sigma k_m(p_f - p_m) \tag{3.9}$$

where, $\tilde{\mu}$ and $\tilde{c}_t$ are dimensionless viscosity and total compressibility (Zhang et al., 2014 and 2016).

3.2.2 DFN Upscaling (Oda's Method)

This study uses a synthetic unconventional dual porosity gas reservoir. This model has been designed using a several clusters randomly distributed in the reservoir map. Two extra ellipsoidal clusters of natural fractures are also put in model to create extra natural fracture density **(Fig. 3.1)**.

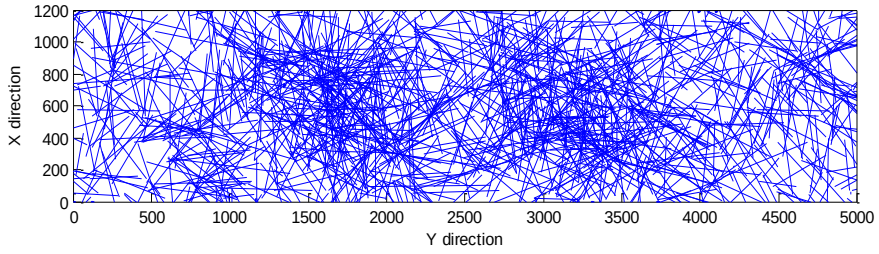


Figure 3.1 Natural Fracture distribution in the base model (Yang et al., 2017)

However, this model needs to be upscaled to corresponding permeability distribution before simulating gas production using FMM based forward simulator. In order to do that, Oda's method (Oda, 1985) was utilized in this study because of its simplicity and speed. Oda (1985) presented the following equation **(Eq. 3.10)** to calculate permeability tensor for a dual permeability dual porosity reservoir model. Equation for calculating permeability tensor in natural fractures is given by:

$$k_{ij}^{(c)} = \lambda\left(P_{kk}\delta_{ij} - P_{ij}\right) + a_{ij} \tag{3.10}$$

where,

λ = dimensionless constant ($0 < \lambda \leq 1/12$)

a_{ij} = correction term

δ_{ij} = Kronecker delta = $= \begin{cases} 0, if\ i \neq j \\ 1, if\ i = j \end{cases}$

$P_{kk} = P_{11} + P_{22} + P_{33}$ = summation of three principal component of the crack tensor P_{ij}

The crack tensor can be derived as,

$$P_{ij} = \frac{\pi\rho}{4}\int_0^\infty \int_0^\infty \int_\Omega r^2 t^3 n_i n_j E(n,r,t) d\Omega dr dt \tag{3.11}$$

where,

r = diameter of natural fracture

t = aperture of natural fractures

n_i, n_j = the components of a unit normal to the fracture

$E(n, r, t)$ = probability density function that describes the number of fractures whose unit vectors n are oriented within a small solid angle $d\Omega$

Ω = entire solid angle corresponding to the surface of a unit sphere

In a naturally fractured reservoir, each natural fracture has two opposing unit normal vectors $n^{(+)}$ and $n^{(-)}$. Dershowitz et al. (2000) presented a simpler way of using Oda's equations. The total number of natural fractures in a grid cell, N is given by:

$$N = \int_{\Omega} n_i n_j E(n) d\Omega \tag{3.12}$$

Plains of permeability are given by,

$$k_{ij} = \frac{1}{12}(F_{kk}\delta_{ij} - F_{ij}) \tag{3.13}$$

where,

F_{ij} = fracture tensor = $\frac{1}{V}\sum_{k=1}^{N} A_k T_k n_{ik} n_{jk}$ (3.14)

V = grid cell volume

A_k = fracture area of k^{th} natural fracture in a grid cell

T_k = transmissivity in k^{th} natural fracture in a grid cell

n_{ik}, n_{jk} = the components of a unit normal to the k^{th} fracture

The fracture system porosity, ϕ_F, is given by:

$$\phi_F = \frac{V_F}{V_{cell}} = \frac{\sum_{k=1}^{N} A_k.e}{V_{cell}} \quad (3.15)$$

where,

V_F = fracture system volume

V_{cell} = grid cell volume

N = number of fractures in a grid cell

A_k = fracture are of k^{th} fracture

e = fracture storage aperture

3.2.3 Hydraulic Fracturing Design

The unconventional shale gas model that is used in this study has a non-uniform permeability distribution due to non-uniform natural fracture density. The objective of this chapter is to optimize the hydraulic fracture design parameters for this reservoir model including location and number of hydraulic fractures, hydraulic fracture half lengths and widths. Economides et al. (2002 and 2012) and Daal and Economides (2006) reported the Unified Fracture Design algorithm to estimate the optimum hydraulic fracture dimensions for a given amount of hydraulic fracture treatment variables such as proppant amount. Propped volume in a single hydraulic fracture, V_p is given by (Economides et al., 2002 and 2012):

$$V_p = 2x_f w_f h_f \quad (3.16)$$

where,

x_f = fracture half-length

w_f= fracture average width respectively

h_f = fracture height

Mass of proppant used per stage, M_p is given by (Economides et al., 2002 and 2012):

$$M_p = V_p(1 - \phi_p)\rho_p \tag{3.17}$$

where,

V_p = propped volume per hydraulic fracture stage

ρ_p = proppant density

ϕ_p = porosity of proppant fracture

For a given fracturing fluid injection flow rate and corresponding pumping time, following equation can be derived keeping in consideration all the fluid losses occurring during fracture propagation (Economides et al., 2002 and 2012):

$$q_i t_e - \kappa(2h_f x_f)C_L\sqrt{t_e} - (2h_f x_f)S_p - x_f w_f h_f = 0 \tag{3.18}$$

where,

q_i = injection rate per half fracture of a bi-winged fracture

t_e = injection time

κ = the opening time distribution factor

C_L = the fluid leak-off coefficient for the formation

S_p = spurt loss coefficient

The total proppant laden slurry volume per stage can be calculated as by (Economides et al., 2002 and 2012):

$$V_{slurry} = 2q_i t_e \tag{3.19}$$

Lastly, the total fracturing fluid volume per fracture stage can be calculated as by (Economides et al., 2002 and 2012):

$$V_{fluid} = 2q_i t_e - \frac{M_p}{\rho_p} \tag{3.20}$$

3.2.4 Genetic Algorithm and Workflow

Since the main objective of this chapter is to optimize the hydraulic fracture design parameters in a given reservoir model, an optimization algorithm is needed to accomplish that. For this study, a class of evolutionary algorithms known as Genetic Algorithms (Holland 1992 and Mitchell 1999) is utilized. A Genetic Algorithm or GA is a derivative free optimization method based on natural selection process that mimics biological evolution. In this algorithm, population members of current generation are evaluated for their objective values and the population members of the next generation is reproduced based on parents from previous generation taking into consideration their corresponding objective values. Cheng et al. (2008) and Yin et al. (2010 and 2011) used GA to solve optimization problems very efficiently. This study follows the same GA algorithm used by Yin et al. (2010 and 2011). **Fig. 3.2** shows the GA approach used by them. A set of parameters are first identified with their minimum, maximum and base values. These parameters are needed to be calibrated in order to optimize the objective function value. Sensitivity analysis is first carried out for each of the parameters that need to be calibrated. Some parameters can then be removed in case the model is not affected much by changing their values. Next, an initial population of preset number of population member size is then created using Latin Hypercube Sampling (LHS) based Design of Experiment (DOE).

This method takes into account the full coverage of parameter ranges provided. Each initial population member is then used to update reservoir model used for this study and FMM based forward simulator is used to generate production profile. The optimization process in this study maximizes the Net Present Value (NPV) of the horizontal well with multiple hydraulic fractures created through it into the reservoir. Therefore, after each model simulation using FMM, NPV is calculated and stored as objective function value for corresponding population member. The GA continues to update by creating new population based on NPV (objective function value) values of previous generation. To create a new generation, fittest members of the previous generation are used for crossover or mutation so as to increase chances of creating better children. The fittest members are chosen based on the corresponding NPV values. Newer generations evolve from previous generations and try to reach optimum value after sufficient generations are reached or if the maximum limit of number of generations are reached as set before optimization process starts.

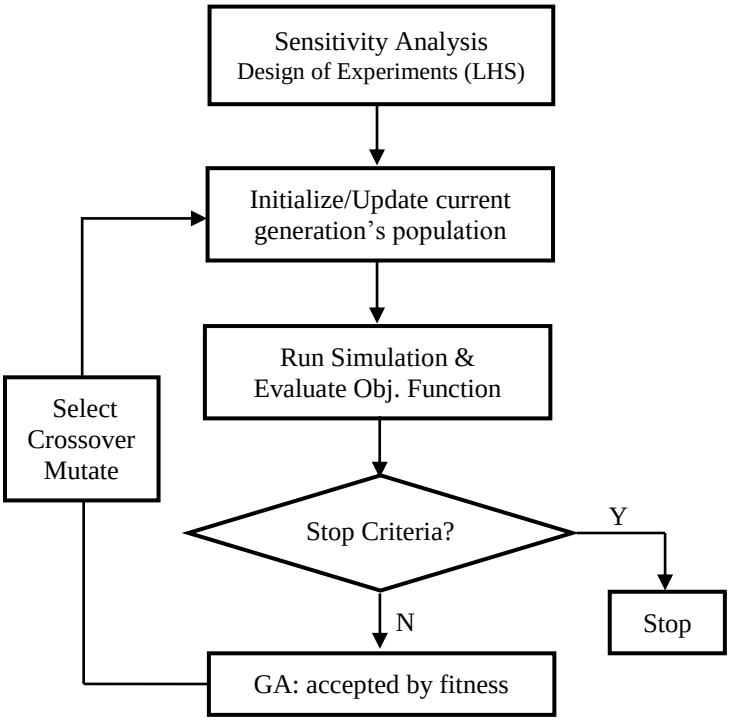


Figure 3.2 General workflow for genetic algorithm (Yang et al., 2017)

Fig. 3.3 shows the steps to calculate NPV in detail. The parameters needed to be optimized in this study are total number of hydraulic fracture, distances between hydraulic fractures, fracture half-length and their widths. Each model in GA is updated using new hydraulic fracture design parameters and corresponding permeability field is generated. Amount of proppant and fracturing fluid required for creating this hydraulic fracturing design can be calculated using **Eqs. 3.16** to **3.20**. Additional costs of equipment rent and horizontal well drilling can be added to fracturing cost to get cost of entire well. The revenue generated from well production can also be calculated based on gas prices and cumulative gas production generated by FMM simulator. Net Present Value, NPV can

then be calculated as the difference between the revenue generated by the well and the cost of well.

$$Revenue = \sum_{i=1}^{T} Revenue_i \tag{3.21}$$

$$Revenue_i = Revenue_{i-1} + (P_i - P_{i-1}) \times Gas\ Price \times \left(1 - \frac{r}{100}\right)^{t_i/365} \tag{3.22}$$

where,

T = total time of production

P_i = cumulative production at i^{th} time step

r = interest rate

$$Cost = Horizontal\ Well\ cost + Production\ Time \times Equipment\ Rent\ cost + Prop.Amount \times Prop.\ \ Price + Frac.Fluid\ Amount \times Frac.Fluid\ \ Price \tag{3.33}$$

$$NPV = Revenue - Cost \tag{3.34}$$

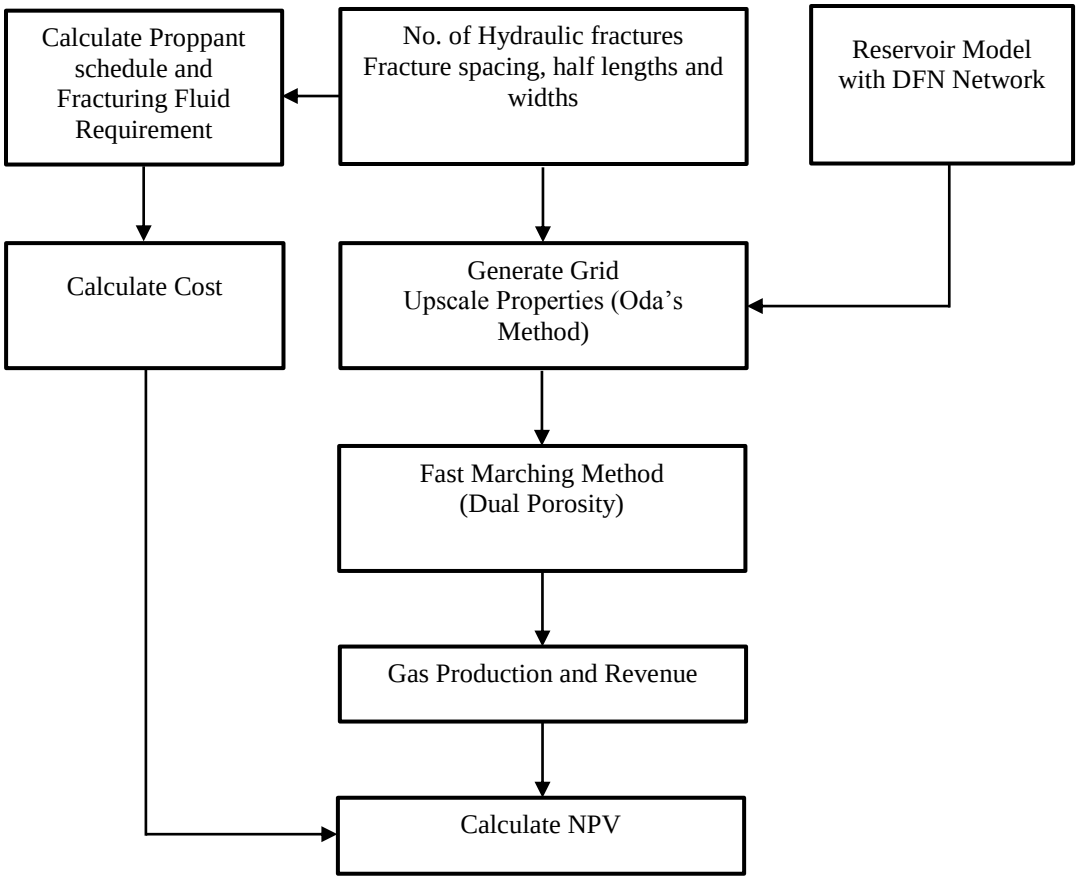
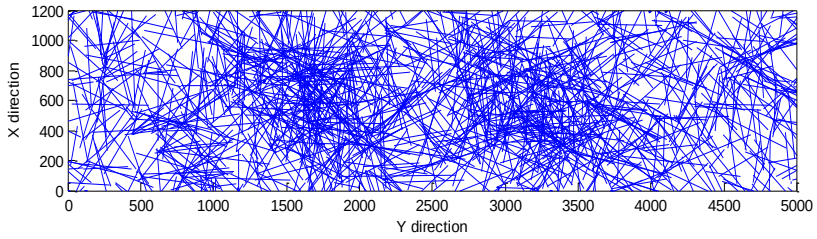


Figure 3.3 Workflow of objective function evaluation for each model (Yang et al., 2017)

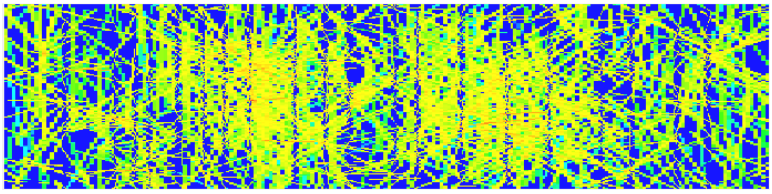
3.3 Results and Discussion

The first objective is to match the FMM prediction results with a commercial simulator Eclipse for the reservoir model. **Fig. 3.4** shows the upscaled permeability field derived from Oda method. Since the optimum values of hydraulic fracture design parameters are unknown, 15 hydraulic fractures with uniform spacing and half lengths are assumed. **Fig. 3.5** shows the comparison between the simulation results from FMM and Eclipse simulators. It may be observed from this figure that FMM is predicting gas rate very close to Eclipse results. However, the main advantage of FMM comes in terms of

time consumed for simulation. In this case, FMM was about 20 times faster that Eclipse making it a more suitable candidate for this optimization study which requires large number of simulations.



(a)



(b)

Figure 3.4 (a) Natural fracture distribution (b) Upscaled reservoir permeability field (Yang et al., 2017)

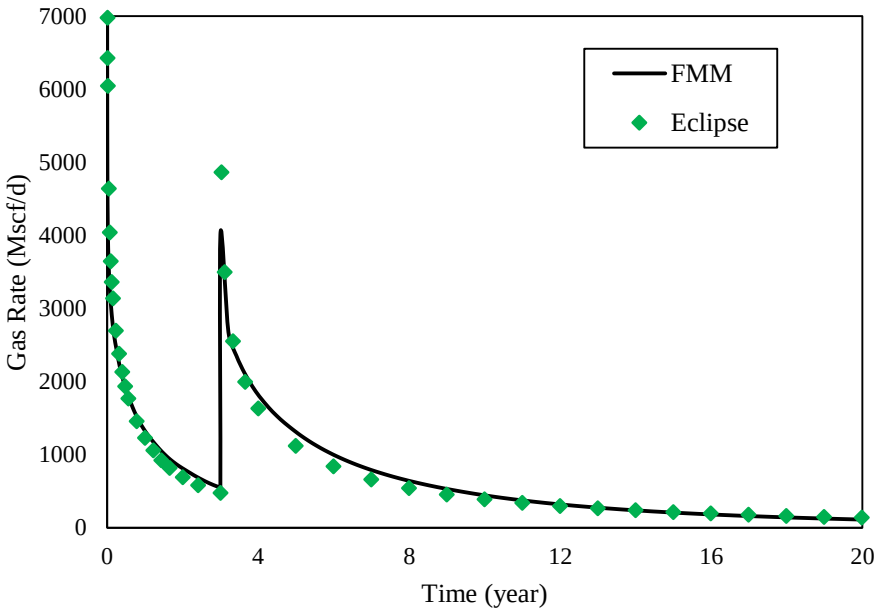


Figure 3.5 FMM versus Eclipse simulated gas production for the base model (Yang et al., 2017)

It should be noticed here that during application of the Oda's method presented in this study, a minimum matrix permeability is assumed to be approximately 10 nd. **Fig. 3.6** shows how the cumulative gas production changes with perturbation of this assumed cut-off value. **Table 3.1** shows the variation in NPV due to changing this minimum matrix permeability cut-off. Since the focus of this study is on the workflow for optimization and not studying the effect of variation of this matrix permeability, this study assumes the base value of 10 nd for this purpose.

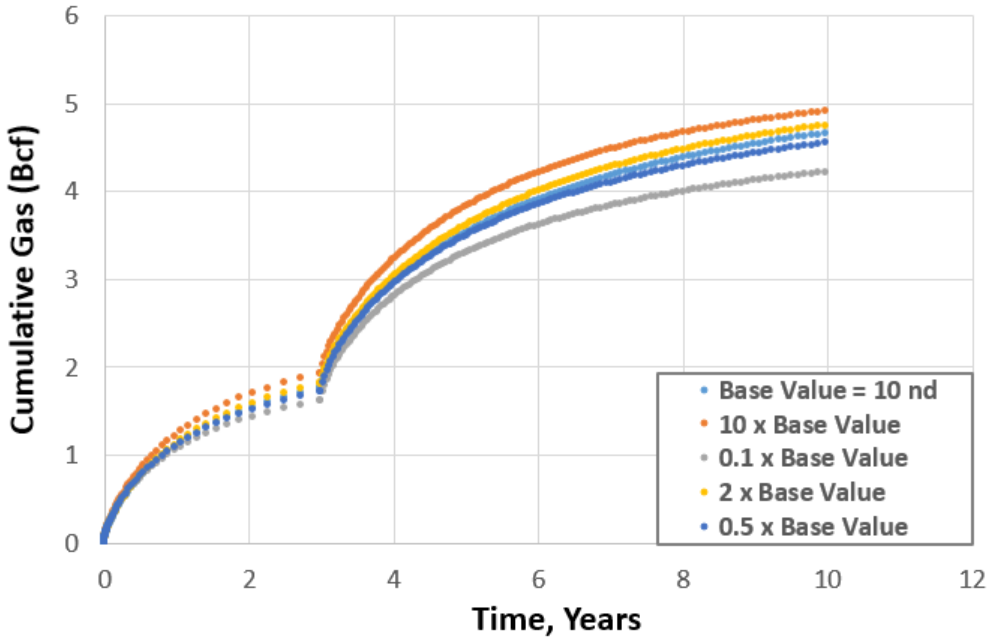


Figure 3.6 Effect of changing minimum matrix permeability during Oda's upscaling

Table 3.1 NPV variation with minimum matrix permeability used

Minimum Matrix Permeability	NPV
10 nd (Base Value)	8.06
10 x Base Value	8.87
0.1 x Base Value	7.07
2 x Base Value	8.33
0.5 x Base Value	7.84

Table 3.2 shows the economic parameters assumed in this study to calculate the Net Present Value (NPV) for a given hydraulic fracturing design. NPV is calculated as the

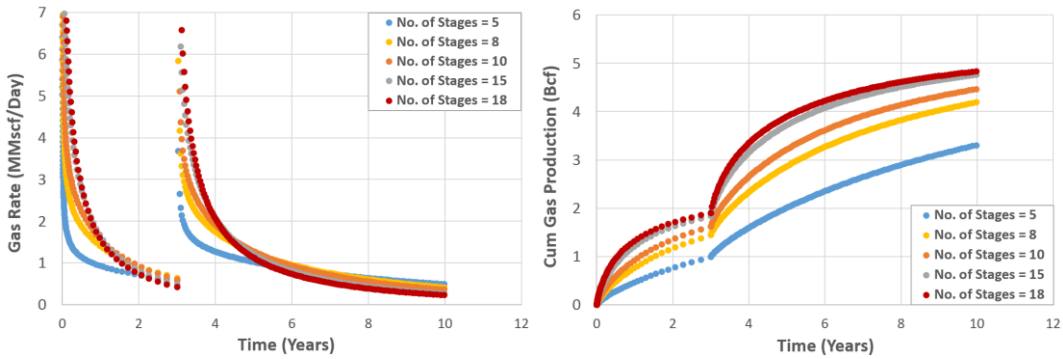
difference between the cumulative revenue generated during a specified period of well production time and the cost of well.

Table 3.2 Economic Parameters for NPV calculations

Properties	Value
Proppant Cost (USD/ton)	550
Fracturing Fluid Cost (USD/gal)	0.4
Horizontal Well Cost (USD/well)	1.2×10^{6}
Equipment Rent Cost (USD/min)	550
Interest Rate (1/year)	10%
Gas Price (USD/Mscf)	3.6

Fig. 3.7 shows the effect of changing the number of uniformly spaced hydraulic fractures on the well's production. It may be observed from this figure that cumulative production increases with increasing the number of hydraulic fracture stages in the reservoir. **Fig. 3.8** shows the effect of increasing number of hydraulic fracture stages on NPV. As can be observed from this figure, NPV increases at the beginning but then decreases with increasing fracture stages further. This is due to the fact that cumulative production does not improve significantly after certain number of stages. However the cost of fracturing increases due to larger amounts of proppant and fracturing fluid utilized for fracking job. Therefore, NPV starts to decline after a certain number of stages. Since a fracturing design problem such as this one involves more than one variables, there is a

need to come up with an optimization workflow which can provide best combination(s) of these variables for maximizing NPV. This study takes advantages of genetic algorithm to present such workflow.



(a) **(b)**

Figure 3.7 a) Gas Rates for various number of fracture stages b) Cumulative Gas Production for different numbers of fracture stages

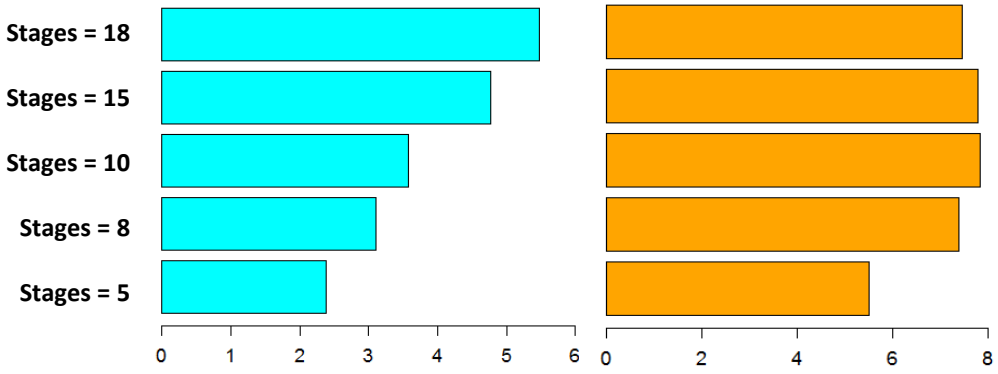


Figure 3.8 Cost and NPV comparison for various cases of number of fracture stages

Table 3.3 presents variable ranges used in this optimization study. For e.g., the number of stages can be between 8 and 18 including the boundaries. This range is decided based on previous results that resulted in maximum NPV within this range. Fracture width range is derived from the assumption that each hydraulic fracture is made up of a collection of 6 cracks on either side of the well and the width of each crack is of the order of thrice the diameter of a commonly used proppant.

Table 3.3 Hydraulic fracture optimization variable ranges

Variable	Min Value	Base Value	Max Value
Stages No.	8	12	18
Average Width (ft)	0.02	0.05	0.08
Fracture half-length (XF1 to XF4) (ft)	150	350	550
Fracture Spacing (DIS2 to DIS25) (ft)	100	250	400

Fig. 3.9 shows the sensitivity analysis results for this study. Each variable is perturbed to its maximum and minimum values as per the variable ranges presented previously and corresponding fractional change in NPV was calculated compared to the base NPV value (NPV resulting from keeping all variables at their base values). It may be observed from this figure that NPV is most sensitive to the average width in the current set of variable ranges. Although fracture spacing is not so sensitive in this figure, they can be more dominant if more than one fracture to fracture spacing is changed during

optimization study. Therefore, current study has kept all the variables for optimization process.

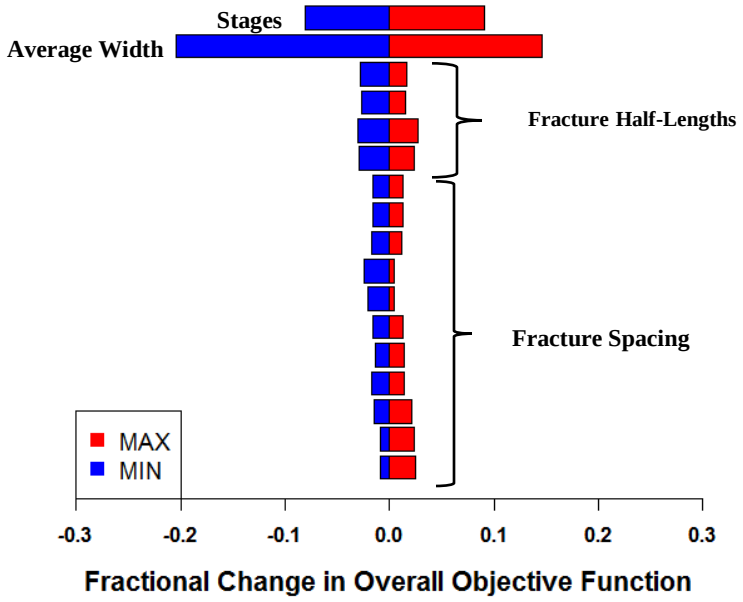


Figure 3.9 Sensitivity analysis of various variables on NPV

Fig. 3.10 shows the results from genetic algorithm based optimization of NPV. As explained in previous section of this chapter, genetic algorithm based optimization consists of updating generations based on previous generations based on cross over and mutation. As can be observed from this figure, subsequent generations tend to be better in terms of objective function NPV. **Fig. 3.11** shows variable distributions in the first generation and the last generation. It may be observed that the first generation consists of all possible values of this variable as provided in **Table 3.3**. However, as we move from first generation to last generation, this variable ranges shrinks. This shows that this algorithm is reaching an optimum set of variable values.

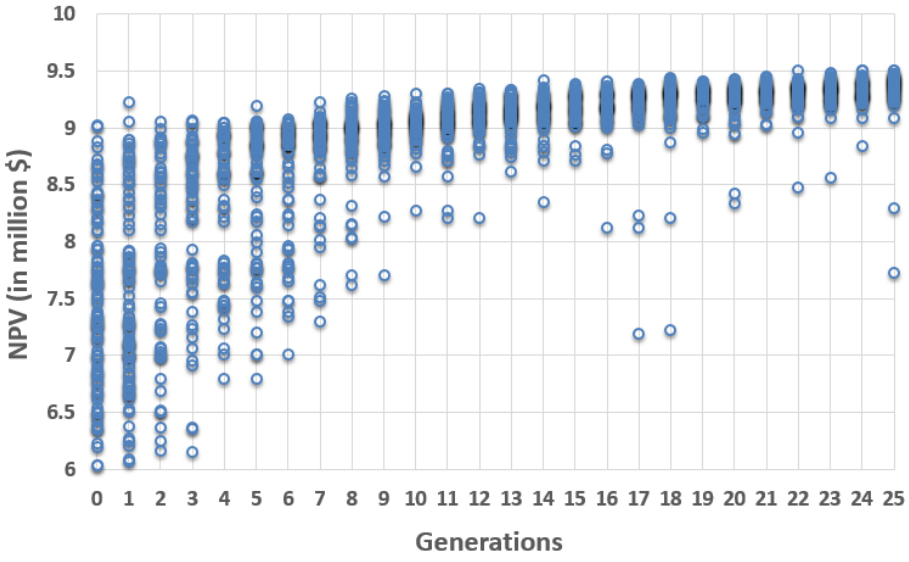


Figure 3.10 NPV distribution in Genetic Algorithm based optimization approach

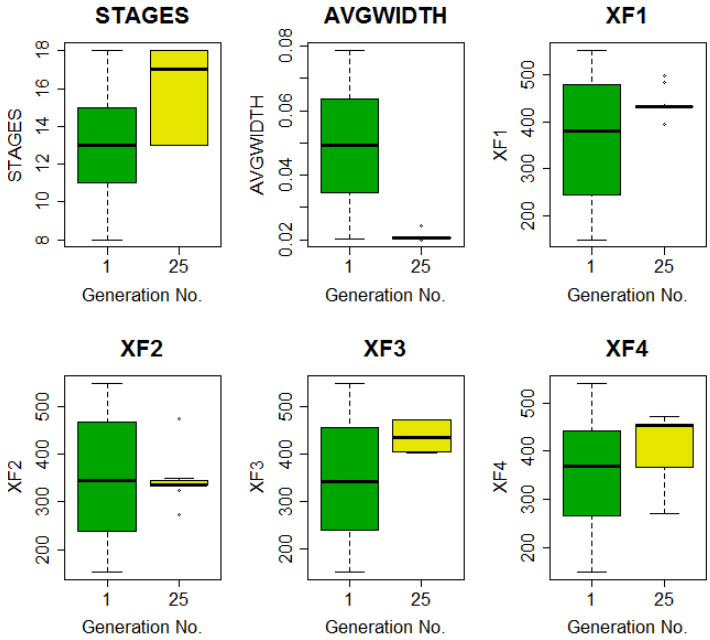


Figure 3.11 Distribution of fracture stages and average widths in generation 1 and generation 25

Fig. 3.12 shows the distribution of stage numbers in the first and the last generations. As can be observed from this figure, two optimum number of stage numbers are available in this problem – 13 and 18.

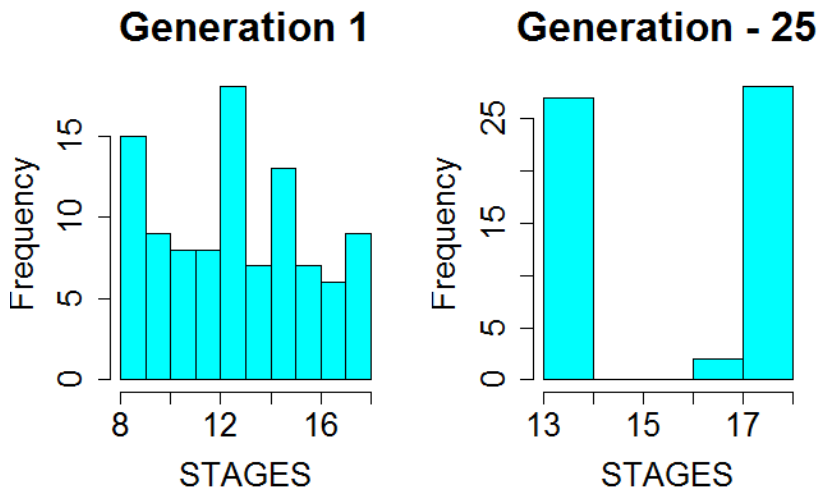
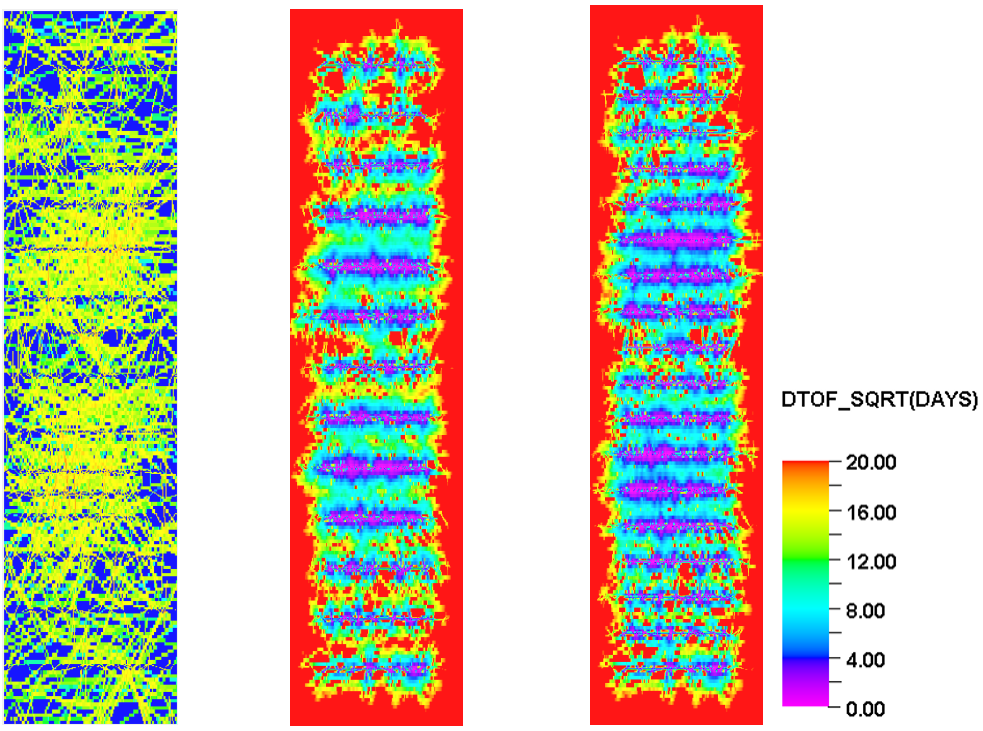


Figure 3.12 Distribution of fracture stages in generation 1 and generation 25

Figs. 3.13 and **3.14** show uniformly placed and optimally placed hydraulic fracture designs. Optimum designs corresponding to both 13 and 18 number of stages are compared in these figures. Comparing NPV values provided in **Figs. 3.13** and **3.14** shows that a reasonable improvement in NPV can be achieved by using the workflow utilized in this study.

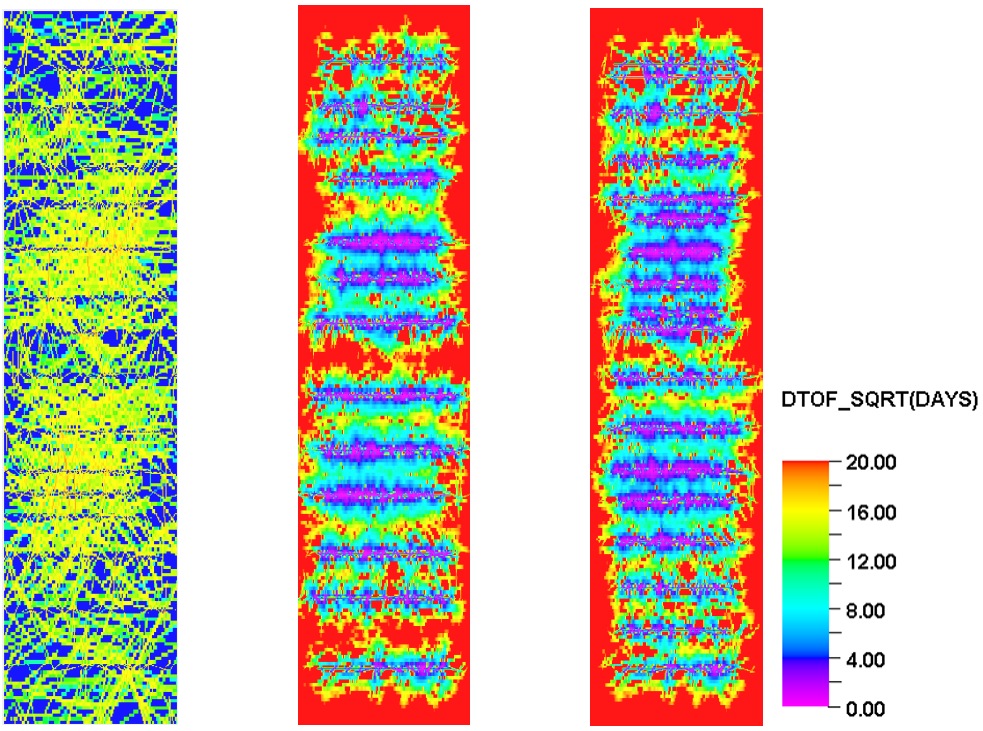


Uniform Design **Uniform Design**

(NPV = $ 7.97 million) **(NPV = $ 7.46 million)**

(13 fracs) **(18 fracs)**

Figure 3.13 NPV from Uniform spaced fractures



Optimum Design - 1 **Optimum Design - 2**

(NPV = $ 9.5 million) **(NPV = $ 9.5 million)**

(13 fracs) **(18 fracs)**

Figure 3.14 Hydraulic fracture placement in optimal design using genetic algorithm

Previous discussion assumed having a good knowledge about natural fracture distribution in the reservoir model. However, if there is some uncertainty present in natural fracture distribution, NPV based on multiple possible realizations can be chosen to be the objective needed to be maximized. **Fig. 3.15** shows possible realizations different from original base model presented before. In this case NPV can be integrated using:

$$NPV = \frac{1}{N}\sum_{i=1}^{N} w_i . NPV_i \tag{3.35}$$

where,

$NPV_i = NPV\ of\ the\ i^{th}\ realization$

$w_i = weights\ assigned\ to\ i^{th}\ NPV$

Fig. 3.16 shows the results from genetic algorithm based maximization of NPV calculated using **Eq. 3.35**. For this study equal weights have been assigned to all reservoir models. It can be observed here that genetic algorithm can successfully converge to a set of models having low variance in NPVs compared to initial set of population. **Fig 3.17** shows the variable distribution in the first and the last generations. It is clear from this figure that variable ranges in the last generation has shrunk compared to the first generation reducing uncertainty in those variables.

Fig. 3.18 shows the most optimum hydraulic fracture design based on multiple realizations when applied to the true model/base model. It can be seen here that the NPV has reduced from $ 9.5 million to 9.48 million when using the six realizations intead of the actual model for optimization problem. This small loss of NPV shows robustness of this algorithm using six realization. **Table. 3.4** shows the variation in NPV if the true model is one of the siz realizations or the base model presented earlier. It may be observed from the numbers provided in this table that moderate uncertainty in true model can have some effect on true NPV, i.e., it may be slightly higher or lower then the expected value. This minor change is however, insignificant compared to the difference between the optimum NPV and the NPV resulting from uniformly placed hydraulic fractures with base variable values presented earlier.

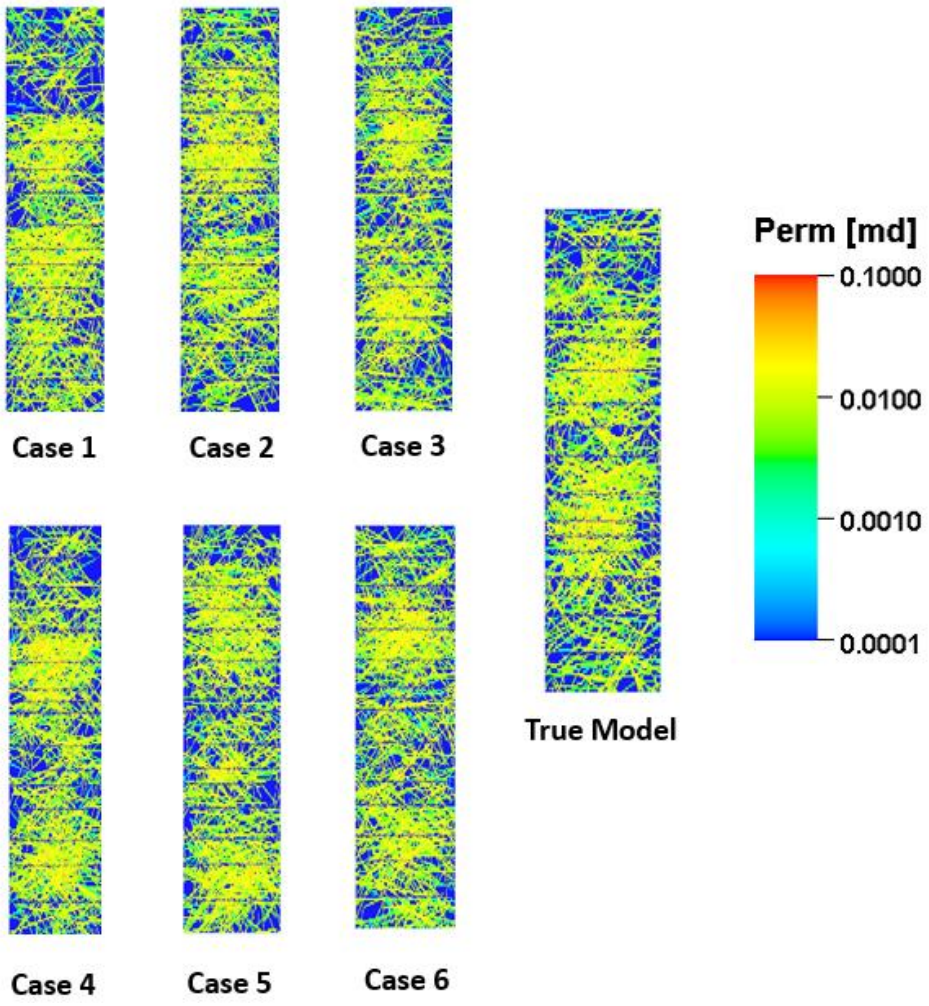


Figure 3.15 Six possible realizations vs true model/base model in case of uncertainty in natural fracture distribution

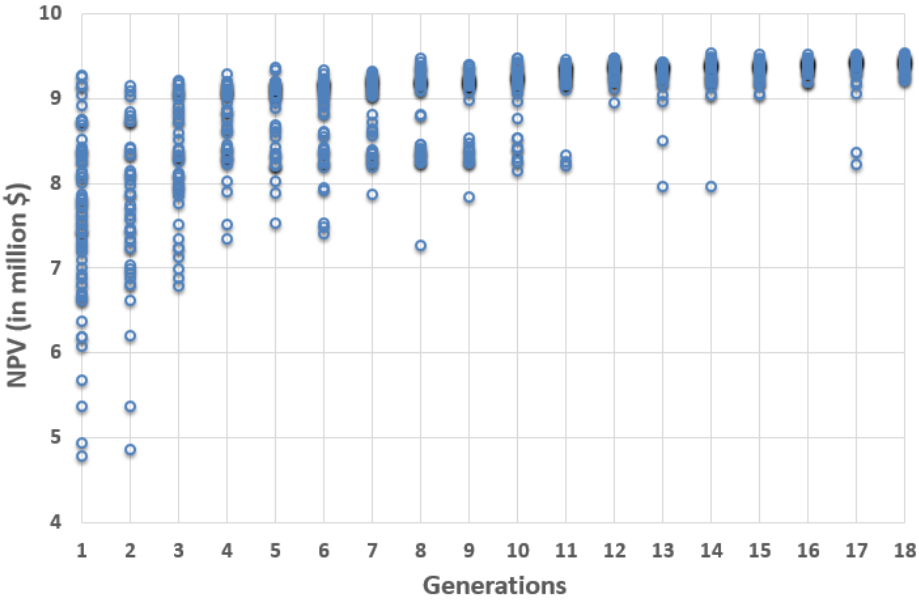


Figure 3.16 Results of genetic algorithm for multiple realization based optimization

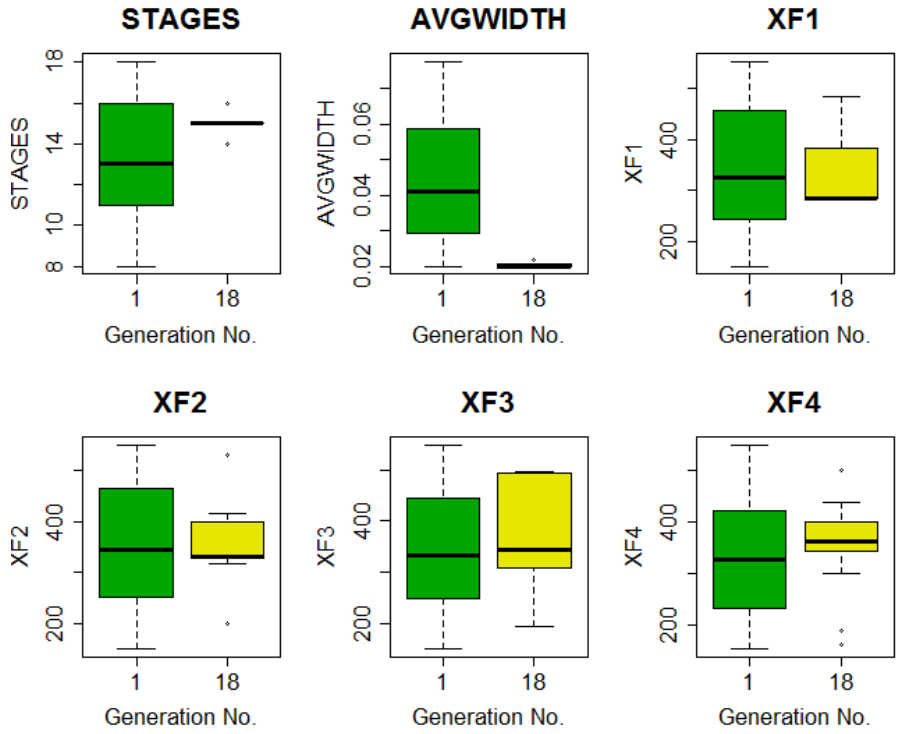


Figure 3.17 Variable distribution in the first generation vs last generation

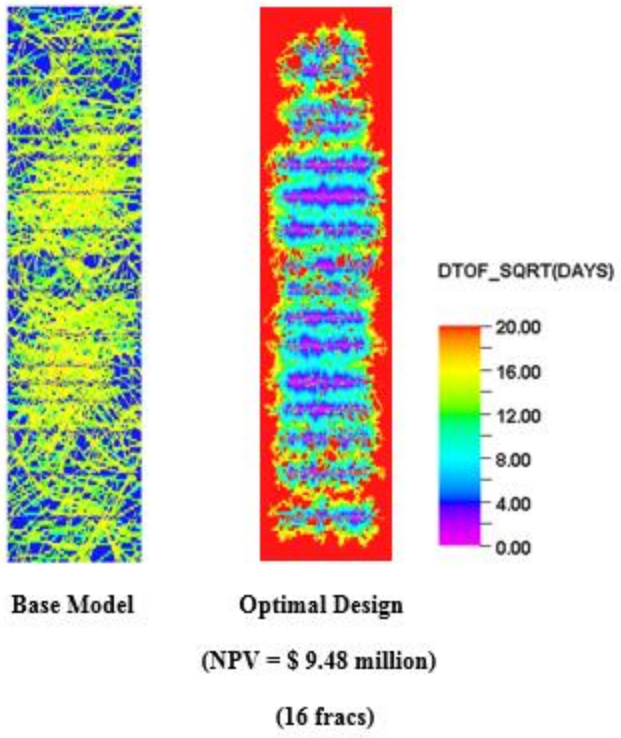


Figure 3.18 Hydraulic fracture placement in optimal design based on multiple realizations

Table 3.4 NPV values correponding to various realizations vs base model or true model

Realization	NPV
Base Model	9.48
Realization 1	9.48
Realization 2	9.44
Realization 3	9.52
Realization 4	9.54
Realization 5	9.61
Realization 6	9.56

3.4 Summary

1. For a given model with known natural fracture distribution, increasing number of hydraulic fractures would increase cumulative production but the corresponding cost of hydraulic fracturing would also increase. There is an optimum number of hydraulic fractures for a given reservoir model.
2. Genetic Algorithm based hydraulic fracture optimization workflow presented in this chapter can be utilized to maximize NPV by optimizing multiple hydraulic fracture variables such as number of hydraulic fractures, widths of hydraulic fractures, fracture half lengths and spacing between hydraulic fractures..
3. This chapter also presents how to deal with uncertainty in natural fracture distribution and presents the modified workflow for such cases. Variance in NPV due uncertainty in true model uncertainty has been presented for example case. Moderate uncertainty in true model can lead to small variation in expected NPV.
4. FMM based simulator has been proven to be an accurate and faster alternative to commercial simulator for an optimization study requiring large number of forward simulations.

CHAPTER IV

A MULTISTAGE GENETIC ALGORITHM FOR HISTORY MATCHING OF SHALE OIL RESERVOIRS: FIELD CASE STUDY[*]

4.1 Background and Introduction

This chapter deals with application of FMM based reservoir simulator in field case reservoir models. Since FMM has already been described earlier in Chapter 2, only the improvements in FMM simulator associated with upgrading it to a three phase and compositional simulator have been presented in this Chapter.

Zhang et al. (2014 and 2016) presented a genetic algorithm (GA) based history matching study in a field case using FMM based reservoir simulator. In their study, the reservoir model was divided into three groups – hydraulic fracture region, Stimulated Reservoir Region (SRV) and outer region. The SRV region is box shaped whose dimensions are needed to be calibrated during history matching. Hydraulic fractures are in transverse direction to the horizontal well and changed in vertical direction only. These hydraulic fractures are divided into several groups such that each group has hydraulic

[*] Parts of the text and data reported in this chapter is reprinted with permission from:

- Iino, A., Vyas, A., Huang, J., Datta-Gupta, A., Fujita, Y., Bansal, N. and Sankaran, S., April, 2017. Efficient Modeling and History Matching of Shale Oil Reservoirs Using the Fast Marching Method: Field Application and Validation. SPE Western Regional Meeting held in Bakersfield, California, USA. Copyright 2017 Society of Petroleum Engineers (SPE)
- Iino, A., Vyas, A., Huang, J., Datta-Gupta, A., Fujita, Y. and Sankaran, S., July, 2017. Rapid Compositional Simulation and History Matching of Shale Oil Reservoirs Using the Fast Marching Method. Unconventional Resources Technology Conference held in Austin, Texas, USA. Copyright 2017 Unconventional Resources Technology Conference (URTeC)

fractures with similar history matching parameters. This is done to reduce the number of parameters needed to be calibrated during history matching. This chapter study follows a similar approach of dividing current field case model into various regions before applying genetic algorithm based history matching.

4.2 Methodology

The methodology followed here is similar to Chapter 3 of this dissertation using GA. However different versions of FMM based reservoir simulators are applied in this studied incorporating both three phase and compositional field case models. A short description of dual porosity based two phase FMM simulator is provided in Chapter 3 of this dissertation. This study involves extending application of FMM based simulator to field case scenario for history matching purpose. Necessary updates in FMM based simulator have been incorporated (Iino et al. (2017)) and the newer versions of these simulators are applied in the field case study.

In the three phase FMM algorithm, single phase diffusivity is replaced by multiphase diffusivity (Iino et al. 2017):

$$\alpha_{mp} = \frac{\lambda_t k}{\phi c_t} \tag{4.1}$$

where,

λ_t = total mobility

c_t = total compressibility

Kazemi et. al. (1976) and Gilman and Kazemi (1983) reported following equations for mass balance in dual porosity model:

Mass balance equation for oil phase:

$$\frac{\partial}{\partial t}\left(\phi_f \frac{S_{of}}{B_o}\right) = \nabla.\left(\boldsymbol{k}_f \frac{k_{rof}}{B_o \mu_o} \nabla p_f\right) + \frac{\tilde{q}_o}{B_o} - \frac{\Gamma_o}{B_o} \tag{4.2}$$

Mass balance equation for water phase:

$$\frac{\partial}{\partial t}\left(\phi_f \frac{S_{wf}}{B_w}\right) = \nabla.\left(\boldsymbol{k}_f \frac{k_{rwf}}{B_w \mu_w} \nabla p_f\right) + \frac{\tilde{q}_w}{B_w} - \frac{\Gamma_w}{B_w} \tag{4.3}$$

Mass balance equation for gas phase:

$$\frac{\partial}{\partial t}\left[\phi_f \left(\frac{S_{gf}}{B_g} + R_s \frac{S_{of}}{B_o}\right)\right] = \nabla.\left[\boldsymbol{k}\left(\frac{k_{rgf}}{B_g \mu_g} + R_s \frac{k_{rof}}{B_o \mu_o}\right) \nabla p_f\right] + \left(\frac{\tilde{q}_g}{B_g} + R_s \frac{\tilde{q}_o}{B_o}\right) - \left(\frac{\Gamma_g}{B_g} + R_s \frac{\Gamma_o}{B_o}\right) \tag{4.4}$$

where,:

$$\Gamma_j = fluid\ transfer\ term = \sigma k_m \left(\frac{k_{rj}}{\mu_j}\right)\left(p_f - p_m\right) \tag{4.5}$$

σ = shape factor that depends on connectivity between matrix and surrounding fractures

j = phase type: oil/water/gas

To transform coordinate system from physical coordinates to τ coordinate, **Eq. 4.6** is used (Iino et al, 2017):

$$\nabla.\left(\boldsymbol{k}_f \frac{k_{rj}}{B_j \mu_j} \nabla p_f\right) \equiv -\frac{\phi_{f,ref}}{w(\tau)} \frac{\partial}{\partial \tau}\left[w(\tau)\left(\frac{c_t}{\lambda_t}\right)_{ref} \frac{k_{rj}}{B_j \mu_j} \frac{\partial p}{\partial \tau}\right] \tag{4.6}$$

The new mass balance equations for oil, water and gas phases then become (Iino et al., 2017):

Mass balance equation for oil phase:

$$\frac{\partial}{\partial t}\left(\phi_f \frac{S_{of}}{B_o}\right) = \frac{\phi_{f,ref}}{w(\tau)} \frac{\partial}{\partial \tau}\left(w(\tau)\left(\frac{c_t}{\lambda_t}\right)_{ref} \frac{k_{ro}}{B_o \mu_o} \frac{\partial p_f}{\partial \tau}\right) + \frac{\tilde{q}_o}{B_o} \delta(\tau_{wb}) - \frac{\Gamma_o}{B_o} \tag{4.7}$$

Mass balance equation for water phase:

$$\frac{\partial}{\partial t}\left(\phi_f \frac{S_{wf}}{B_w}\right) = \frac{\phi_{f,ref}}{w(\tau)}\frac{\partial}{\partial \tau}\left(w(\tau)\left(\frac{c_t}{\lambda_t}\right)_{ref}\frac{k_{rwf}}{B_w\mu_w}\frac{\partial p_f}{\partial \tau}\right) + \frac{\tilde{q}_w}{B_w}\delta(\tau_{wb}) - \frac{\Gamma_w}{B_w} \quad (4.8)$$

Mass balance equation for gas phase:

$$\frac{\partial}{\partial t}\left[\phi_f\left(\frac{S_{gf}}{B_g} + R_s\frac{S_{of}}{B_o}\right)\right] = \frac{\phi_{f,ref}}{w(\tau)}\frac{\partial}{\partial \tau}\left[w(\tau)\left(\frac{c_t}{\lambda_t}\right)_{ref}\left(\frac{k_{rgf}}{B_g\mu_g} + R_s\frac{k_{rof}}{B_o\mu_o}\right)\frac{\partial p_f}{\partial \tau}\right] + \left(\frac{\tilde{q}_g}{B_g} + R_s\frac{\tilde{q}_o}{B_o}\right)\delta(\tau_{wb}) - \left(\frac{\Gamma_g}{B_g} + R_s\frac{\Gamma_o}{B_o}\right) \quad (4.9)$$

Eqs. 4.7 to **4.9** show that mass balance equations can be solved w.r.t 1-D τ coordinate system. These equations can be solved using a finite difference method to calculate oil, water and gas rates. A detailed description of this FMM based reservoir simulator is provided in Iino et al. (2017). The compositional FMM version follows similar concept except that it incorporates compositional effects (Iino et. al, 2017)

The field case under investigation in this chapter is used to match history data and to forecast future production. The history matching problem in this chapter is based on Genetic Algorithm (GA). However instead of maximizing the objective function (NPV) as in the case of Chapter 2, the objective function in this study (mismatch error) is to be minimized in this chapter. The objective function or error function, $f(m)$, to be minimized in this study is given by:

$$f(m) = ln|\Delta Cum\_Oil| + ln|\Delta Cum\_Water| + ln|\Delta Cum\_Gas| \quad (4.1)$$

where,

$\Delta Cum\_Oil, \Delta Cum\_Water$ and ΔCum_Gas = root mean squared errors of observed cumulative production and simulated cumulative production for corresponding phases: oil/gas/water

Iino et al. (2017) presented history matching results using three phase FMM and compositional FMM. This study uses the same reservoir model but applies a slightly different approach of GA based workflow. **Fig. 4.1** shows various steps in history matching using this modified GA consisting of various GA stages. First, the objective function is tested for sensitivity w.r.t various reservoir model parameters needed to be calibrated for history matching. To calculate sensitivity, a parameter is perturbed to its maximum and minimum values keeping all other parameters at their base values. The relative change in the objective function compared to the base model (in which all parameters are kept at their base values) is calculated. This is repeated for all parameters to be calibrated one at a time and compared together in the end. Finally, an engineering judgement is made to decide if any parameter is needed to be removed from further study. If one or more parameters are not affecting the objective function significantly, they can be discarded for next GA stage. Once GA results show no further significant improvement (in terms of variable ranges and objective error values), the GA is stopped and a collection of best models are selected. Next, the updated variable ranges for the variables included in the previous GA stage is utilized for next GA stage. Also, the variables that were discarded in previous stage are also incorporated. Similar process is repeated in the next GA until reasonably good history matching results are observed.

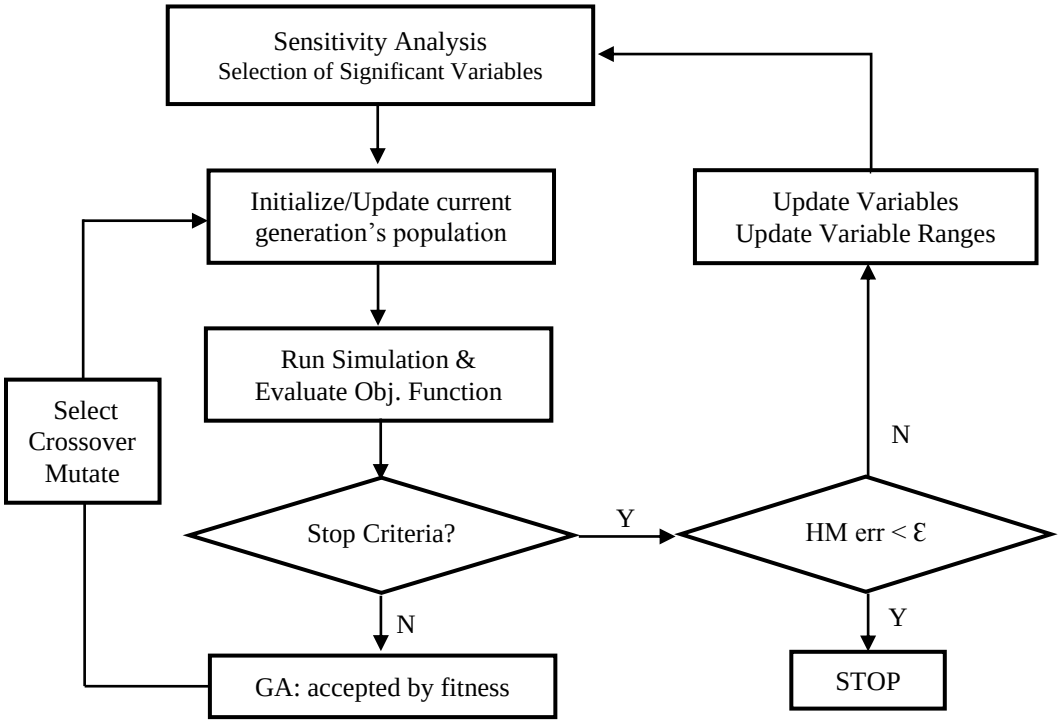


Figure 4.1 General workflow for genetic algorithm (GA)

4.3 Results and Discussion

The field case dual porosity model studied here is dimensioned 7,100 ft × 2,500 ft × 180 ft. The reservoir model has 71 × 25 × 13 (= 23,075) grid blocks. Initial reservoir pressure is 3,953 psi with bubble point pressure of 2,930 psi and therefore the reservoir is initially under saturated. The model has a single horizontal well with ten stages of hydraulic fractures. The model is divided mainly in three regions - Hydraulic Fractures, Stimulated Reservoir Volume (SRV) and non-SRV region (outer region) **(Fig. 4.2)**. **Table 4.1** lists various variables where the uncertainty exists with corresponding minimum and maximum values. The base values are the best estimate of a given variable. These variable ranges are determined with active discussions with the operator of this field.

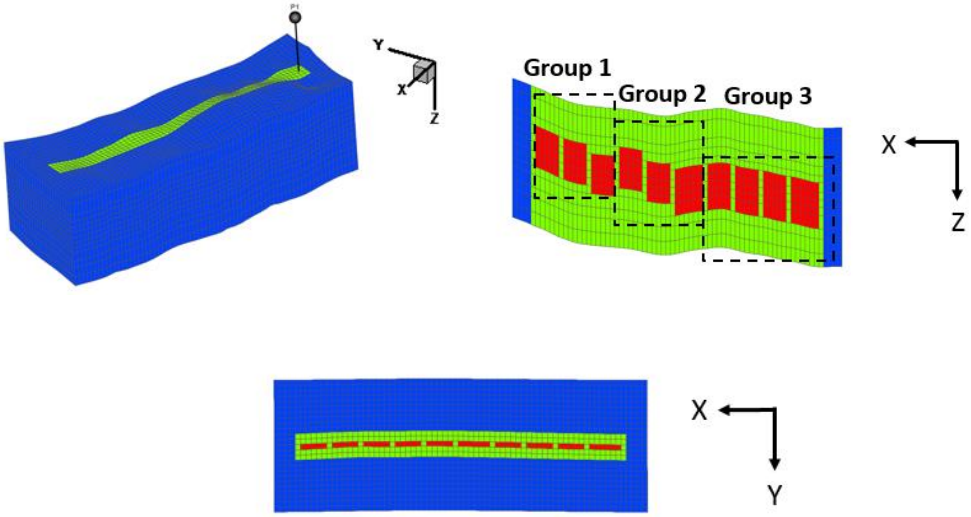


Figure 4.2 Three regions in the field case reservoir model

Table 4.1 Uncertainty in Model parameters and their base values for Sensitivity Analysis (Iino et al., 2017)

Region	Uncertain Parameters	Low	High	Base
Hydraulic Fracture	Porosity (HF_poro1, HF_poro2, HF_poro3)	0.005	0.02	0.01
	Permeability (mD) (HF_perm1, HF_perm2, HF_perm3)	0.2	3.0	0.55
	Water saturation (HF_S_{wi})	0.75	0.95	0.85
	Compaction table (HF_comp)	2	12	2
	Shape factor (ft^{-2}) (HF_sigma1, HF_sigma2, HF_sigma3)	0.0025	0.5	0.005
	Fracture half length (ft) (HF_Xf1, HF_Xf2, HF_Xf3)	50	150	50
	Fracture height (ft) (HF_h1, HF_h2, HF_h3)	40	100	60
	Stage length (ft) (HF_len1, HF_len2, HF_len3)	300-400	500-600	500-600
SRV	Porosity (SRV_poro1, SRV_poro2, SRV_poro3)	0.005	0.012	0.01
	Permeability (mD) (SRV_perm1, SRV_perm2, SRV_perm3)	0.01	0.2	0.1
	Water saturation (SRV_ S_{wi}1, SRV_ S_{wi}2, SRV_ S_{wi}3)	0.175	0.7	0.35
	Compaction table (SRV_ comp)	2	12	2
	Shape factor (ft^{-2}) (SRV_ sigma1, SRV_ sigma2, SRV_ sigma3)	1.25×10^{-4}	0.02	1.25×10^{-3}
	SRV_Width (ft) (SRV_W1, SRV_W2, SRV_W3)	300	900	500
Matrix	Porosity (Mat_poro)	0.059	0.094	0.08
	Permeability (Mat_perm), mD	2.3×10^{-7}	1.3×10^{-4}	2.7×10^{-5}
	Water saturation (Mat_S_{wi})	0.3	0.77	0.41
	Connate water saturation (Mat_S_{wc})	0.5*S_{wi}	1.0*S_{wi}	1.0*S_{wi}

4.3.1 History matching results based on GA and three phase FMM

Iino et al. (2017) presented a FMM based three phase unconventional reservoir simulator that is multiple times faster than a commercially available finite difference based reservoir simulator. This study applied FMM as a suitable candidate for history matching problem involving large number of simulations. Current study also utilizes the advantages of FMM for history matching. To test accuracy of FMM relative to Eclipse, simulations have been conducted for both FMM based simulator and Eclipse for the field case model under investigation using the base values of each variable. **Fig. 4.3** shows the well constraint utilized here which is tubing head pressure. **Figs. 4.4** to **4.9** present the comparison plots of the simulation results using three phase FMM simulator and Eclipse 100 simulator. It is clear from these figures that FMM and Eclipse are reasonably close to each other and therefore, FMM can be a good candidate for further history matching simulations due to faster simulations (Iino et. al, 2017).

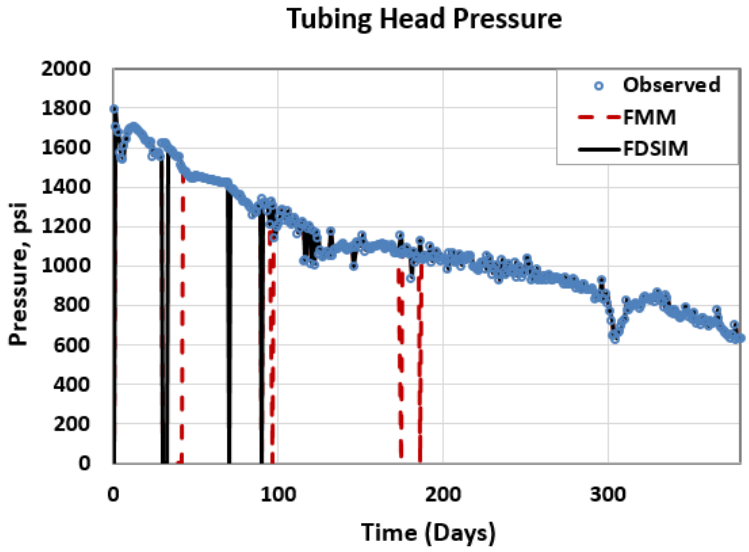


Figure 4.3 Well constraint Tubing Head Pressure during well production period

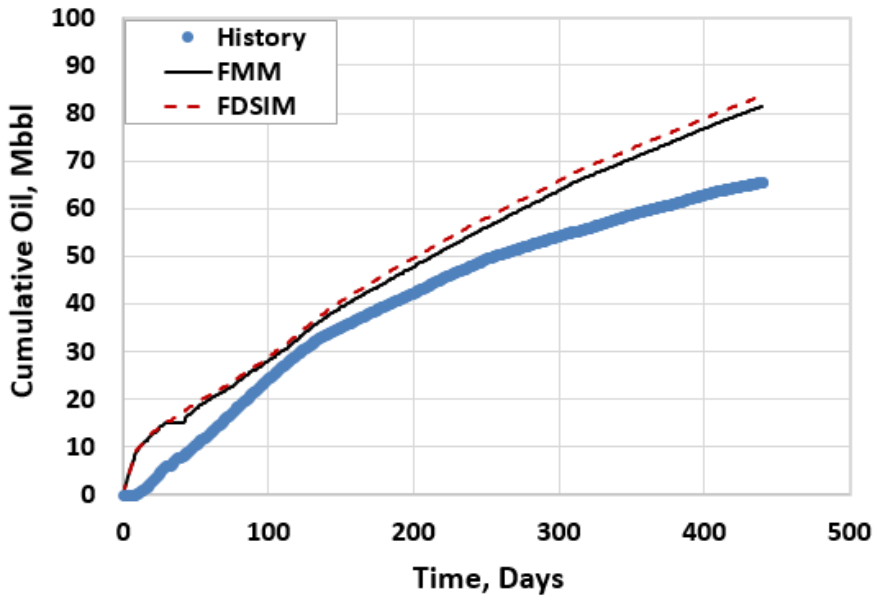


Figure 4.4 Cumulative Oil Production of FMM and Eclipse as compared to History data with base case variables (three phase FMM)

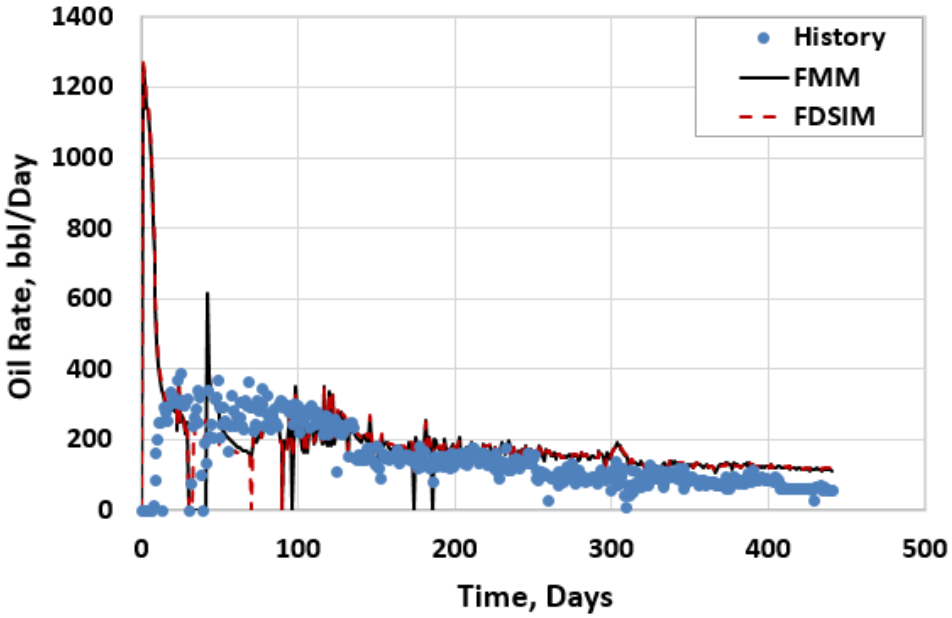


Figure 4.5 Oil Rate Production of FMM and Eclipse as compared to History data with base case variables (three phase FMM)

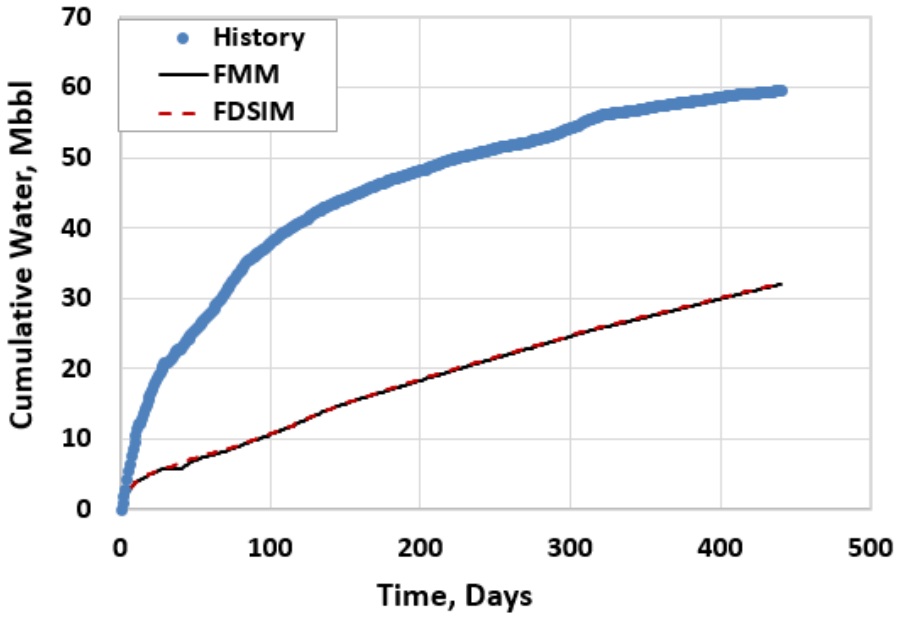


Figure 4.6 Cumulative Water Production of FMM and Eclipse as compared to History data with base case variables (three phase FMM)

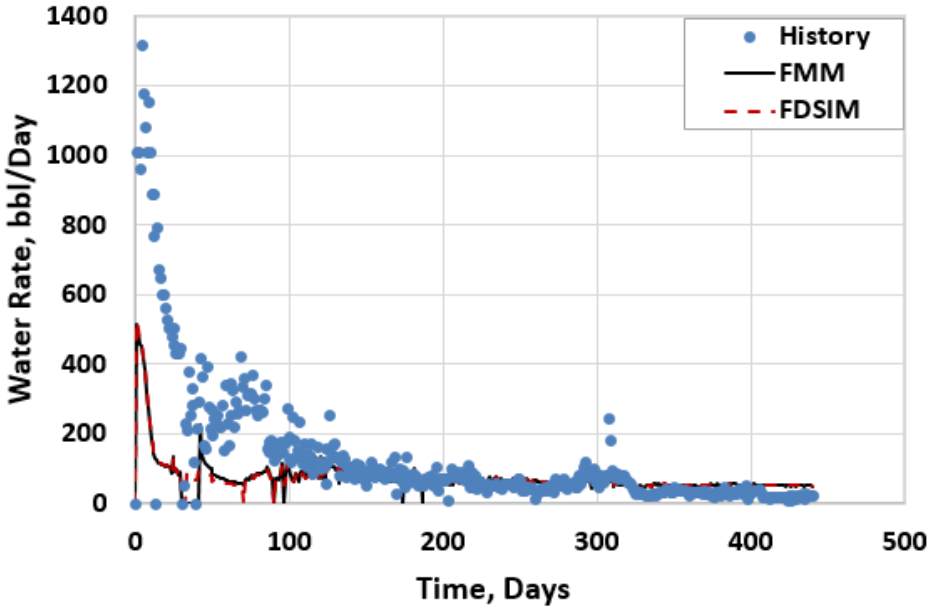


Figure 4.7 Water Rate Production of FMM and Eclipse as compared to History data with base case variables (three phase FMM)

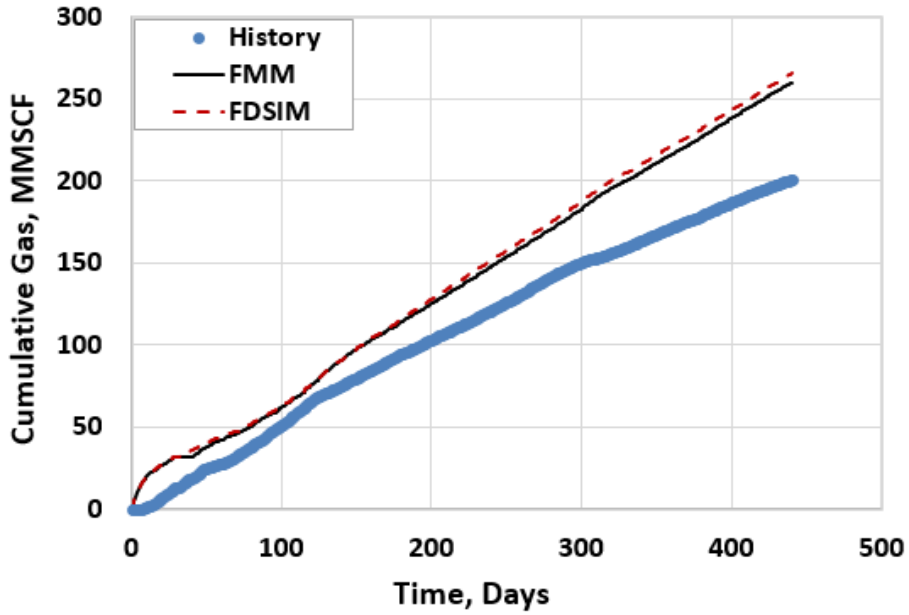


Figure 4.8 Cumulative Gas Production of FMM and Eclipse as compared to History data with base case variables (three phase FMM)

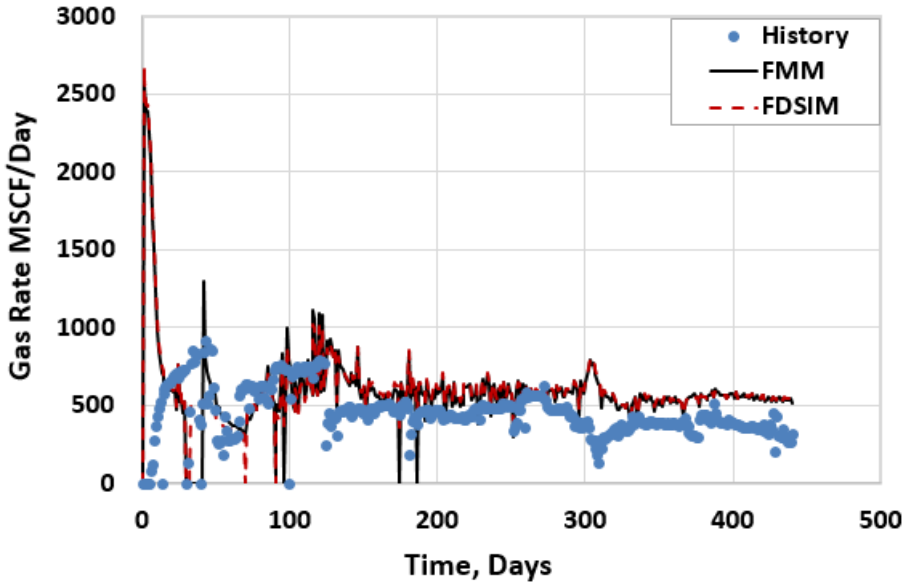


Figure 4.9 Gas Rate Production of FMM and Eclipse as compared to History data with base case variables (three phase FMM)

As presented in the previous section of this chapter, a multi-stage GA approach has been utilized for this study. In stage 1, sensitivity analysis is done and relative importance of variaous variables are checked. Heavy hitter variables or the variables making relatively larger impact on the objective error functions are identified and rest of the variables are discarded for this stage. **Fig 4.10** shows the results of sensitivity analysis. Parameters not included for this stage GA are shown in green boxes.

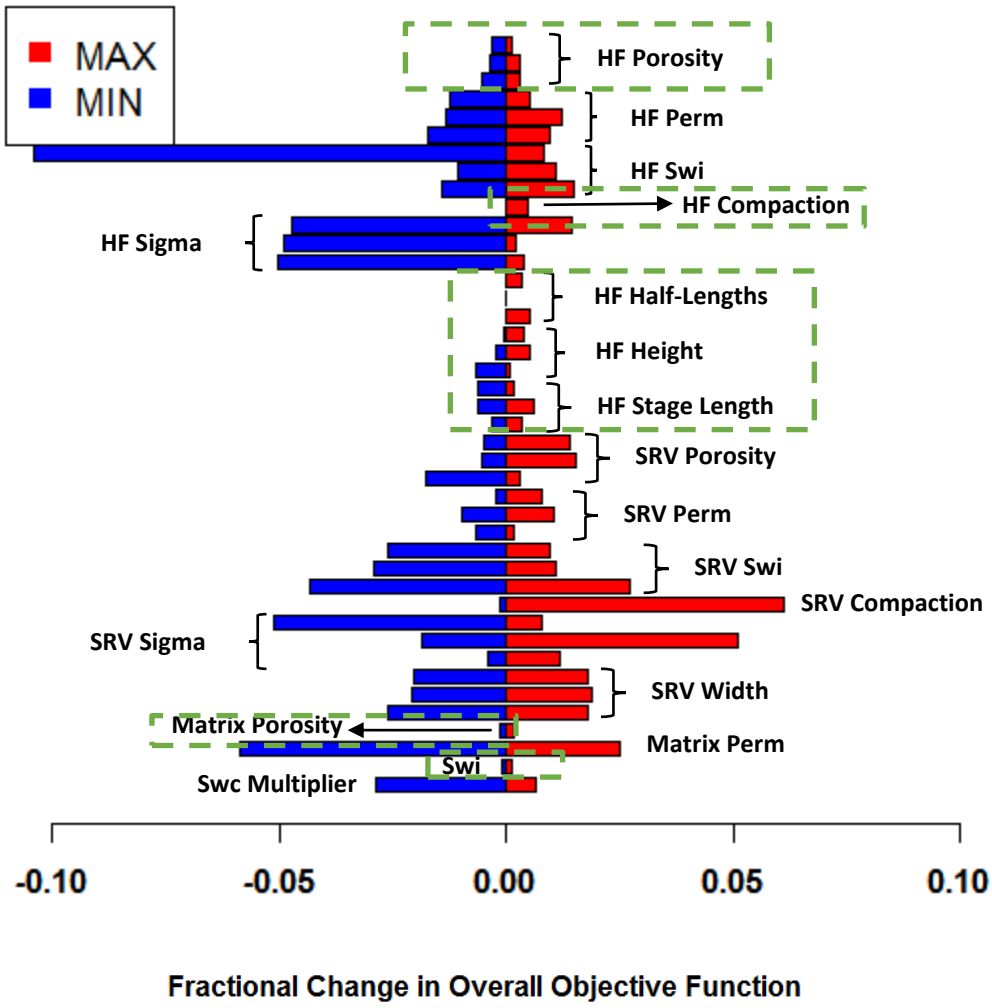


Figure 4.10 Sensitivity analysis at the beginning of Stage 1 (three phase FMM)

Fig. 4.11 shows the results of GA in stage 1. As can be observed from this figure, after multiple generations, improvement in objective error function reduces. Also, since variables in this GA operation show large shrinkage in their ranges from generation 1 to generation 12 **(Figs. 4.12 to 4.18)**, GA was stopped at this point and a collection of best models was selected **(Fig. 4.11).** These best models are chosen to derive new variable ranges of the variables included for the next GA stage. **Figs. 4.19 to 4.25** show the variable distribution in generation 1 of this stage while **Figs. 4.26 to 4.32** show the variable ranges in the best models selected at the end of this GA stage. It may be observed that a relatively uniform variable distribution transforms into a narrower and close to normal distribution.

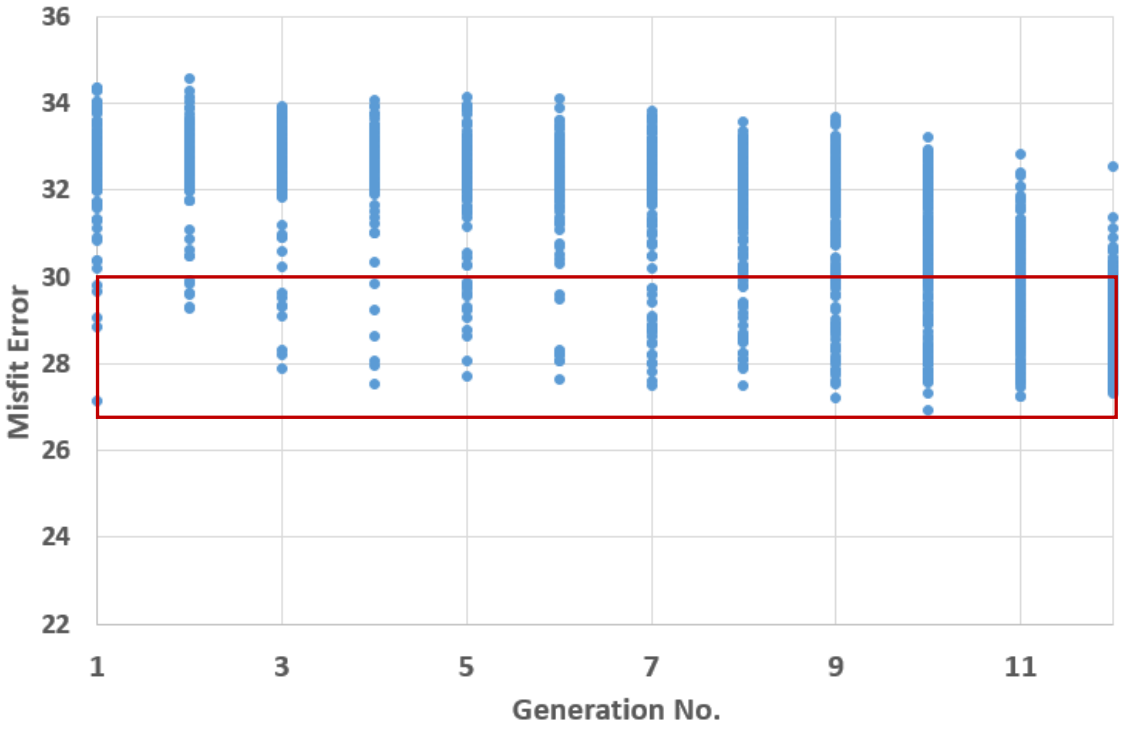


Figure 4.11 GA results for Stage 1 (three phase FMM)

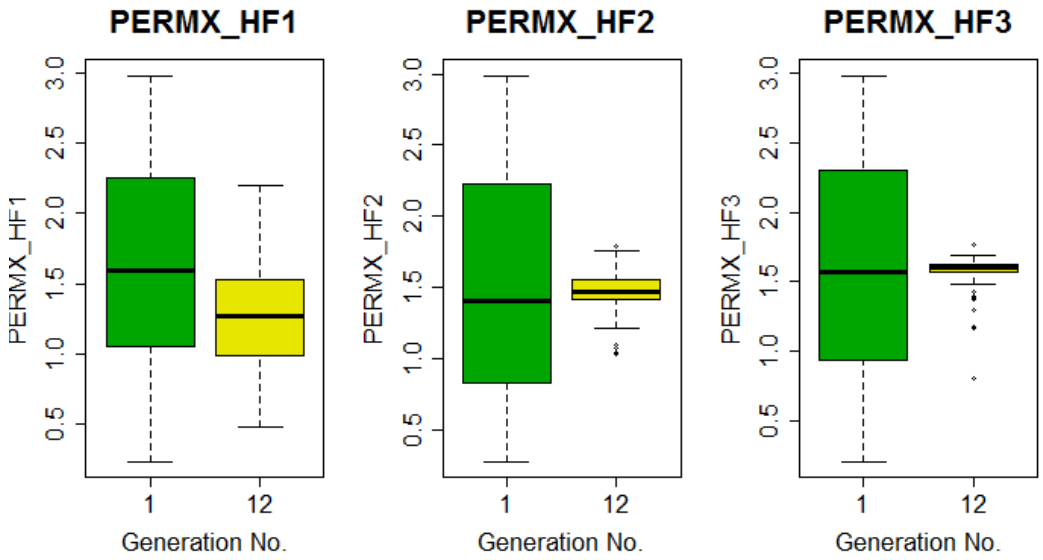


Figure 4.12 Uncertainty reduction in hydraulic fracture permeability during GA - Stage 1 (three phase FMM)

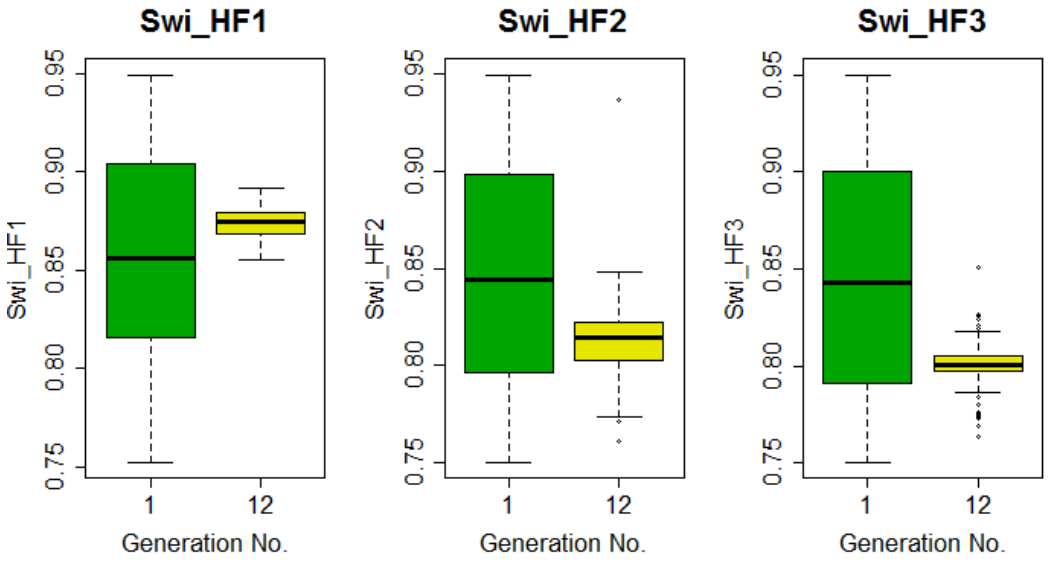


Figure 4.13 Uncertainty reduction in hydraulic fracture initial water saturation during GA - Stage 1 (three phase FMM)

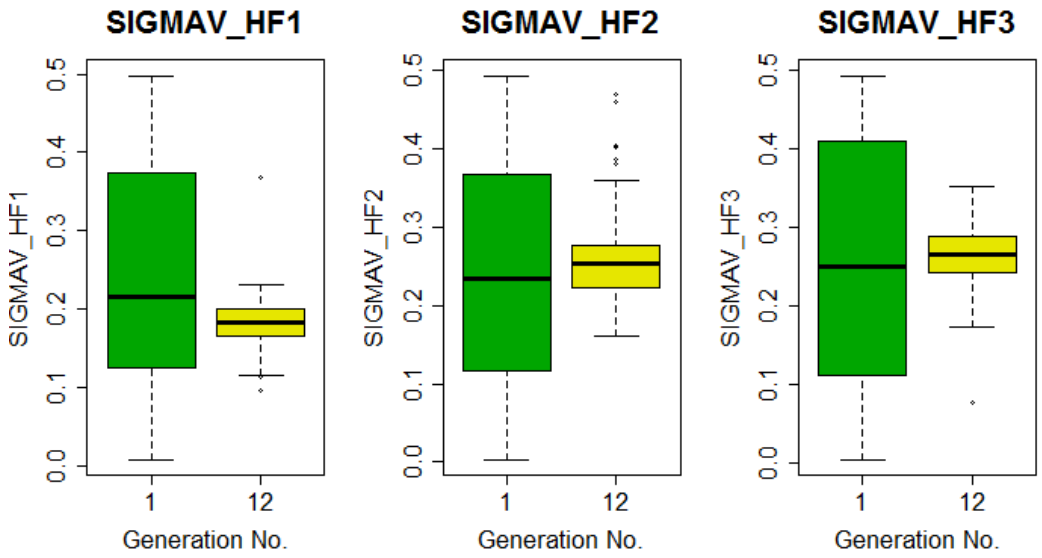


Figure 4.14 Uncertainty reduction in hydraulic fracture shape factor during GA - Stage 1 (three phase FMM)

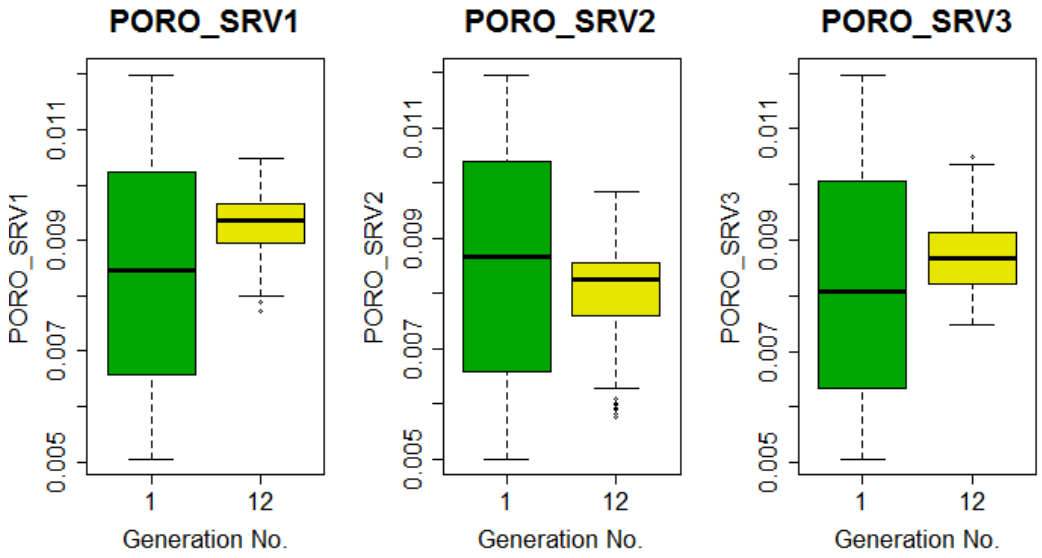


Figure 4.15 Uncertainty reduction in SRV porosity during GA - Stage 1 (three phase FMM)

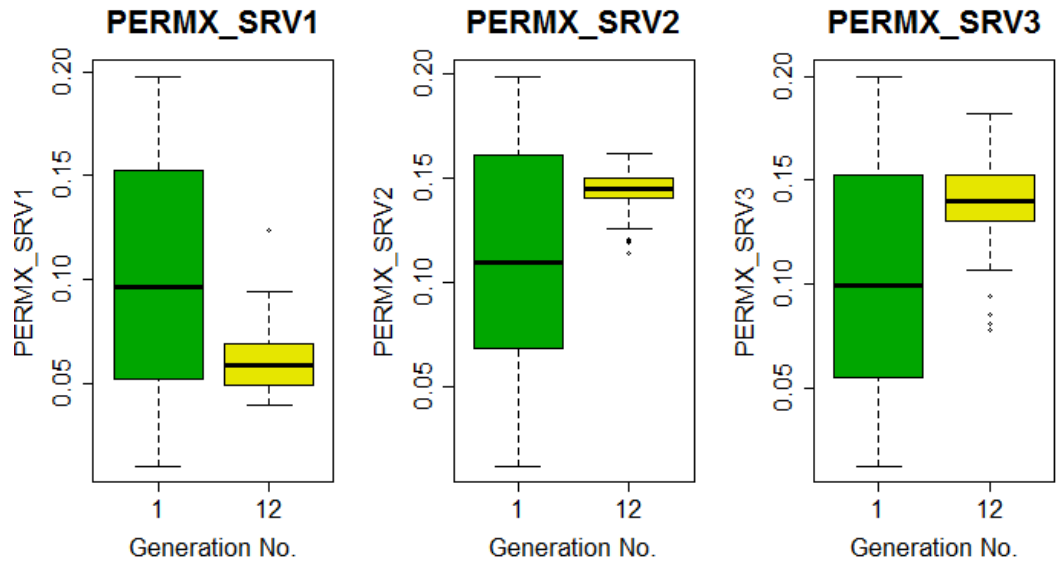


Figure 4.16 Uncertainty reduction in SRV permeability during GA - Stage 1 (three phase FMM)

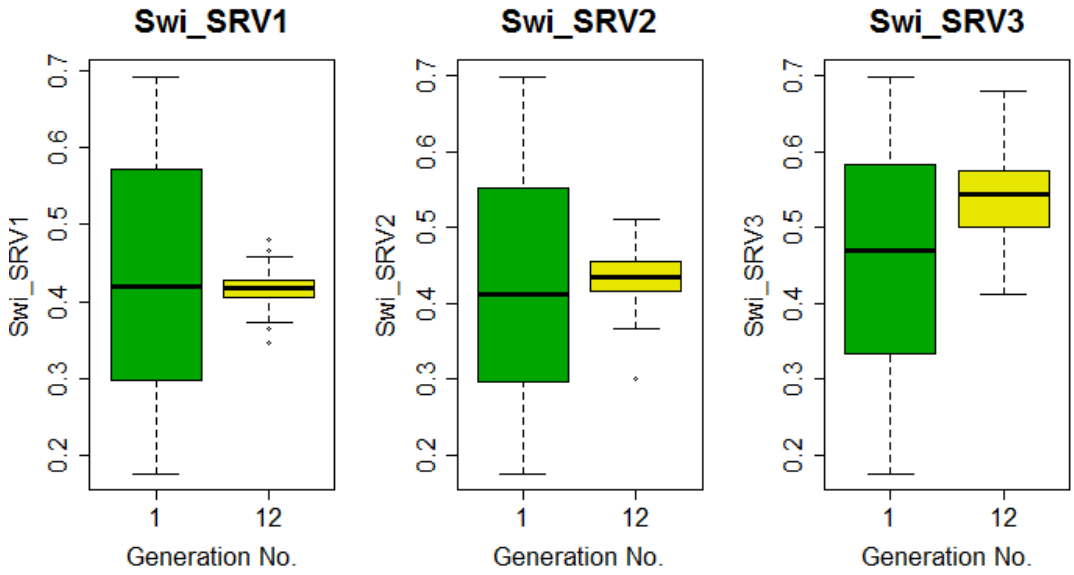


Figure 4.17 Uncertainty reduction in SRV initial water saturation during GA - Stage 1 (three phase FMM)

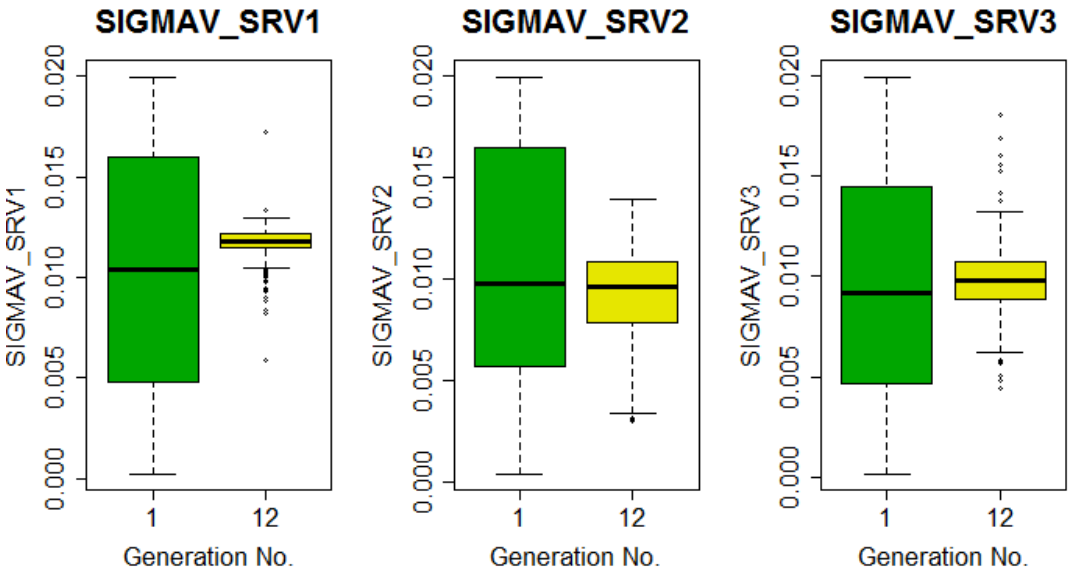


Figure 4.18 Uncertainty reduction in SRV shape factor during GA - Stage 1 (three phase FMM)

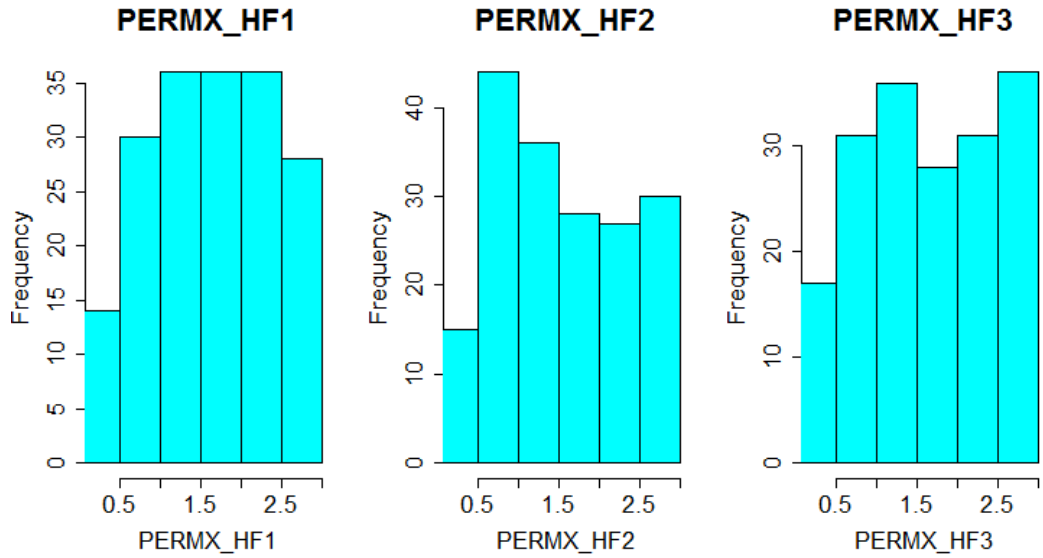


Figure 4.19 Variable distribution of hydraulic fracture permeability in the first generation of GA - Stage 1 (three phase FMM)

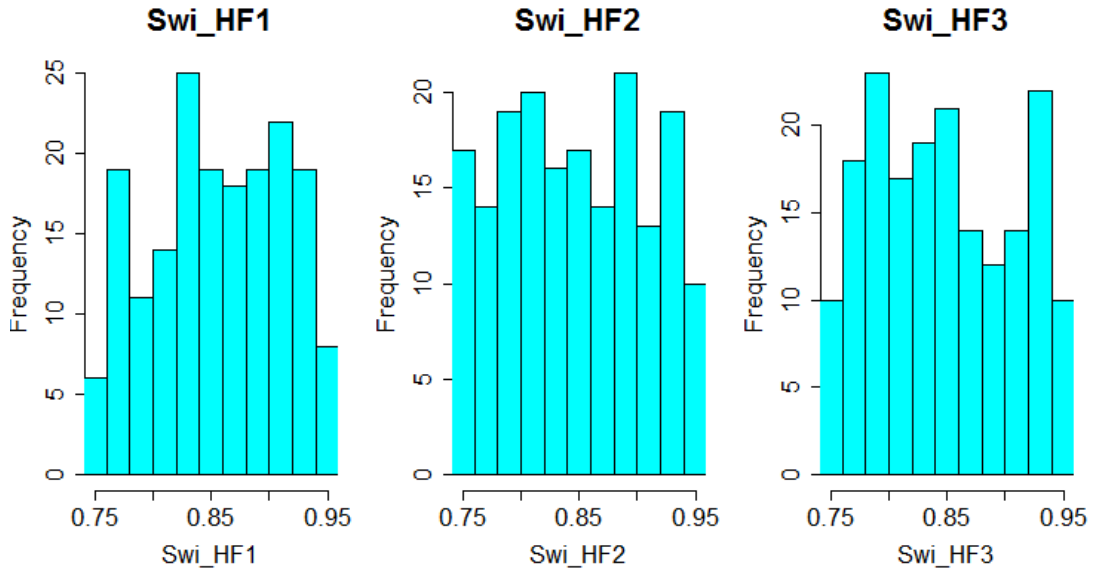


Figure 4.20 Variable distribution of hydraulic fracture initial water saturation in the first generation of GA - Stage 1 (three phase FMM)

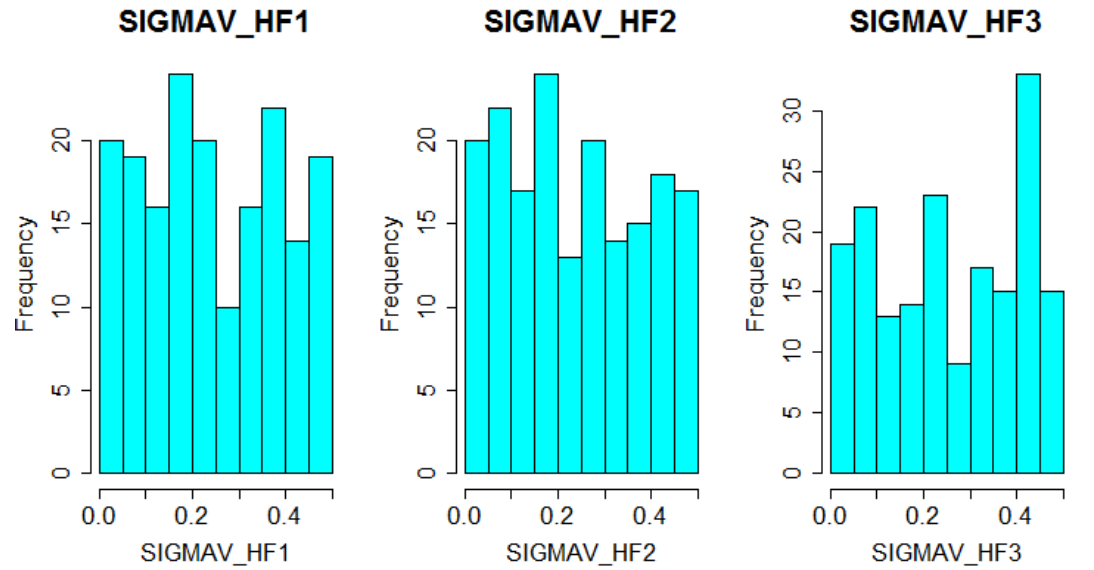


Figure 4.21 Variable distribution of hydraulic fracture shape factor in the first generation of GA - Stage 1 (three phase FMM)

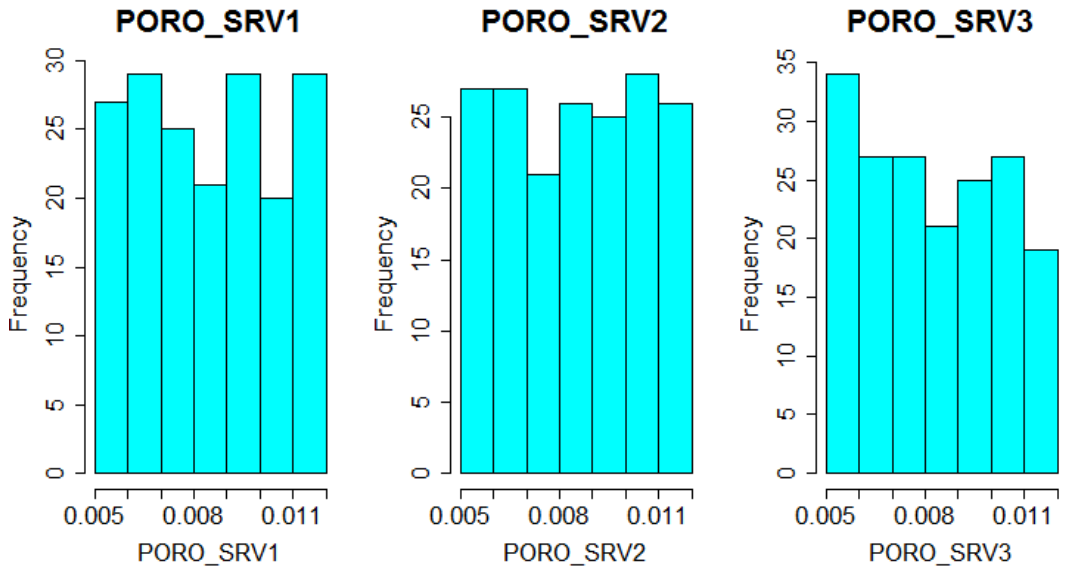


Figure 4.22 Variable distribution of SRV porosity in the first generation of GA - Stage 1 (three phase FMM)

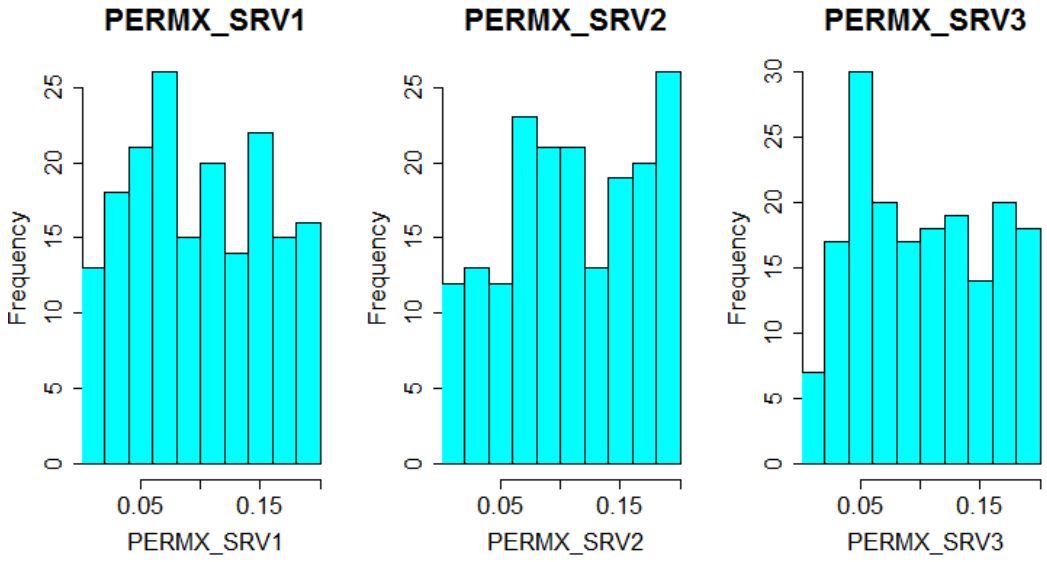


Figure 4.23 Variable distribution of SRV permeability in the first generation of GA - Stage 1 (three phase FMM)

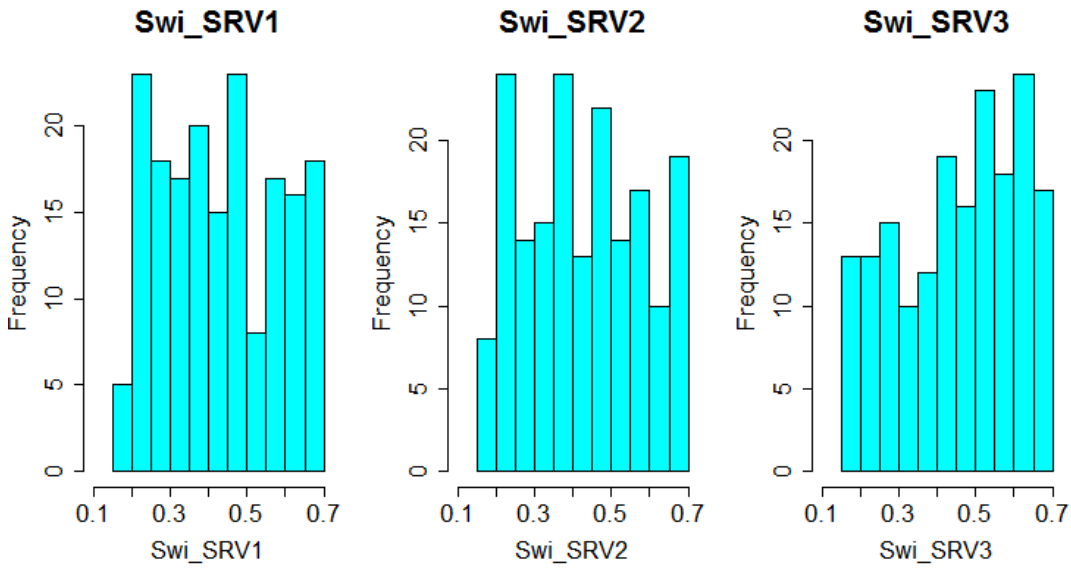


Figure 4.24 Variable distribution of SRV initial water saturation in the first generation of GA - Stage 1 (three phase FMM)

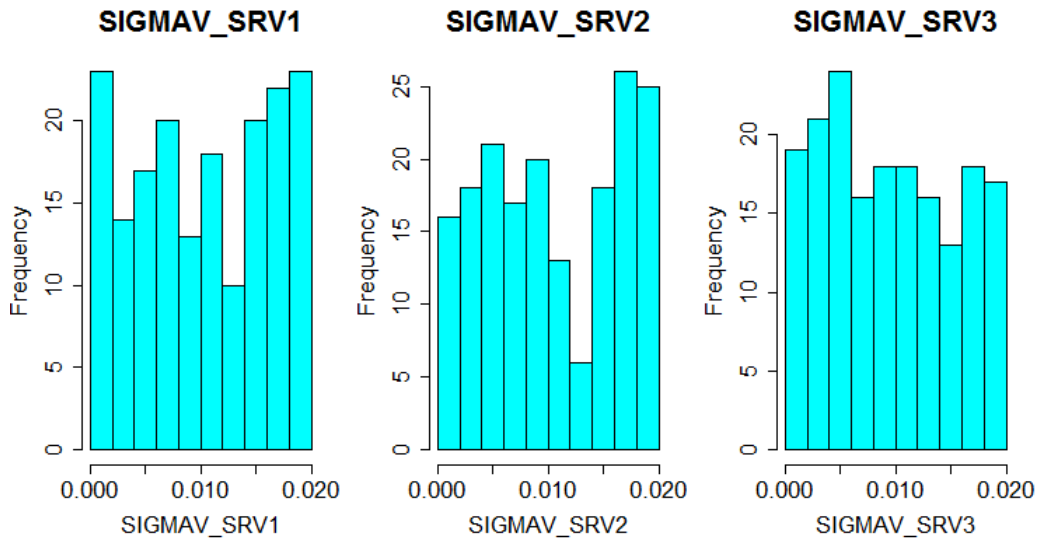


Figure 4.25 Variable distribution of SRV shape factor in the first generation of GA - Stage 1 (three phase FMM)

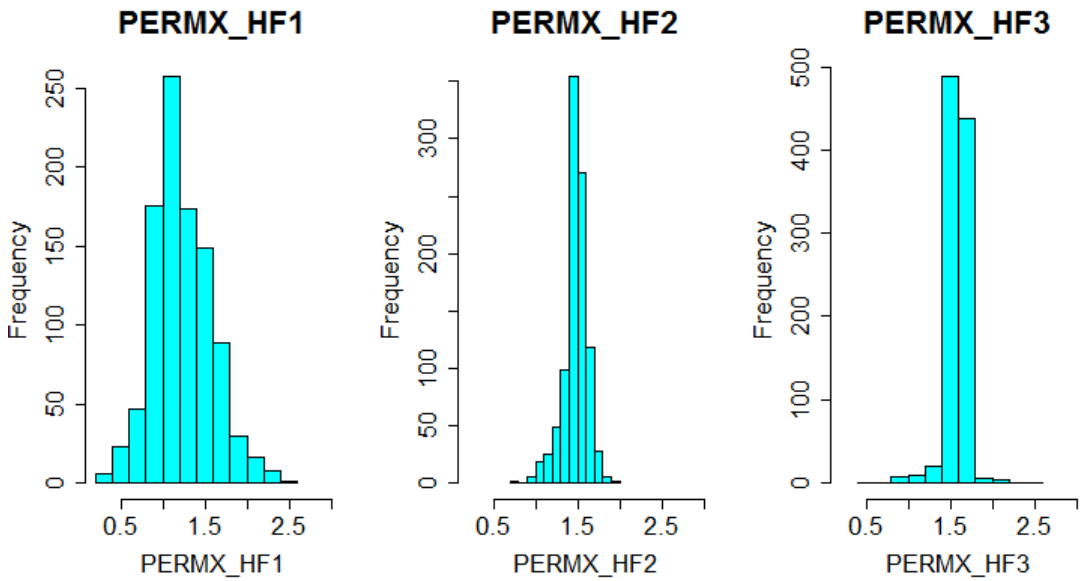


Figure 4.26 Variable distribution of hydraulic fracture permeability in the best selected models of GA - Stage 1 (three phase FMM)

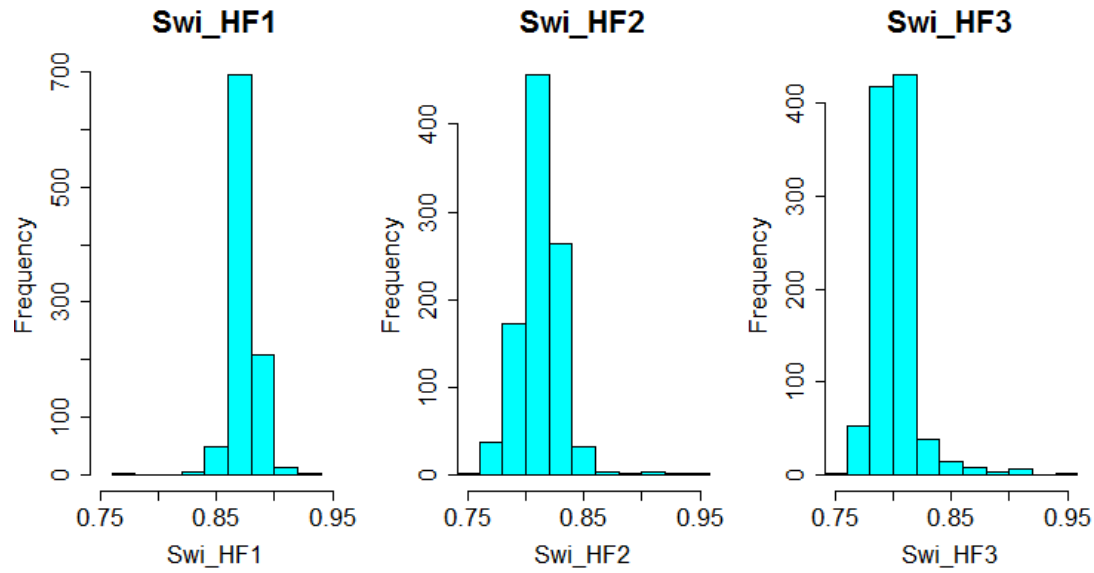


Figure 4.27 Variable distribution of hydraulic fracture initial water saturation in the best selected models of GA - Stage 1 (three phase FMM)

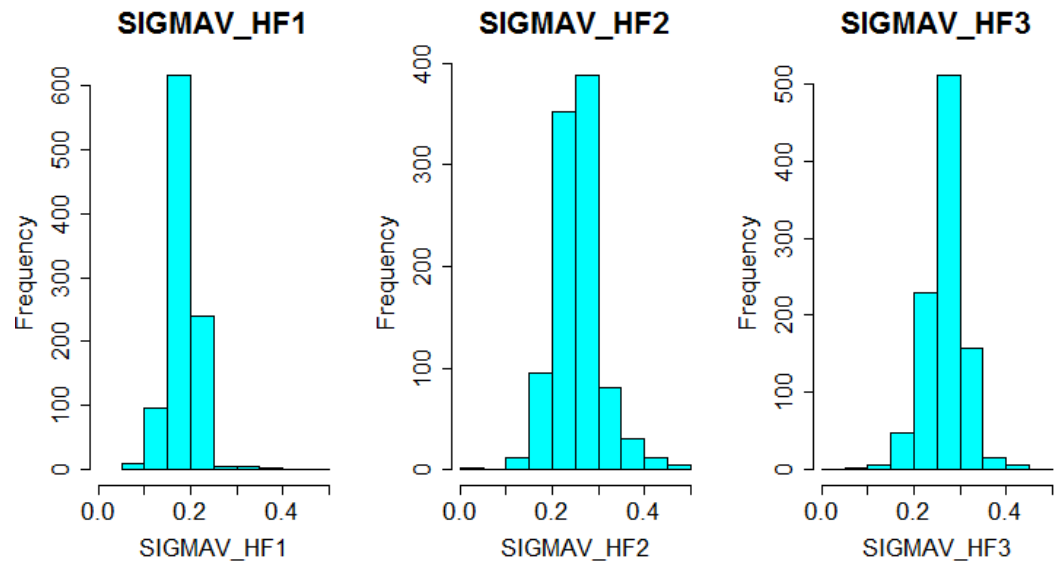


Figure 4.28 Variable distribution of hydraulic fracture shape factor in the best selected models of GA - Stage 1 (three phase FMM)

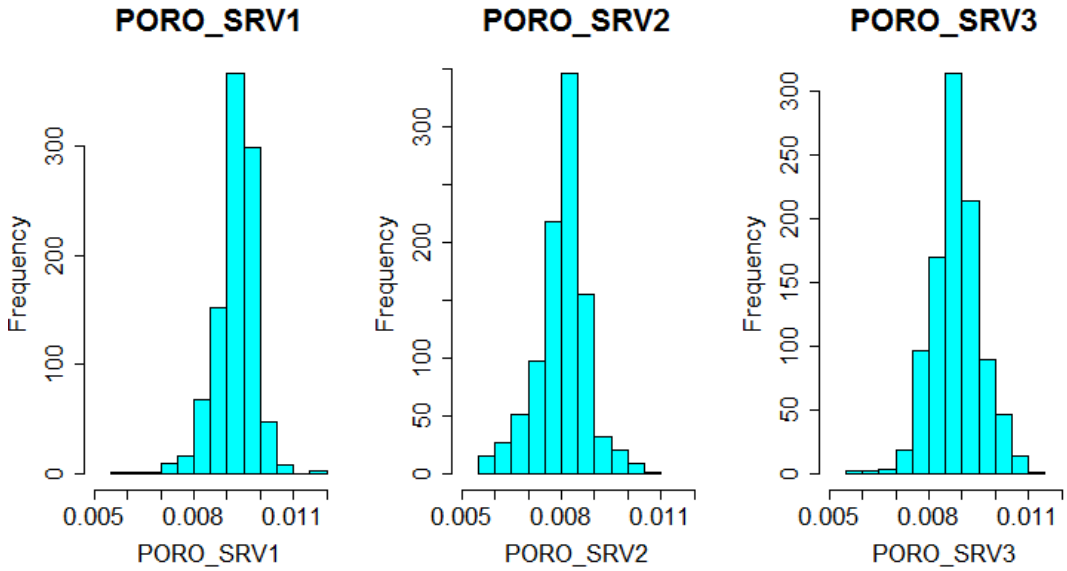


Figure 4.29 Variable distribution of SRV porosity in the best selected models of GA - Stage 1 (three phase FMM)

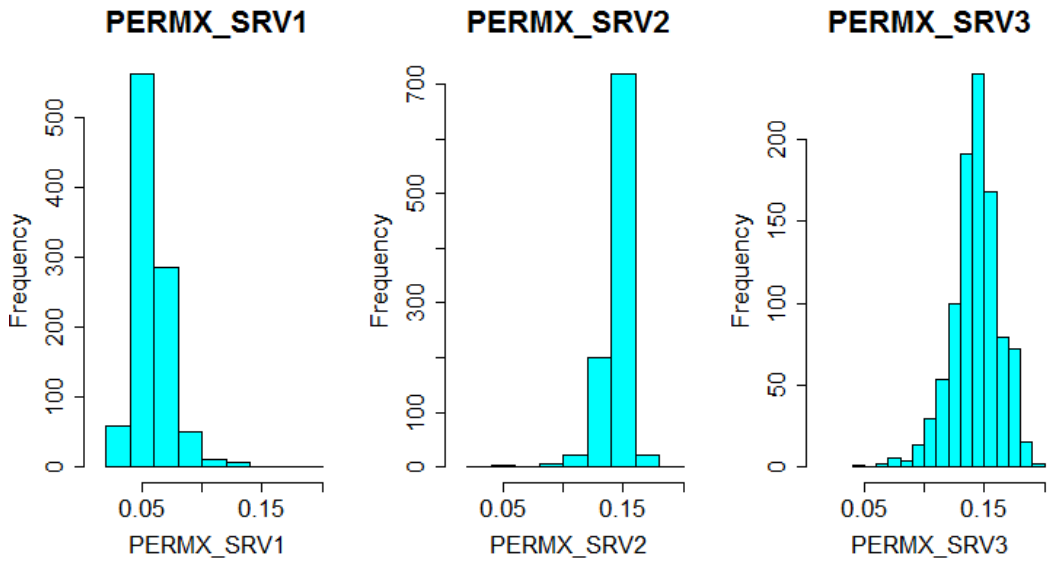


Figure 4.30 Variable distribution of SRV permeability in the best selected models of GA - Stage 1 (three phase FMM)

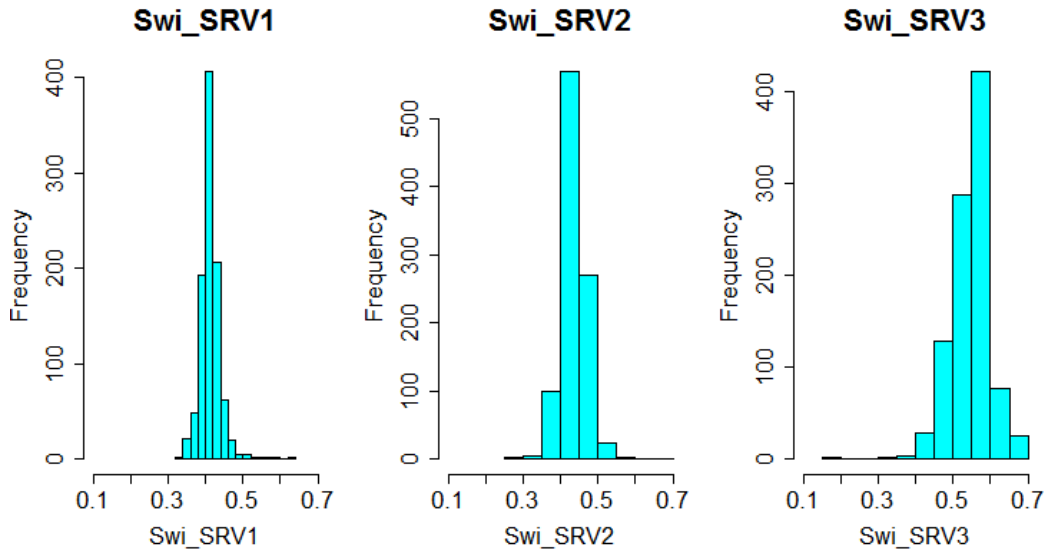


Figure 4.31 Variable distribution of SRV initial water saturation in the best selected models of GA - Stage 1 (three phase FMM)

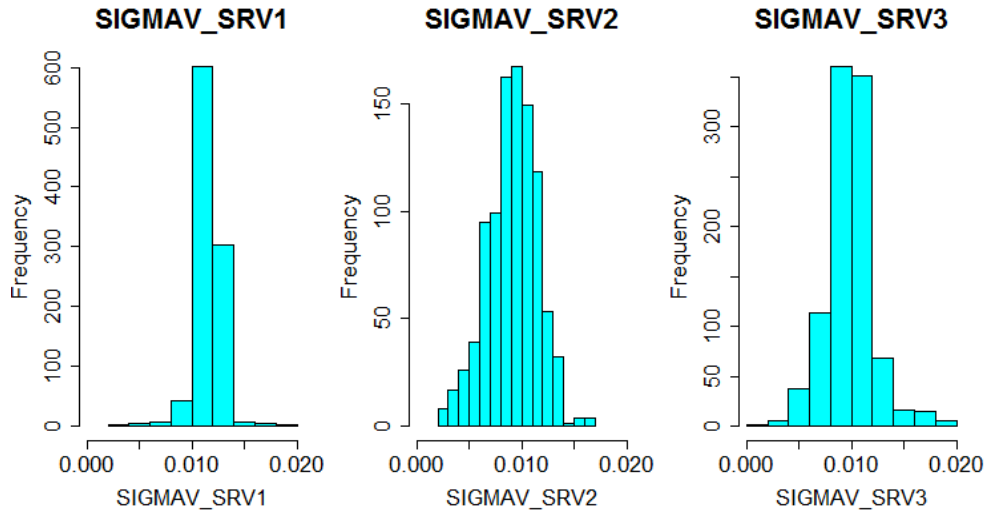


Figure 4.32 Variable distribution of SRV shape factor in the best selected models of GA - Stage 1 (three phase FMM)

In the next GA stage, the variables of stage 1 are kept with updated ranges based on best models selected previously and the previously discarded variables are also included. **Fig. 4.33** shows the new sensitivity plot. It can be observed that this time, more uniformity is seen in terms of variable importance.

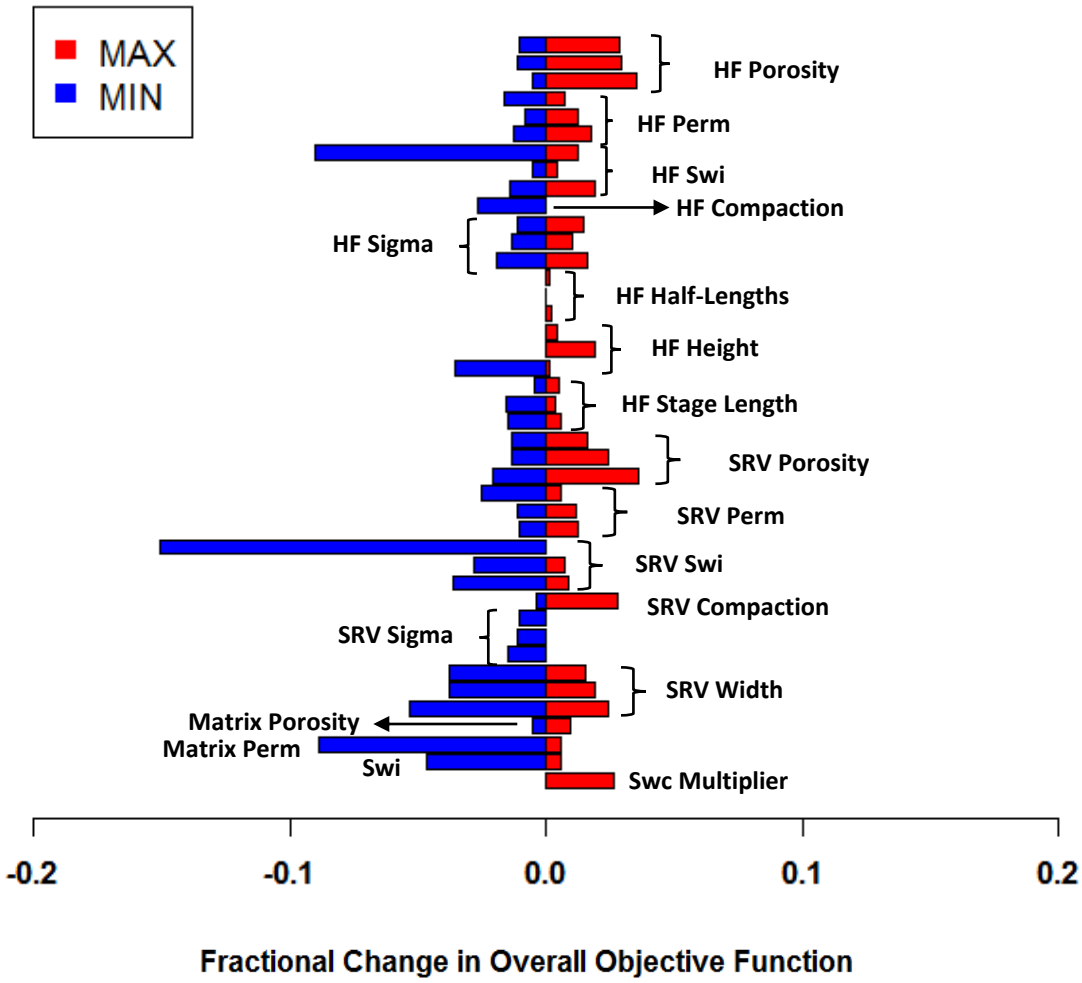


Figure 4.33 Sensitivity analysis at the beginning of Stage 2 (three phase FMM)

Fig. 4.34 shows the results of GA in stage 2. As can be observed from this figure, after multiple generations, improvement in objective error function reduces. Also, since variables in this GA operation show large shrinkage in their ranges from generation 1 to generation 12 **(Figs. 4.35 to 4.42)**, GA was stopped at this point and a collection of best models was selected **(Fig. 4.34).** These best models are chosen to derive new variable ranges of the variables included for this GA stage. **Figs. 4.43 to 4.50** show the variable ranges in the best models selected at the end of this GA stage. It may be observed that distributions of the variables common with previous stage have become narrower showing further reduction in uncertainty.

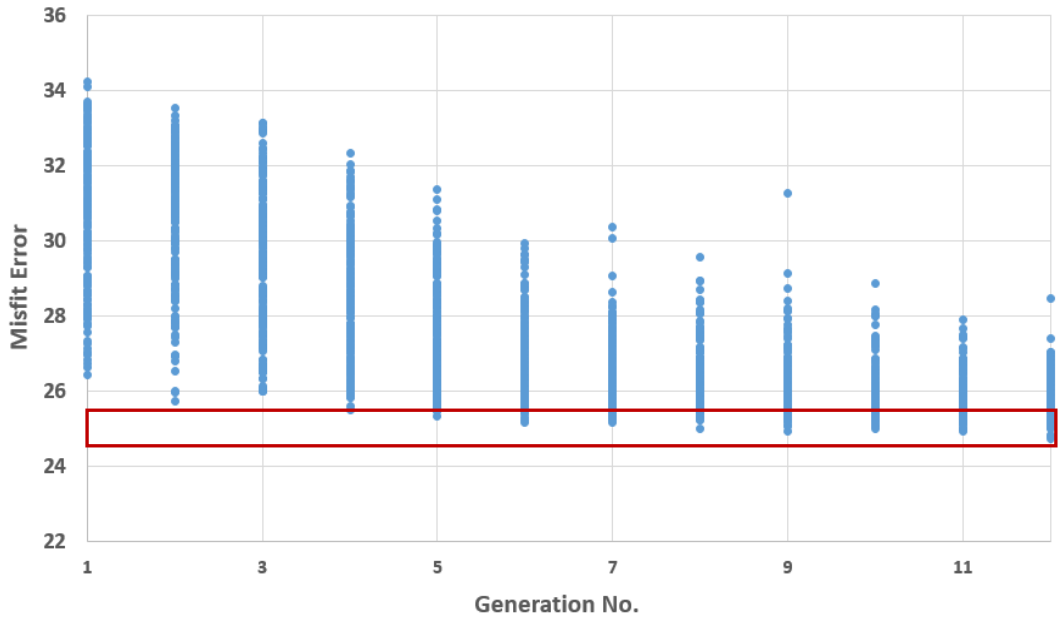


Figure 4.34 GA results for Stage 2 (three phase FMM)

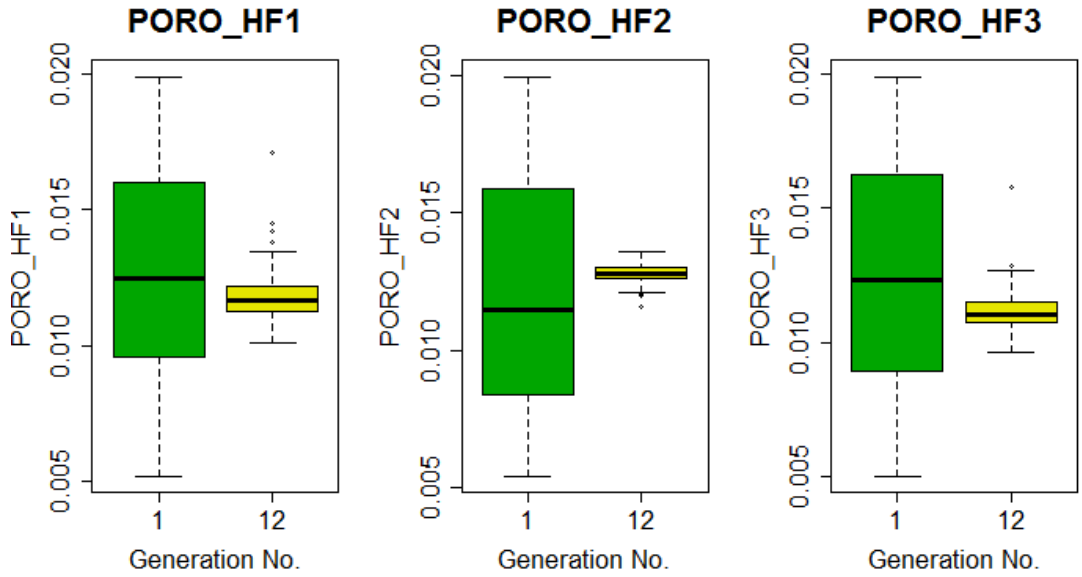


Figure 4.35 Uncertainty reduction in hydraulic fracture porosity during GA - Stage 2 (three phase FMM)

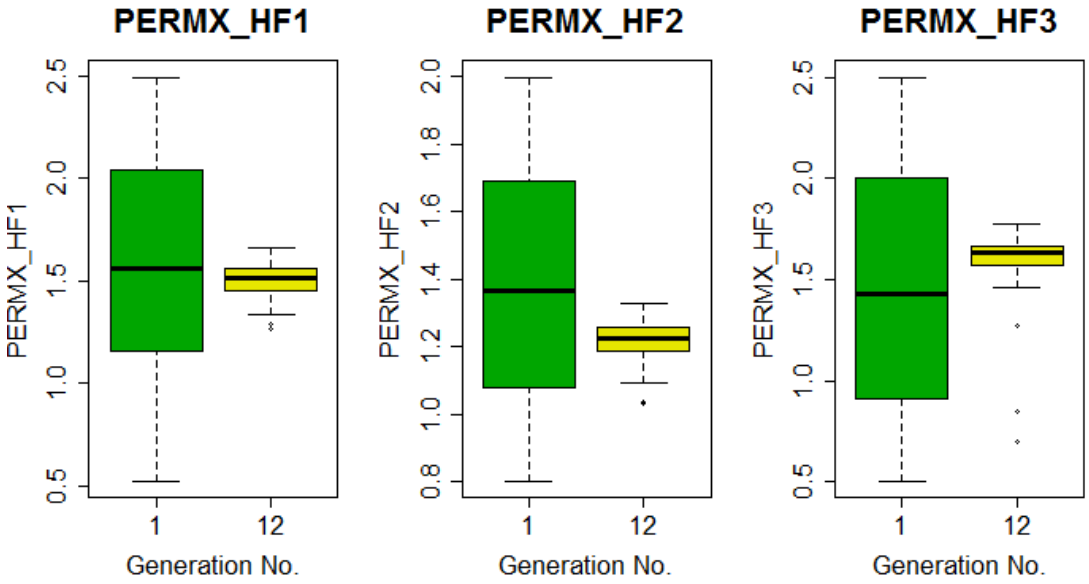


Figure 4.36 Uncertainty reduction in hydraulic fracture permeability during GA - Stage 2 (three phase FMM)

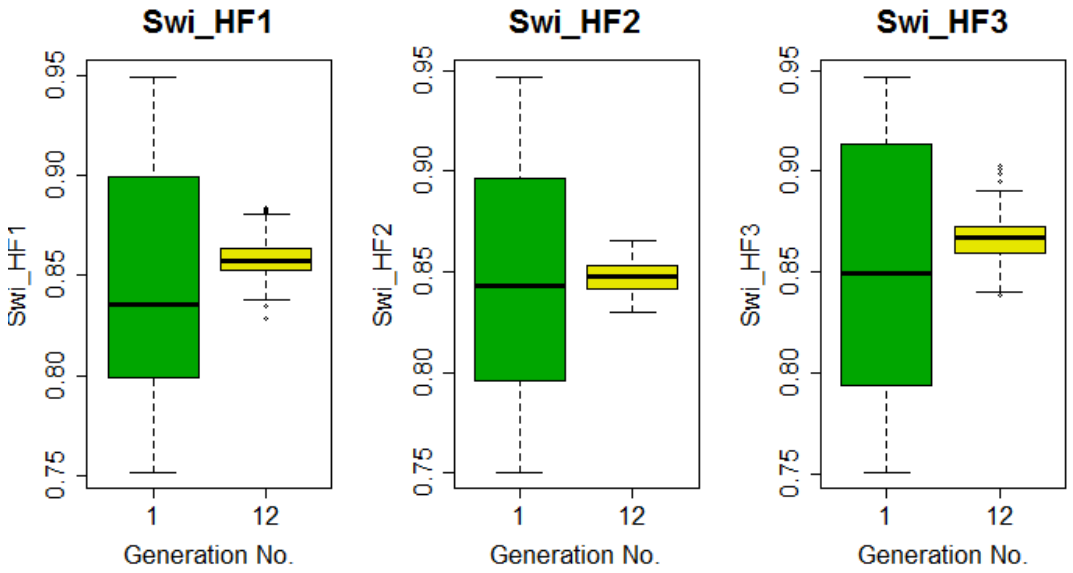


Figure 4.37 Uncertainty reduction in hydraulic fracture initial water saturation during GA - Stage 2 (three phase FMM)

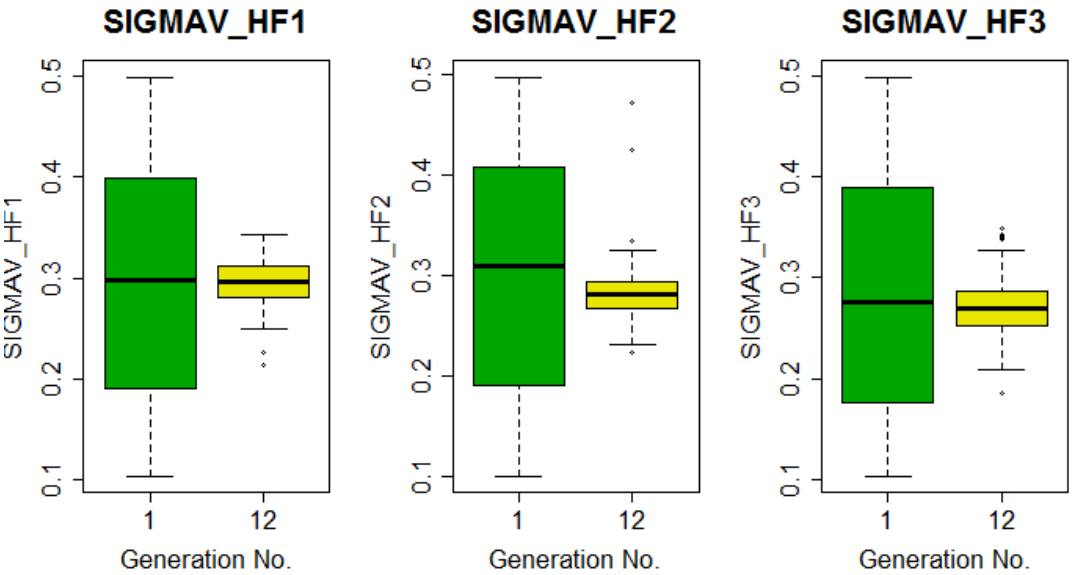


Figure 4.38 Uncertainty reduction in hydraulic fracture shape factor during GA - Stage 2 (three phase FMM)

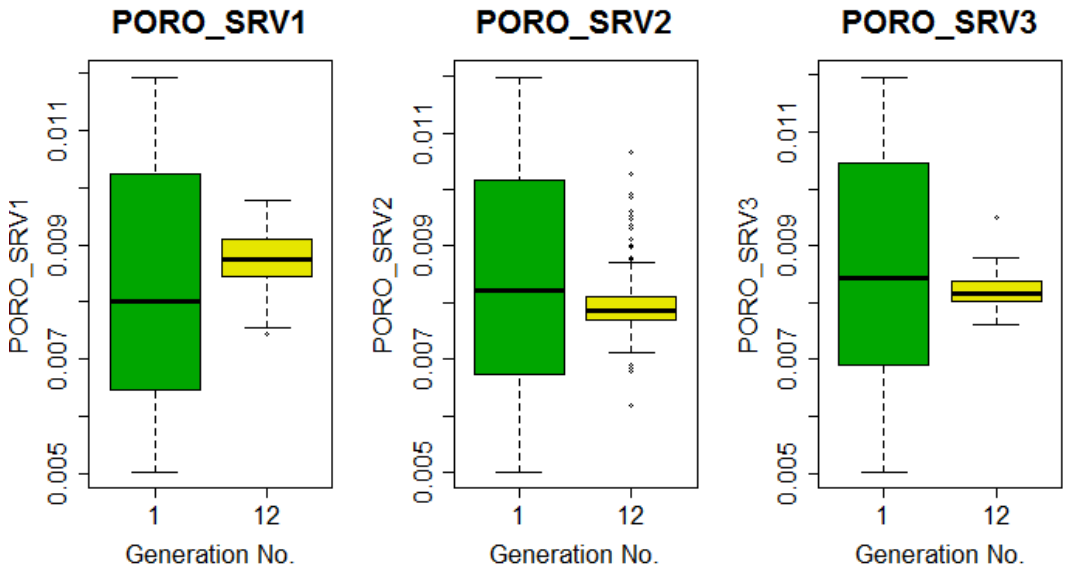


Figure 4.39 Uncertainty reduction in SRV porosity during GA - Stage 2 (three phase FMM)

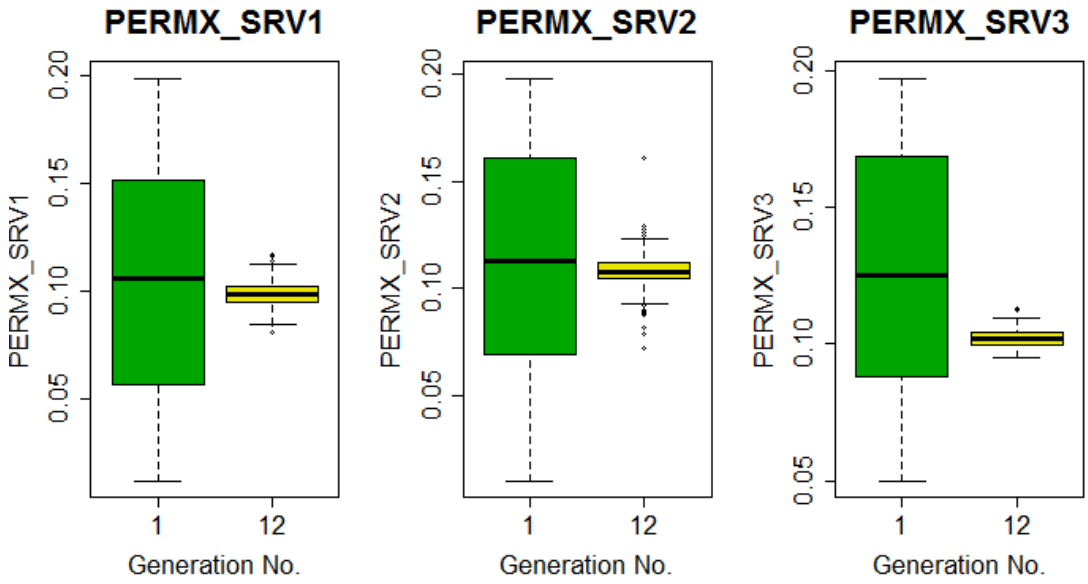


Figure 4.40 Uncertainty reduction in SRV permeability during GA - Stage 2 (three phase FMM)

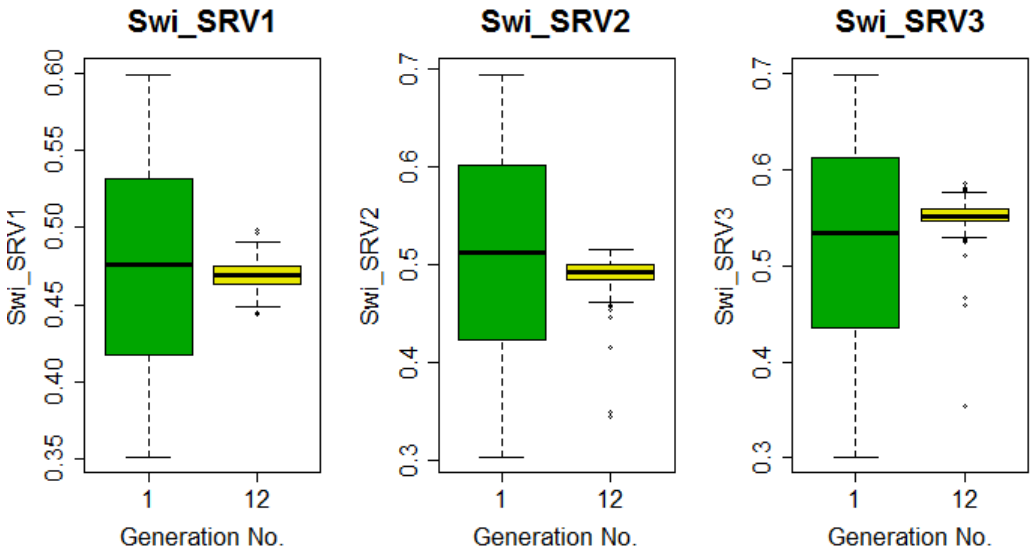


Figure 4.41 Uncertainty reduction in SRV initial water saturation during GA - Stage 2 (three phase FMM)

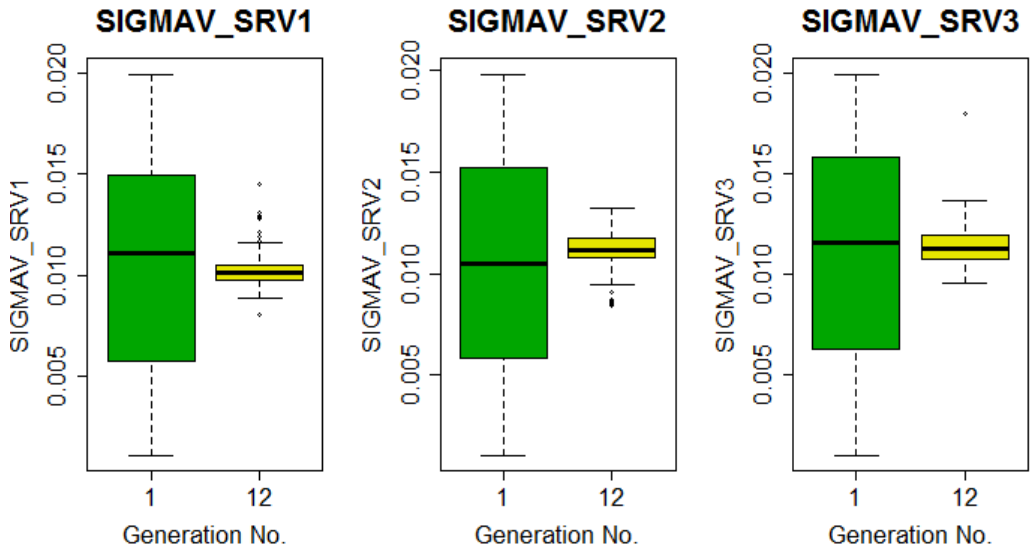


Figure 4.42 Uncertainty reduction in SRV shape factor during GA - Stage 2 (three phase FMM)

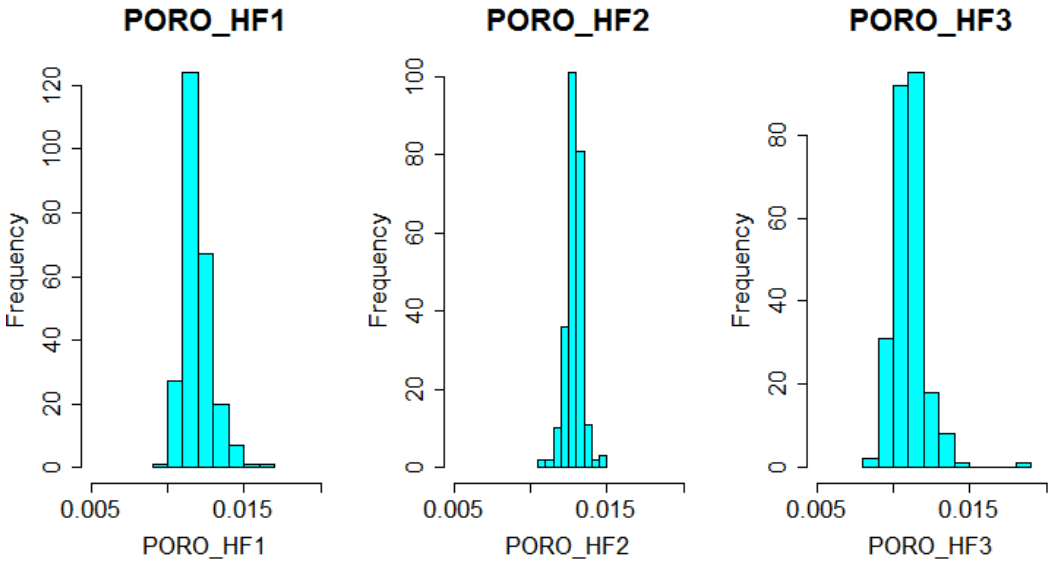


Figure 4.43 Variable distribution of hydraulic fracture porosity in the best selected models of GA - Stage 2 (three phase FMM)

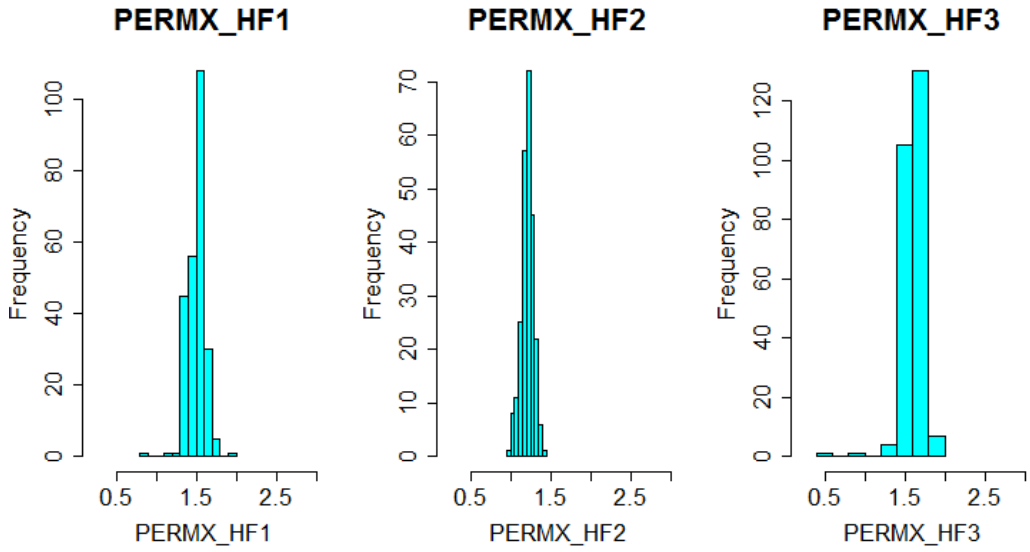


Figure 4.44 Variable distribution of hydraulic fracture permeability in the best selected models of GA - Stage 2 (three phase FMM)

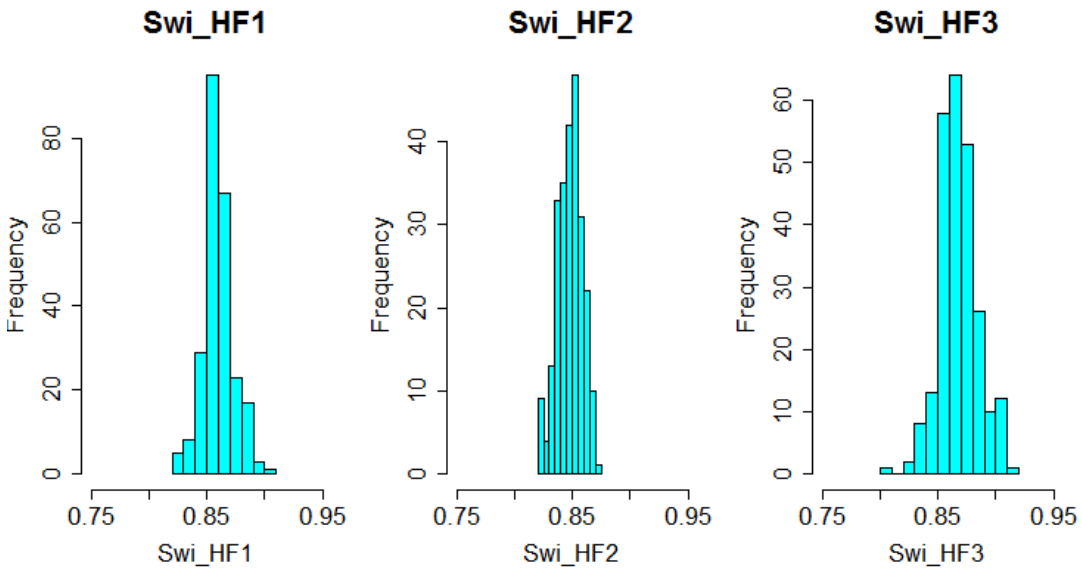


Figure 4.45 Variable distribution of hydraulic fracture initial water saturation in the best selected models of GA - Stage 2 (three phase FMM)

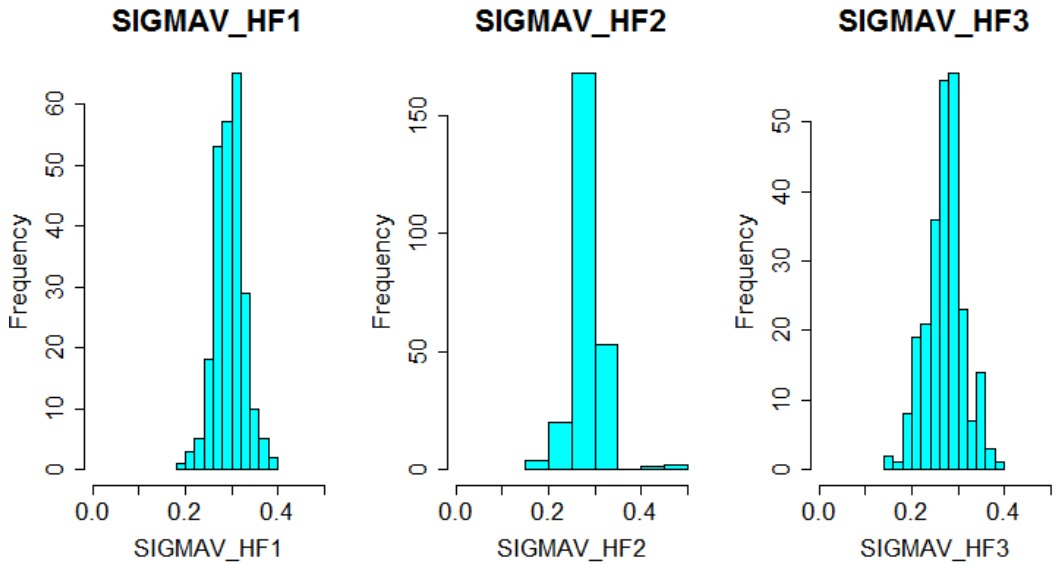


Figure 4.46 Variable distribution of hydraulic fracture shape factor in the best selected models of GA - Stage 2 (three phase FMM)

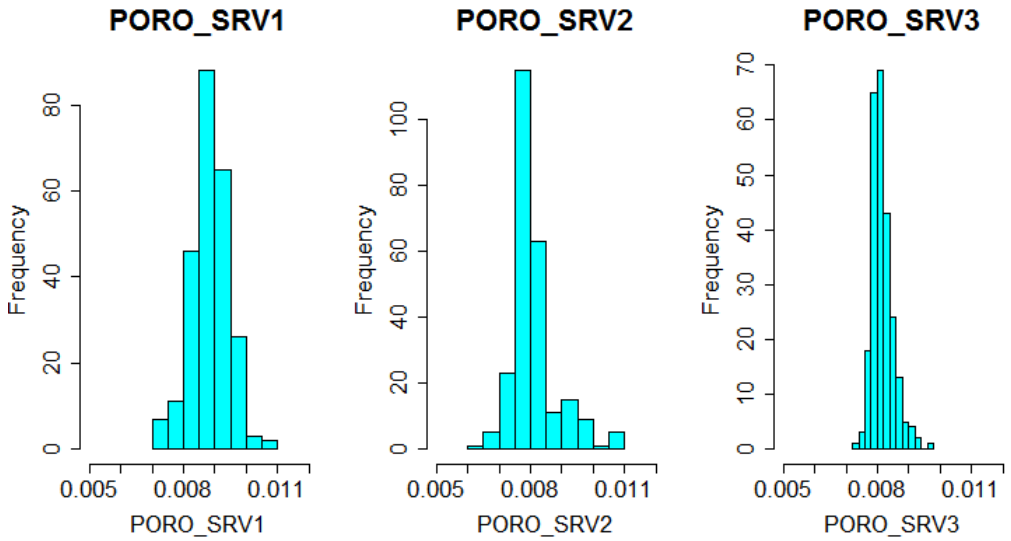


Figure 4.47 Variable distribution of SRV porosity in the best selected models of GA - Stage 2 (three phase FMM)

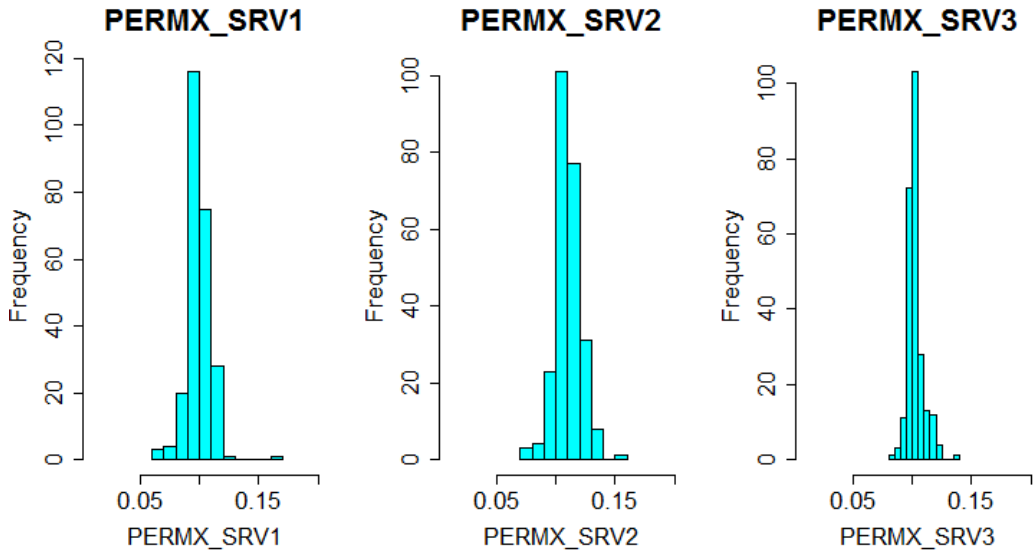


Figure 4.48 Variable distribution of SRV permeability in the best selected models of GA - Stage 2 (three phase FMM)

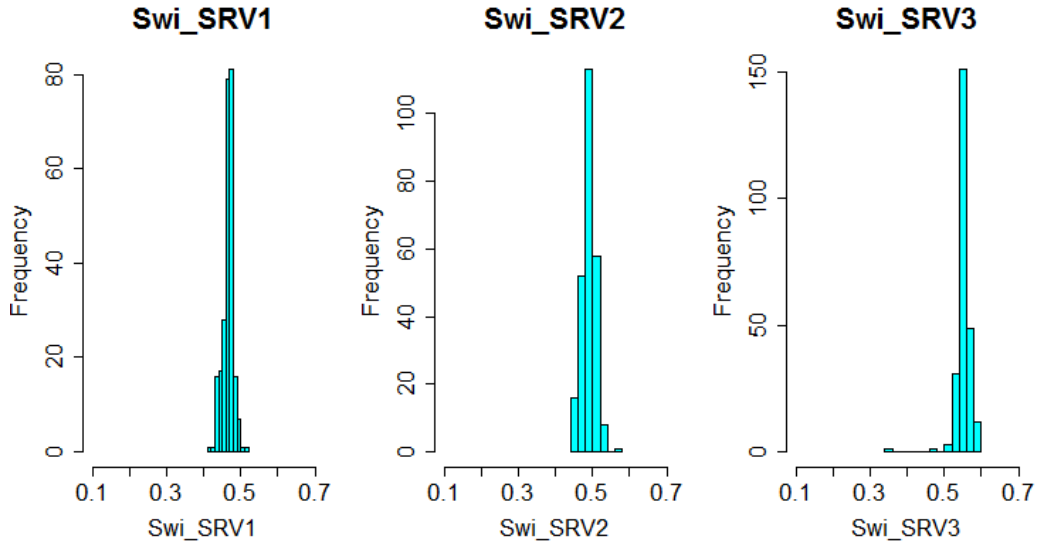


Figure 4.49 Variable distribution of SRV initial water saturation in the best selected models of GA - Stage 2 (three phase FMM)

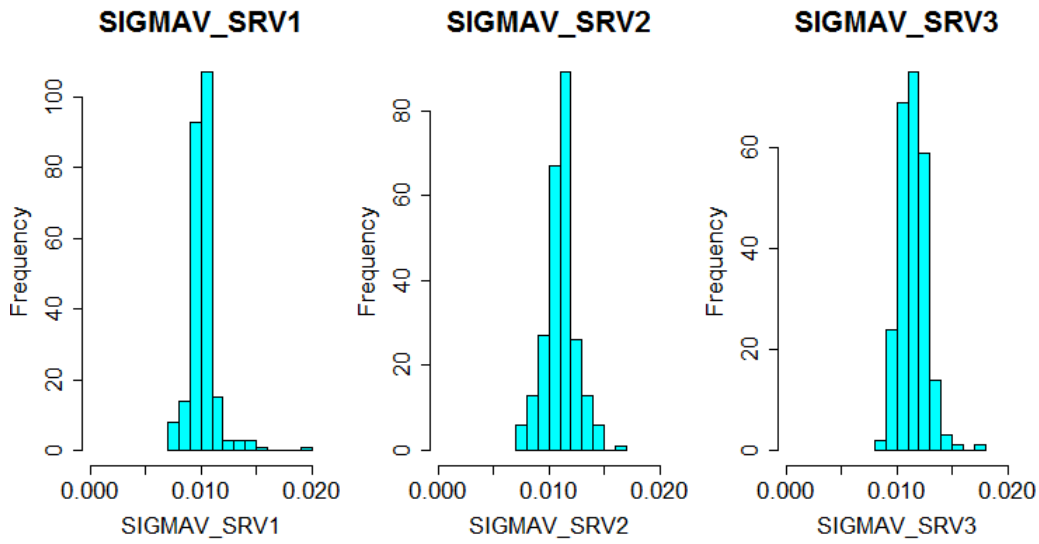


Figure 4.50 Variable distribution of SRV shape factor in the best selected models of GA - Stage 2 (three phase FMM)

In the next GA stage, the variables of the previous stage are kept with updated ranges based on best models selected previously. **Fig. 4.51** shows the new sensitivity plot. It can be observed that this time, some of the variables are not making big impact due to shrinkage of their ranges in the previous GA stages. However, all the variables are included in this GA stage.

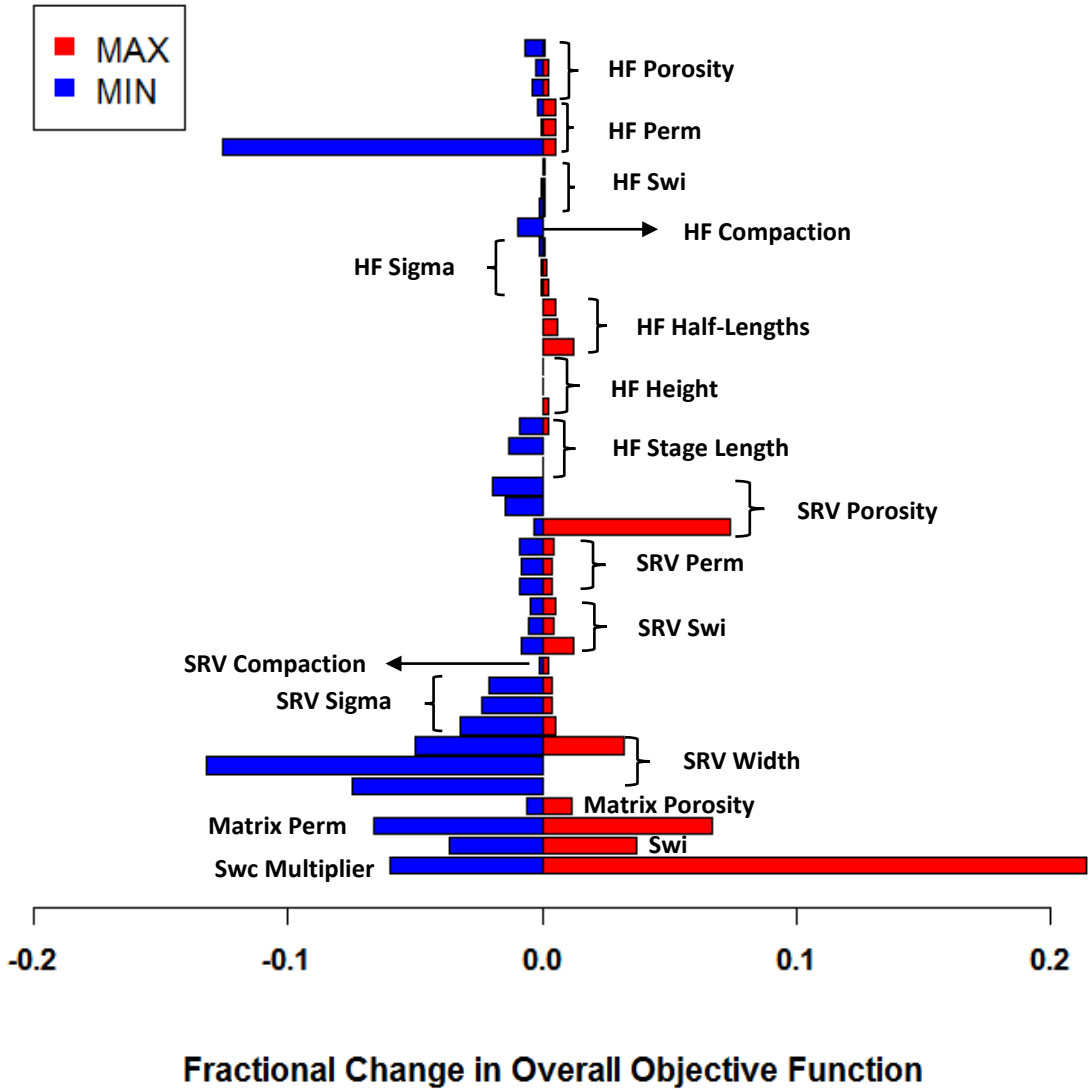


Figure 4.51 Sensitivity analysis at the beginning of Stage 3 (three phase FMM)

Fig. 4.52 shows the results of GA in stage 3. As can be observed from this figure, after multiple generations, improvement in objective error function reduces. Also, since variables in this GA operation show large shrinkage in their ranges from generation 1 to generation 12 **(Figs. 4.53 to 4.60)**, GA was stopped at this point and a collection of best models was selected **(Fig. 4.52).** These best models are chosen to derive new variable ranges of the variables included for this GA stage. **Figs. 4.61 to 4.67** show the variable ranges in the best models selected at the end of this GA stage. It may be observed that distributions of the variables common with previous stage have become narrower showing further reduction in uncertainty.

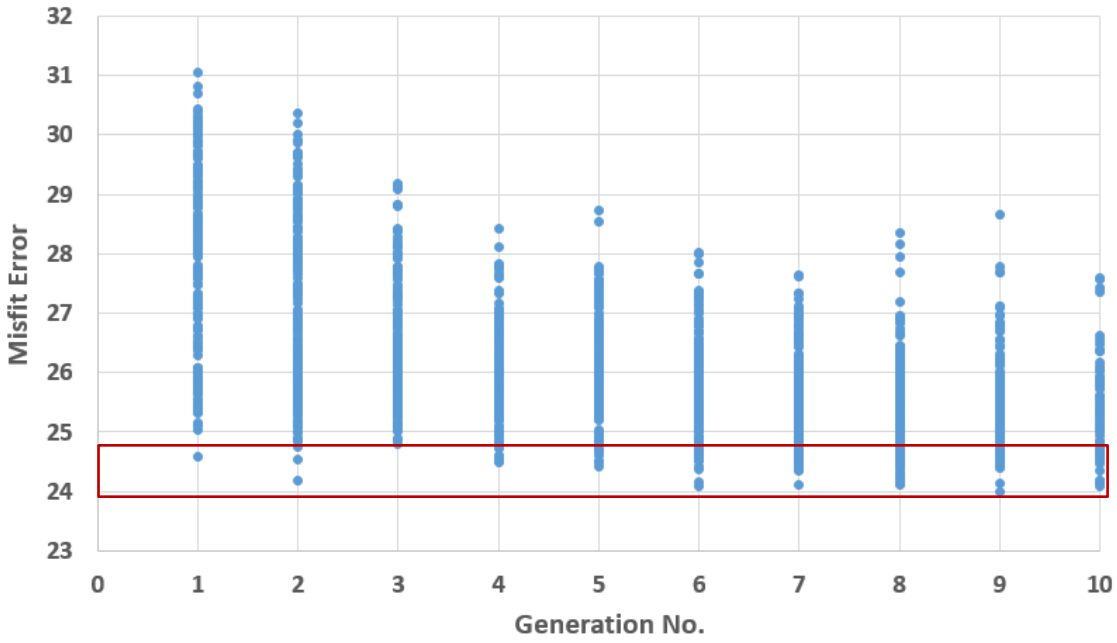


Figure 4.52 GA results for Stage 3 (three phase FMM)

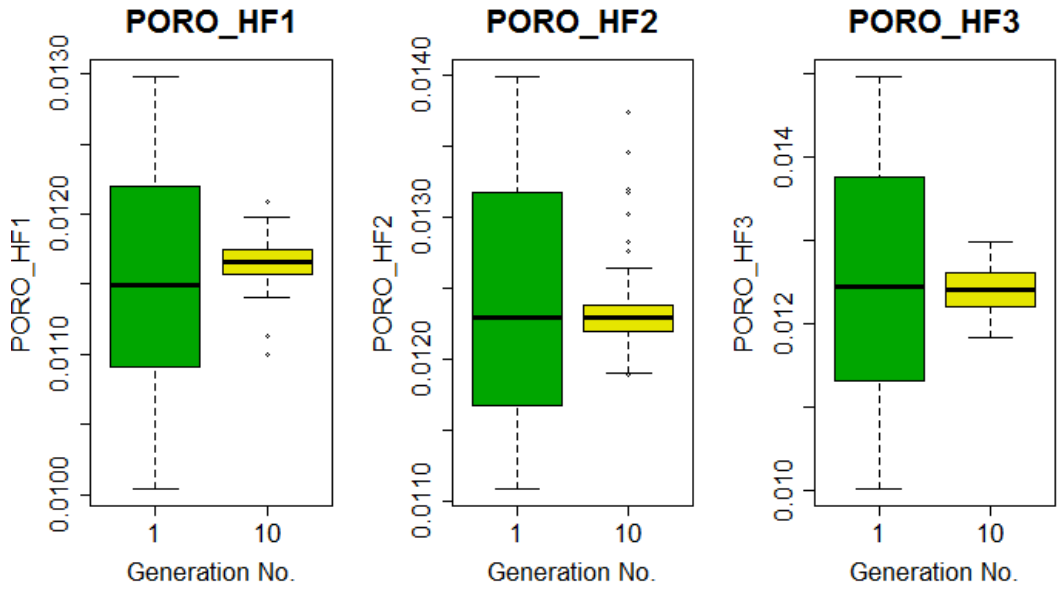


Figure 4.53 Uncertainty reduction in hydraulic fracture porosity during GA - Stage 3 (three phase FMM)

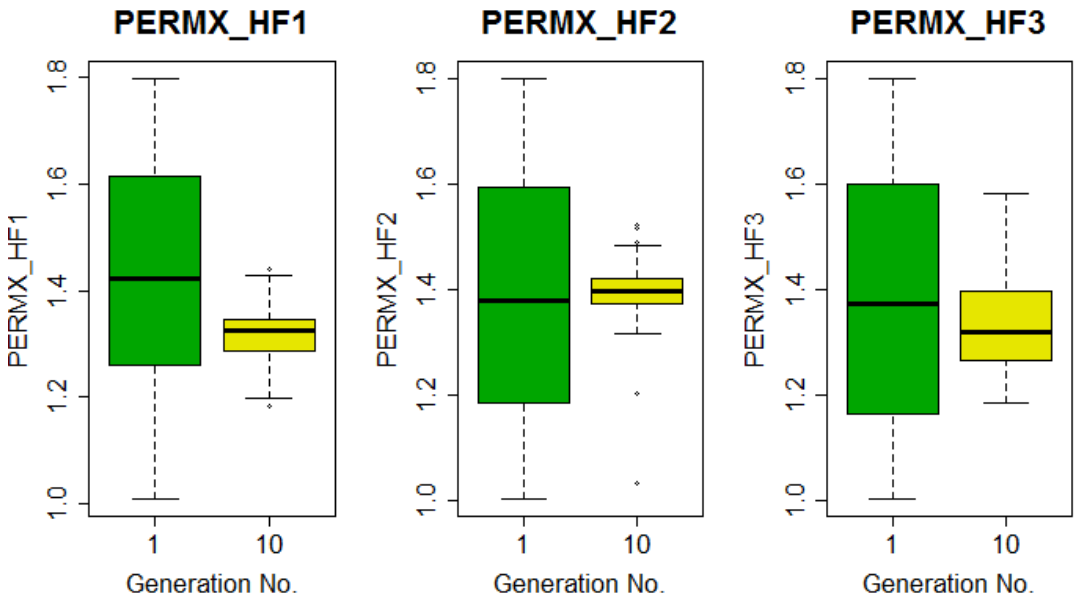


Figure 4.54 Uncertainty reduction in hydraulic fracture permeability during GA - Stage 3 (three phase FMM)

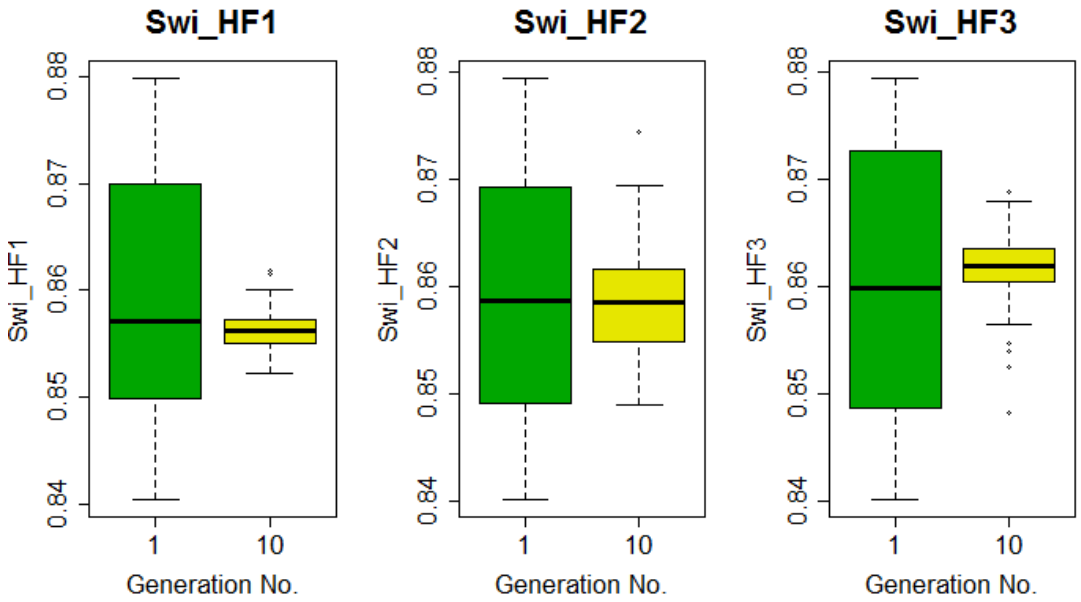


Figure 4.55 Uncertainty reduction in hydraulic fracture initial water saturation during GA - Stage 3 (three phase FMM)

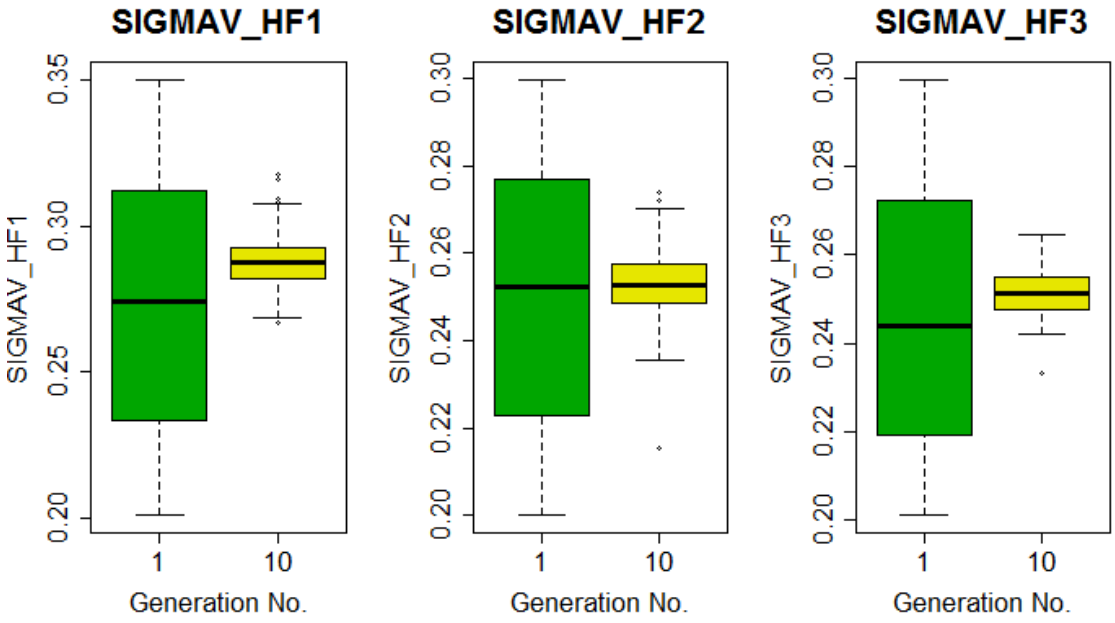


Figure 4.56 Uncertainty reduction in hydraulic fracture shape factor during GA - Stage 3 (three phase FMM)

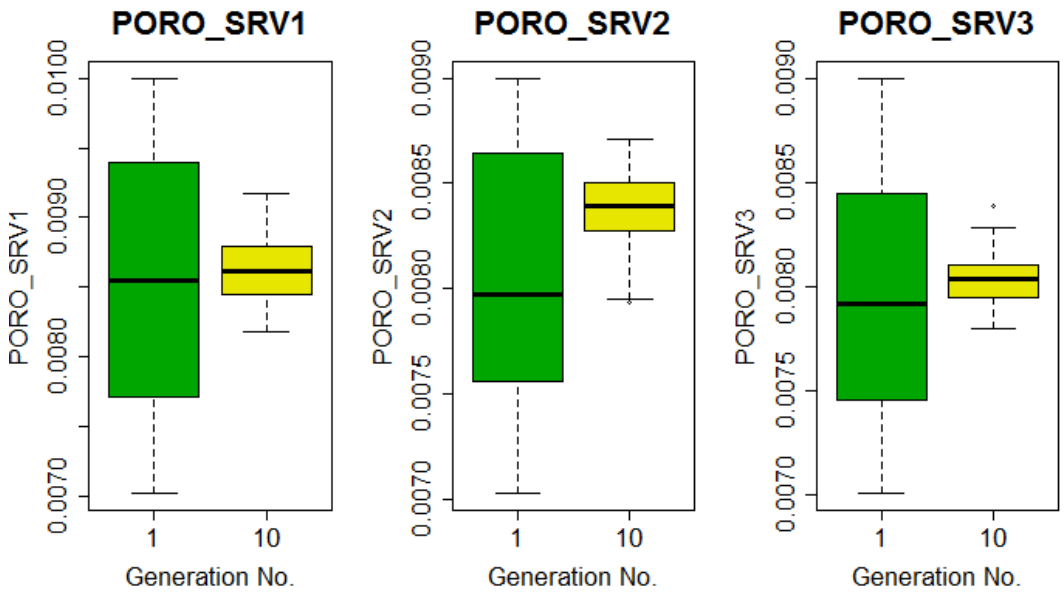


Figure 4.57 Uncertainty reduction in SRV porosity during GA - Stage 3 (three phase FMM)

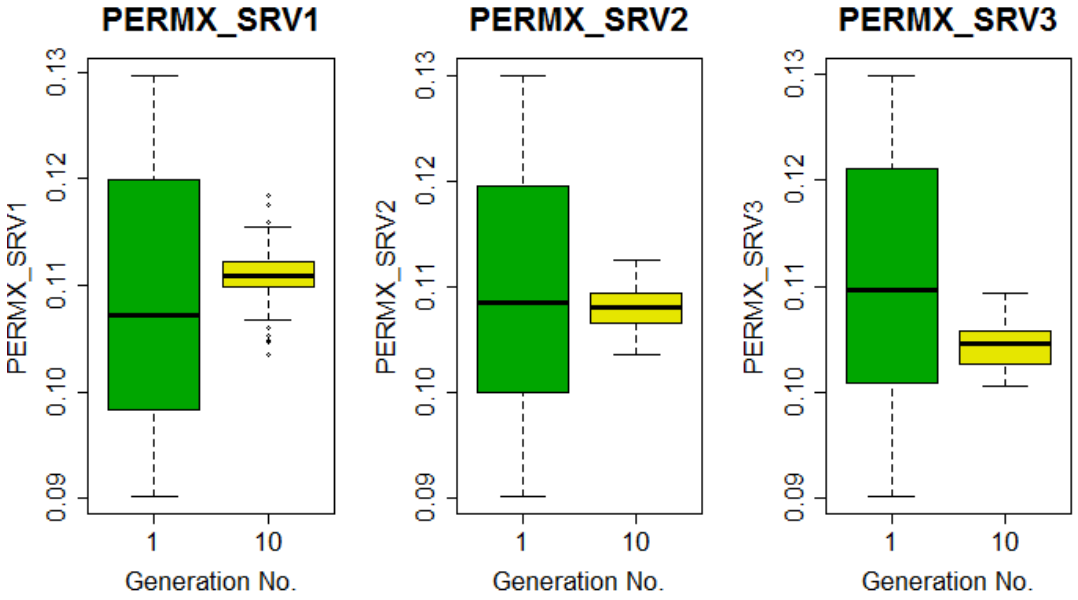


Figure 4.58 Uncertainty reduction in SRV permeability during GA - Stage 3 (three phase FMM)

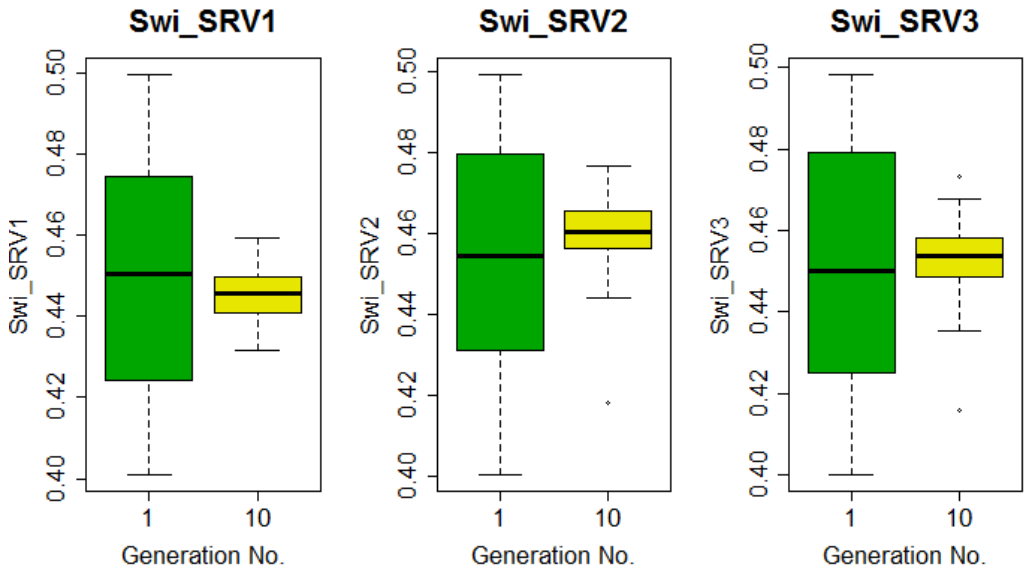


Figure 4.59 Uncertainty reduction in SRV initial water saturation during GA - Stage 3 (three phase FMM)

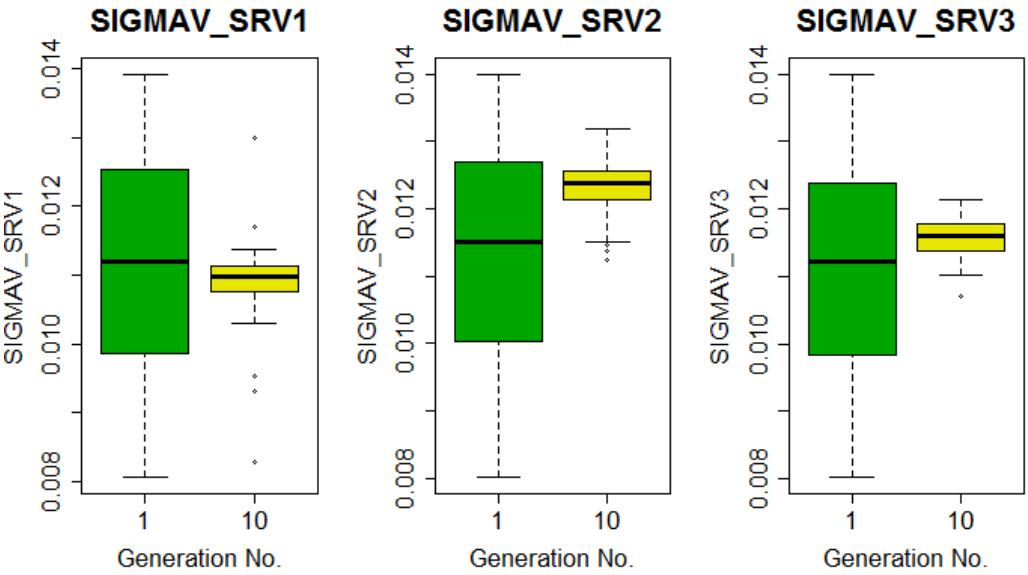


Figure 4.60 Uncertainty reduction in SRV shape factor during GA - Stage 3 (three phase FMM)

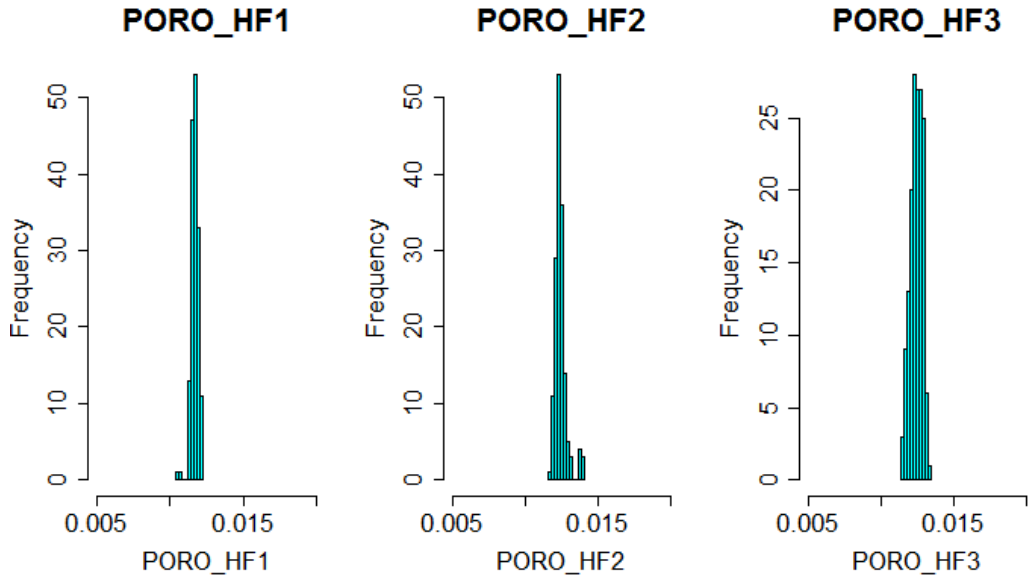


Figure 4.61 Variable distribution of hydraulic fracture porosity in the best selected models of GA - Stage 3 (three phase FMM)

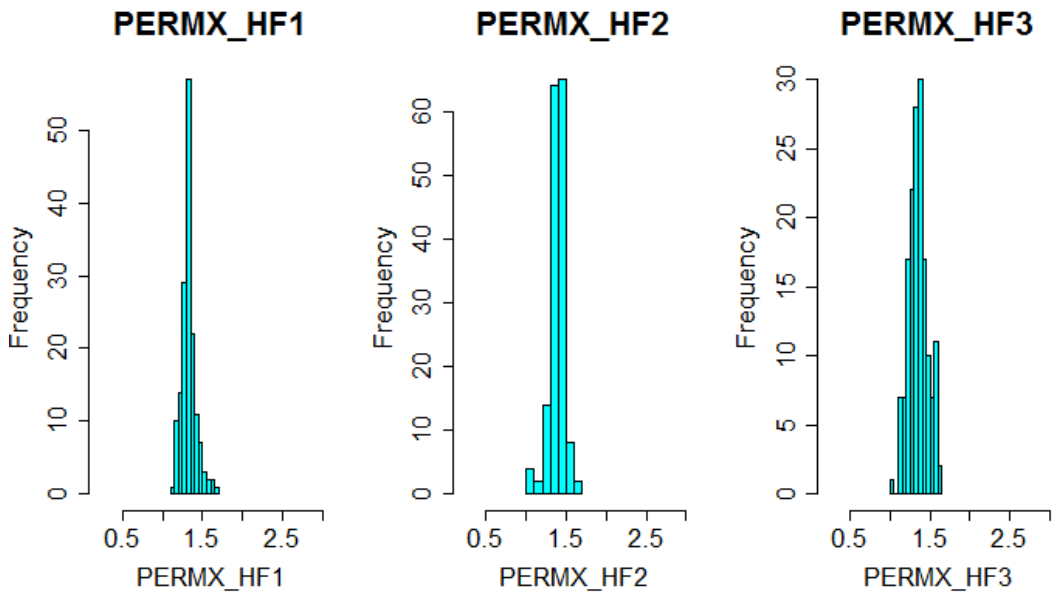


Figure 4.62 Variable distribution of hydraulic fracture permeability in the best selected models of GA - Stage 3 (three phase FMM)

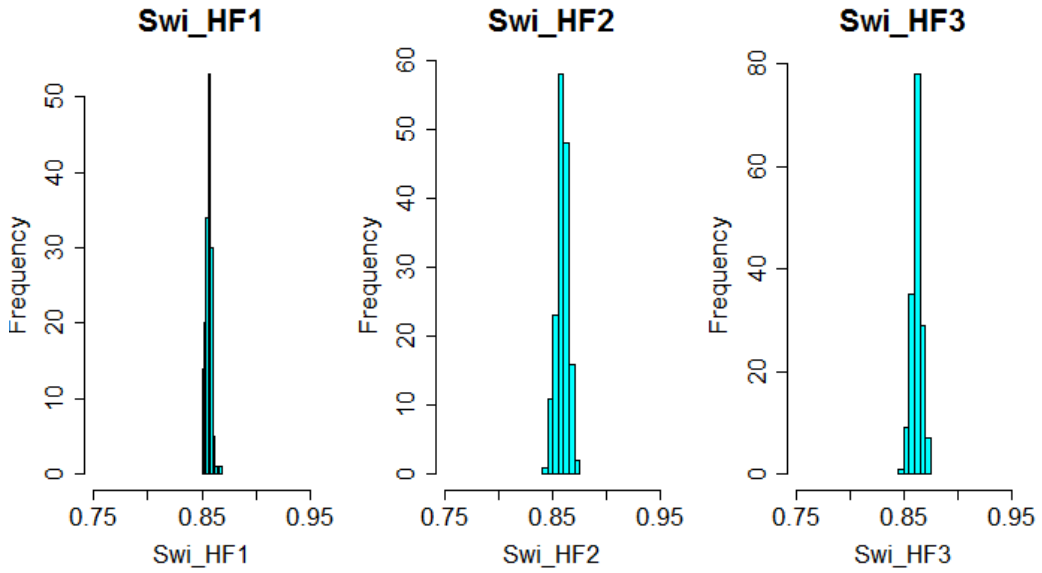


Figure 4.63 Variable distribution of hydraulic fracture initial water saturation in the best selected models of GA - Stage 3 (three phase FMM)

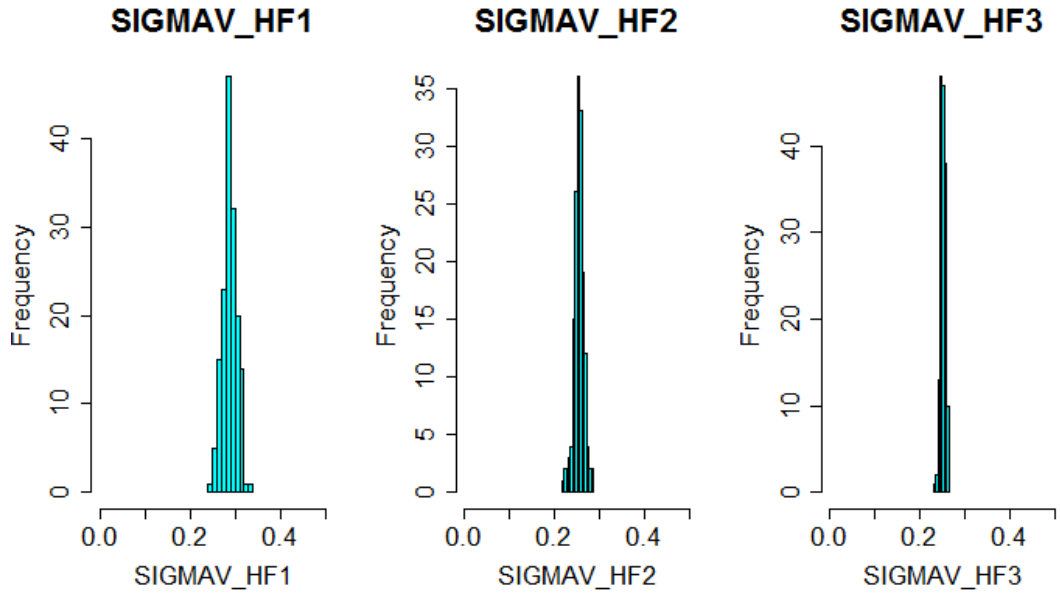


Figure 4.64 Variable distribution of hydraulic fracture shape factor in the best selected models of GA - Stage 3 (three phase FMM)

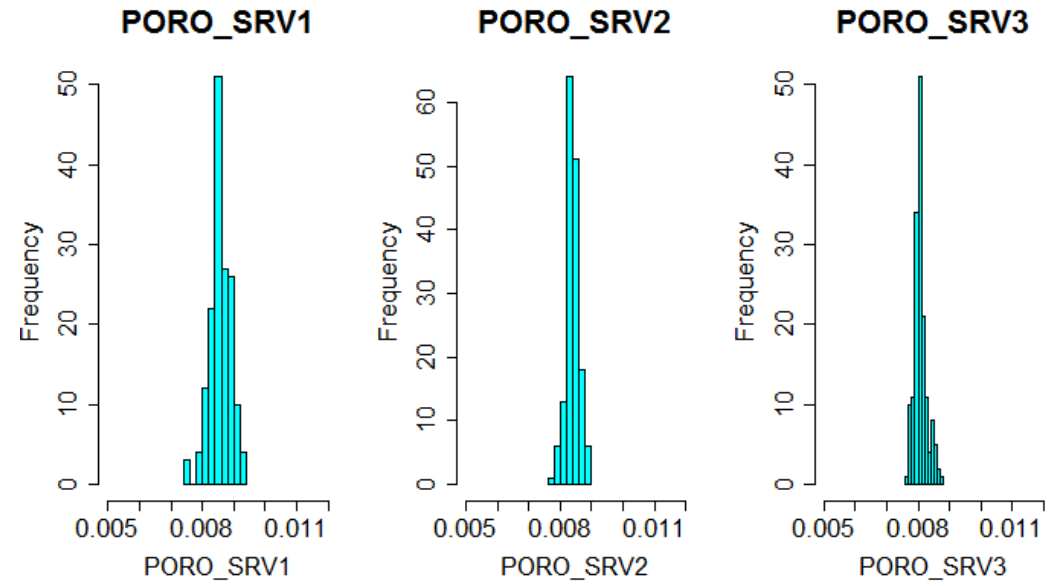


Figure 4.65 Variable distribution of SRV porosity in the best selected models of GA - Stage 3 (three phase FMM)

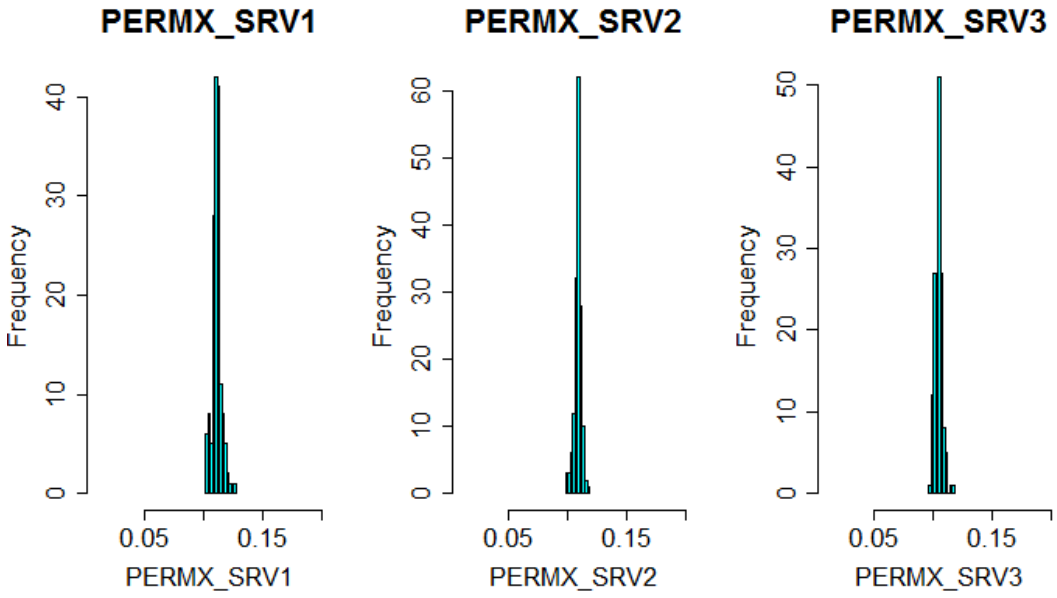


Figure 4.66 Variable distribution of SRV permeability in the best selected models of GA - Stage 3 (three phase FMM)

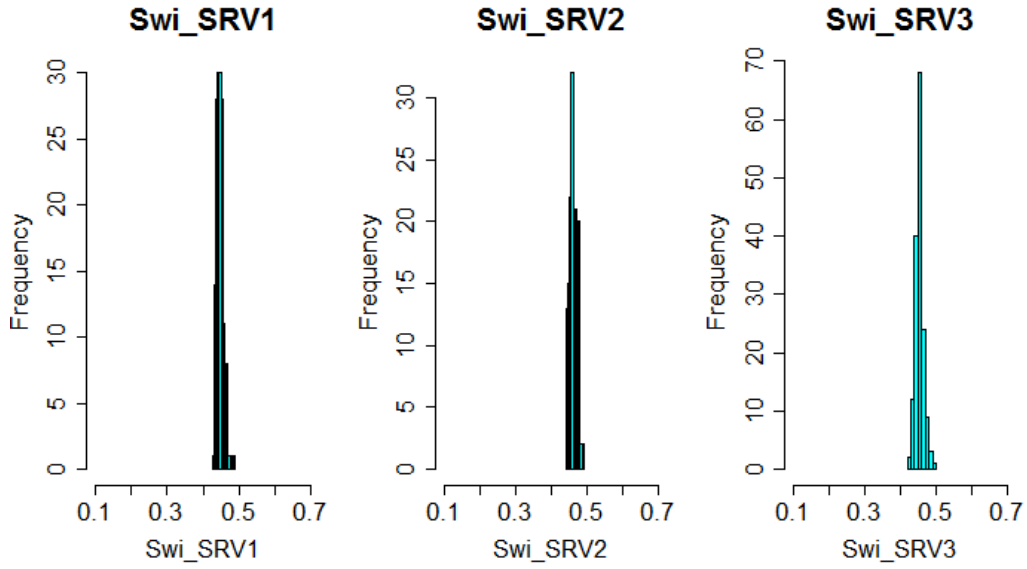


Figure 4.67 Variable distribution of SRV initial water saturation in the best selected models of GA - Stage 3 (three phase FMM)

Fig. 4.68 shows the combined plot showing all GA stages. It may be observed that there is significant improvement from one GA stage to the next one. At this point the best models are selected as mentioned previously and plotted against history data **(Figs. 4.69 to 4.74)**.

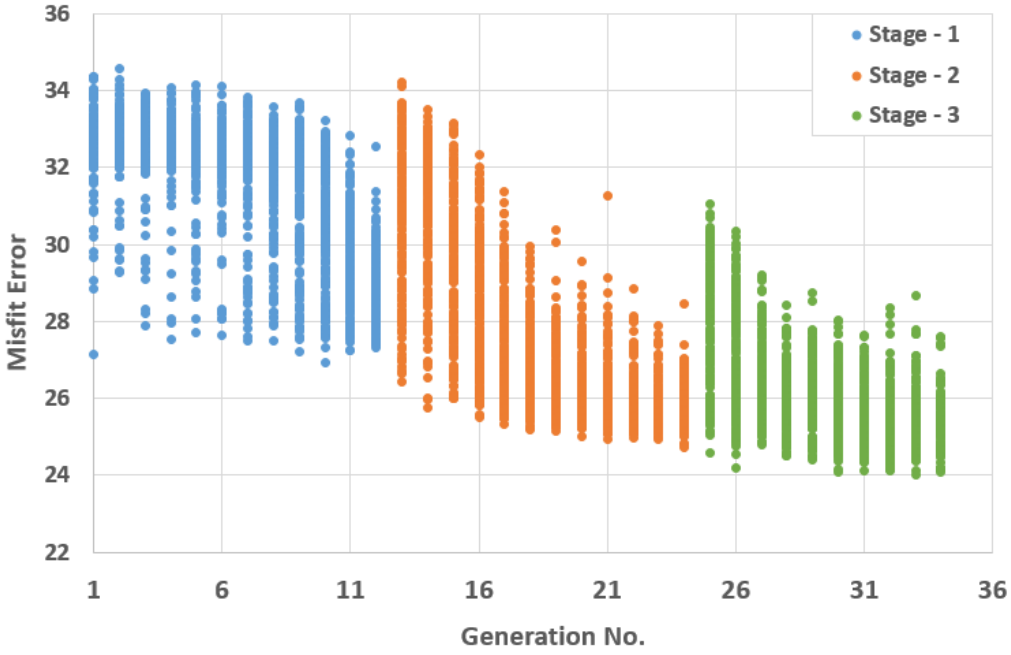
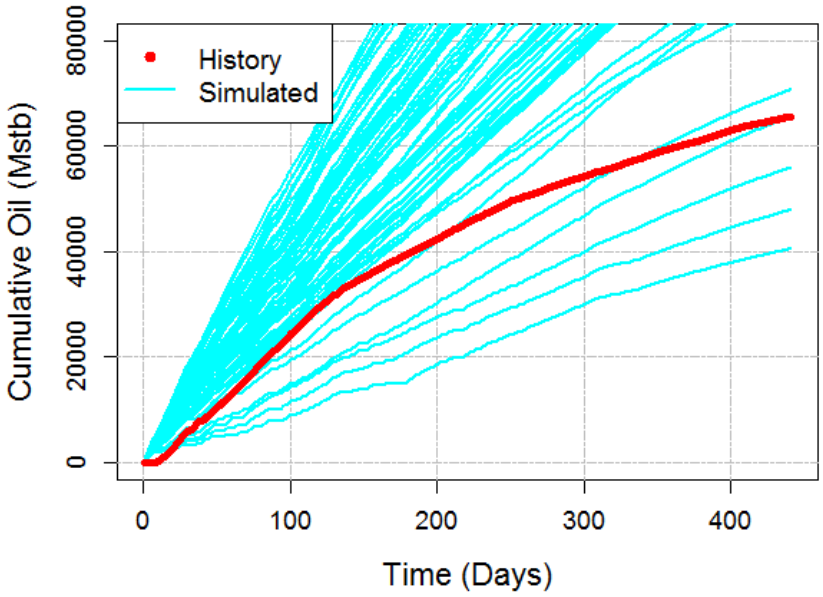
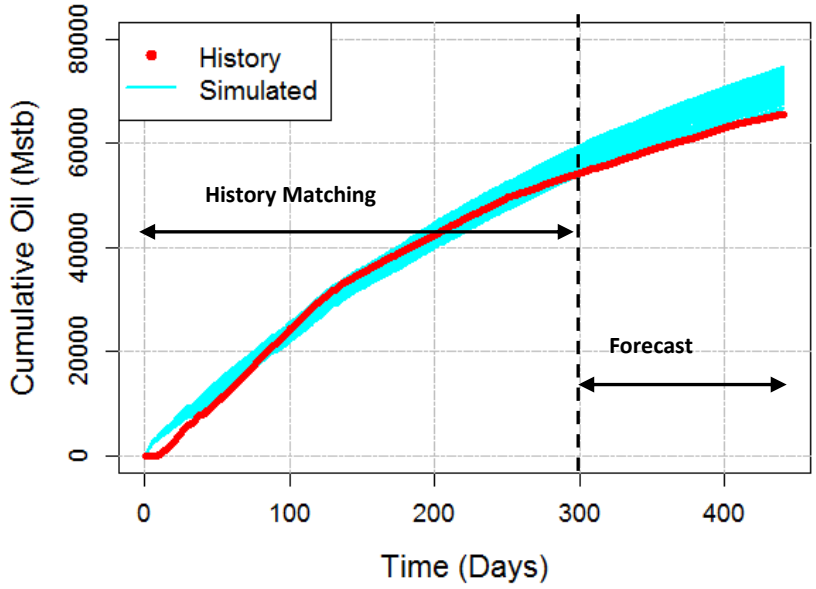


Figure 4.68 Combined GA results for all stages (three phase FMM)

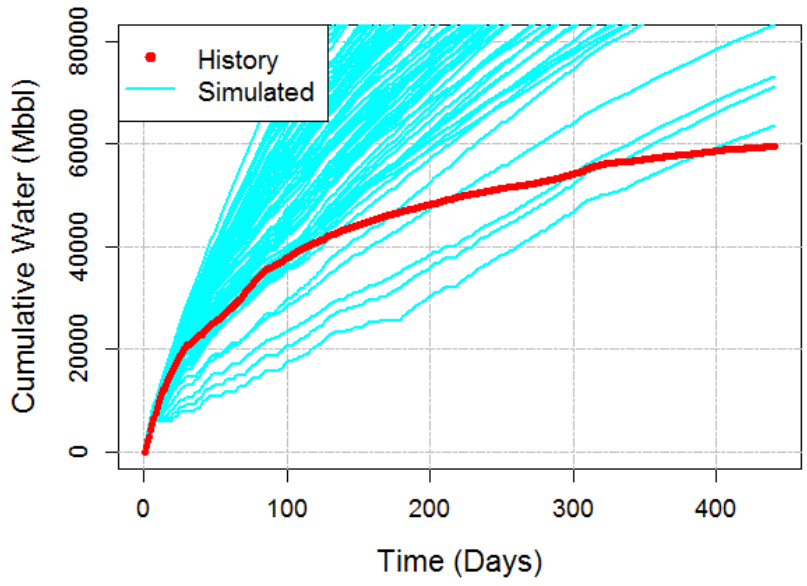


(a)

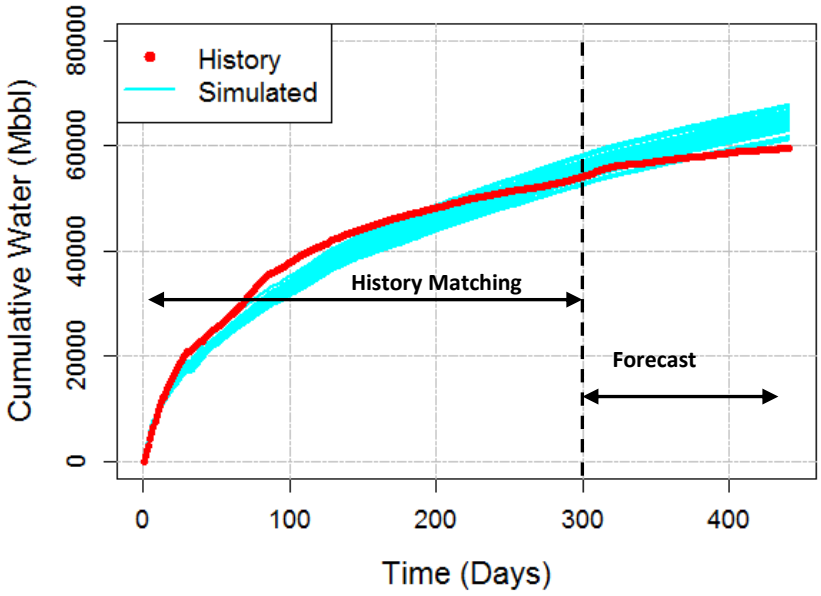


(b)

Figure 4.69 Cumulative oil history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (three phase FMM)

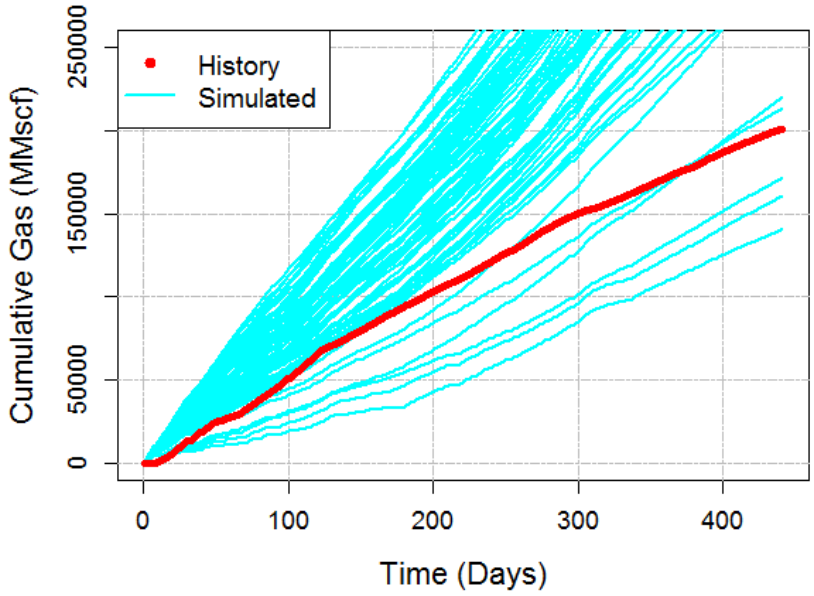


(a)

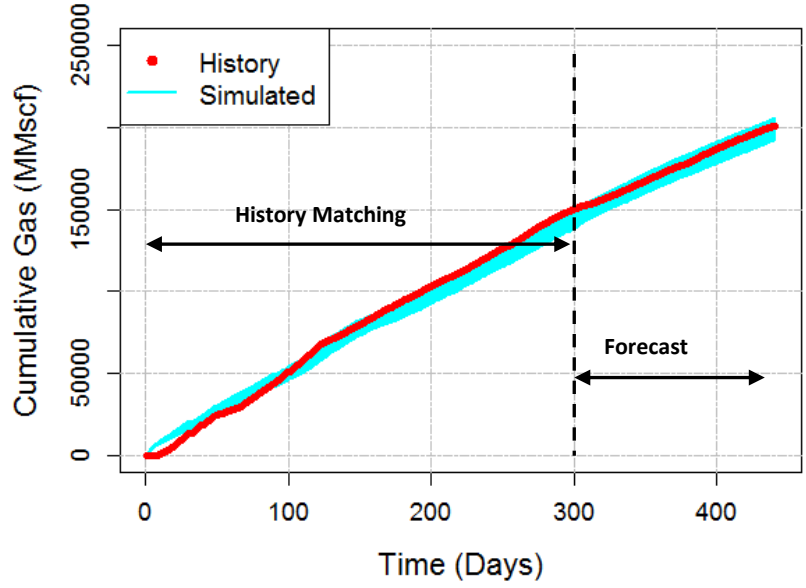


(b)

Figure 4.70 Cumulative water history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (three phase FMM)

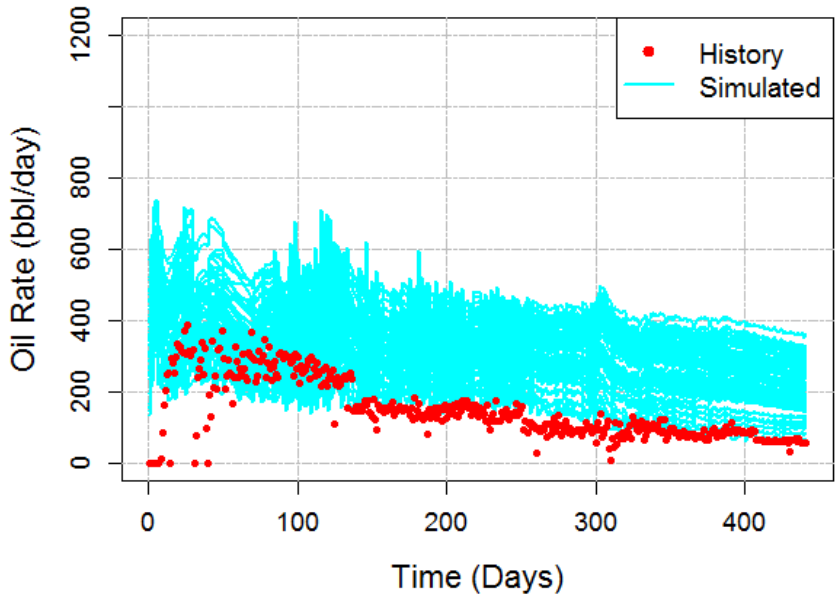


(a)

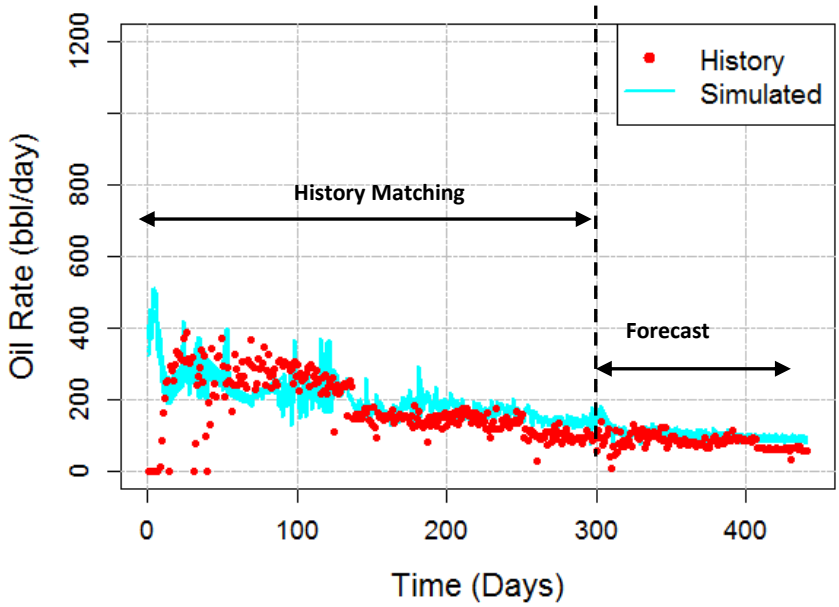


(b)

Figure 4.71 Cumulative gas history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (three phase FMM)

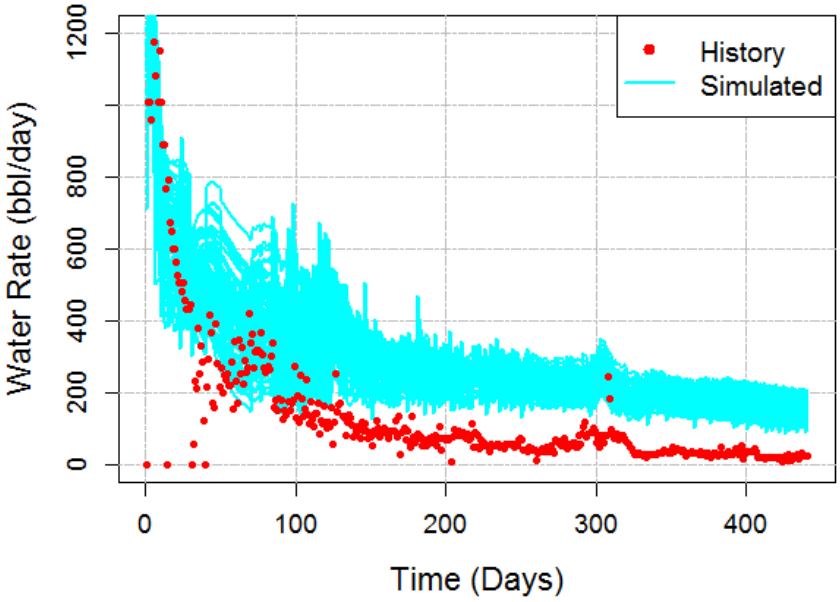


(a)

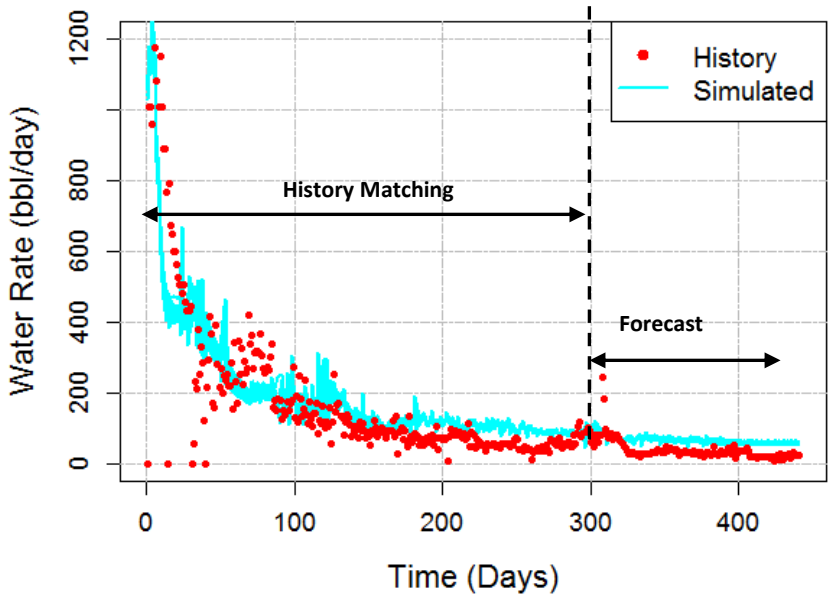


(b)

Figure 4.72 Oil rate history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (three phase FMM)

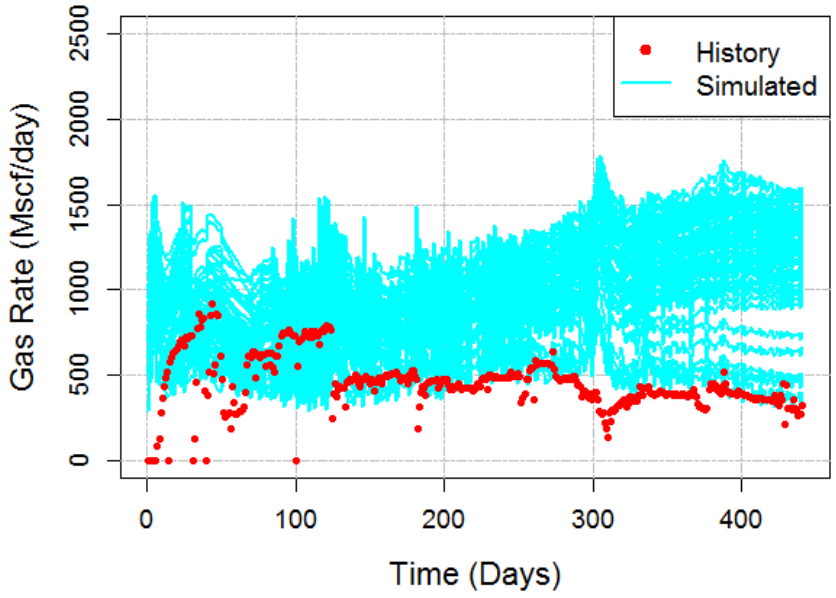


(a)

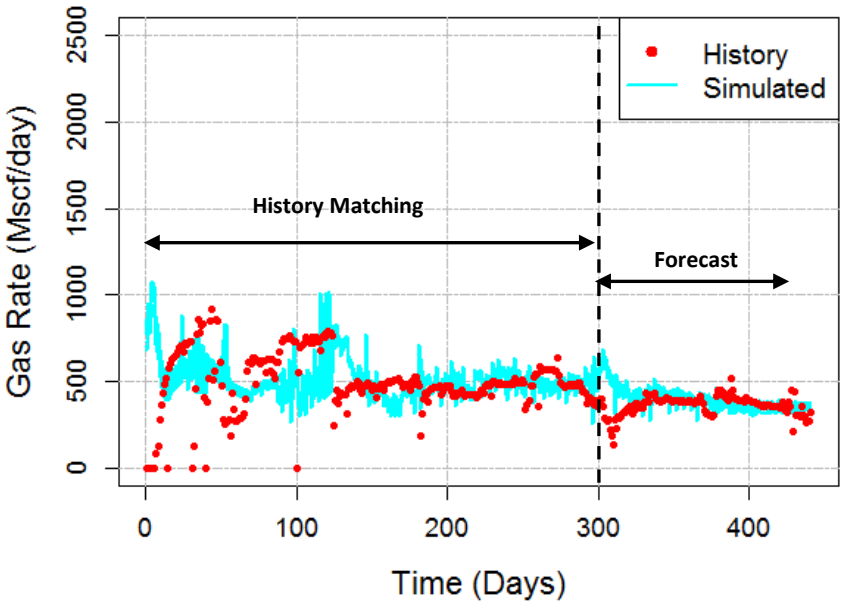


(b)

Figure 4.73 Water rate history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (three phase FMM)



(a)



(b)

Figure 4.74 Gas rate history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (three phase FMM)

4.3.2 History matching results based on GA and compositional FMM

Iino et al. (2017) presented a FMM based compositional unconventional reservoir simulator that is multiple times faster than a commercially available finite difference based reservoir simulator. Their study applied compositional FMM as a suitable candidate for history matching problem involving large number of simulations. Current dissertation study also utilizes the advantages of compositional FMM for history matching. To test accuracy of FMM relative to Eclipse, simulations have been conducted for both FMM based simulator and Eclipse for the field case model under investigation using the base values of each variable. **Figs. 4.75 to 4.80** present the comparison plots of the simulation results using compositional FMM simulator and Eclipse 300 simulator. It is clear from these figures that FMM and Eclipse are reasonably close to each other and therefore, FMM can be a good candidate for further history matching simulations due to faster simulations (Iino et. al, 2017).

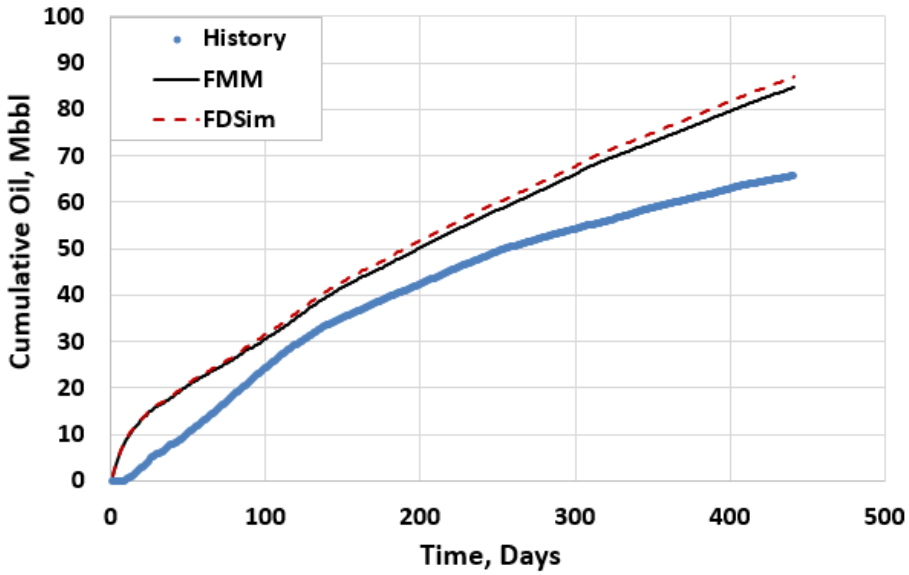


Figure 4.75 Cumulative Oil Production of FMM vs Eclipse as compared to History data with base case variables (compositional FMM)

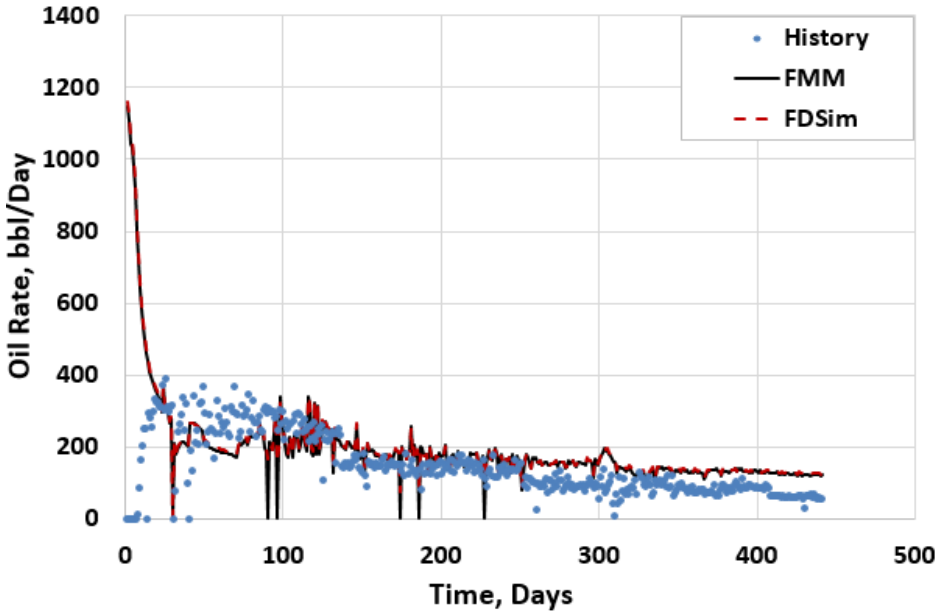


Figure 4.76 Oil Rate Production of FMM vs Eclipse as compared to History data with base case variables (compositional FMM)

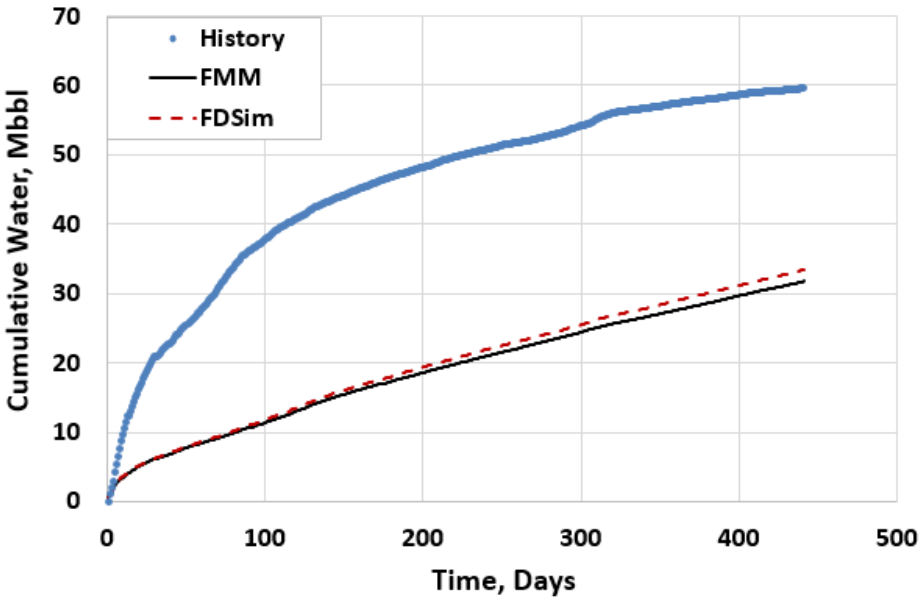


Figure 4.77 Cumulative Water Production of FMM vs Eclipse as compared to History data with base case variables (compositional FMM)

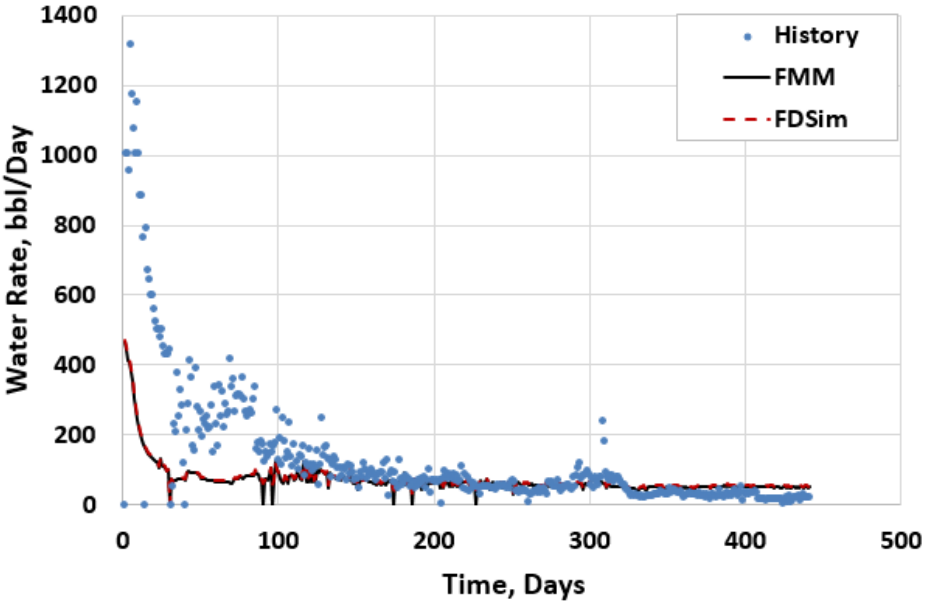


Figure 4.78 Water Rate Production of FMM vs Eclipse as compared to History data with base case variables (compositional FMM)

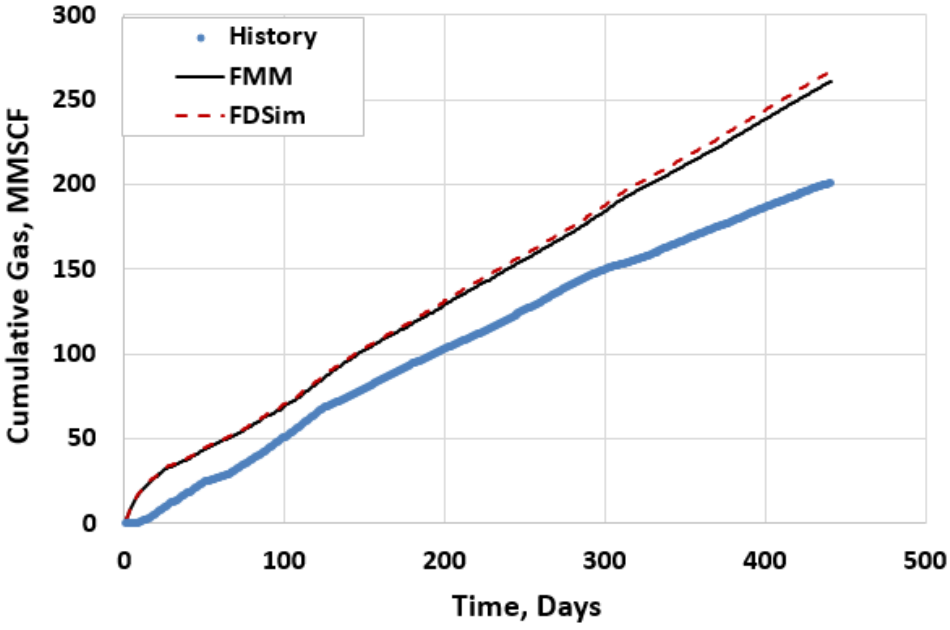


Figure 4.79 Cumulative Gas Production of FMM vs Eclipse as compared to History data with base case variables (compositional FMM)

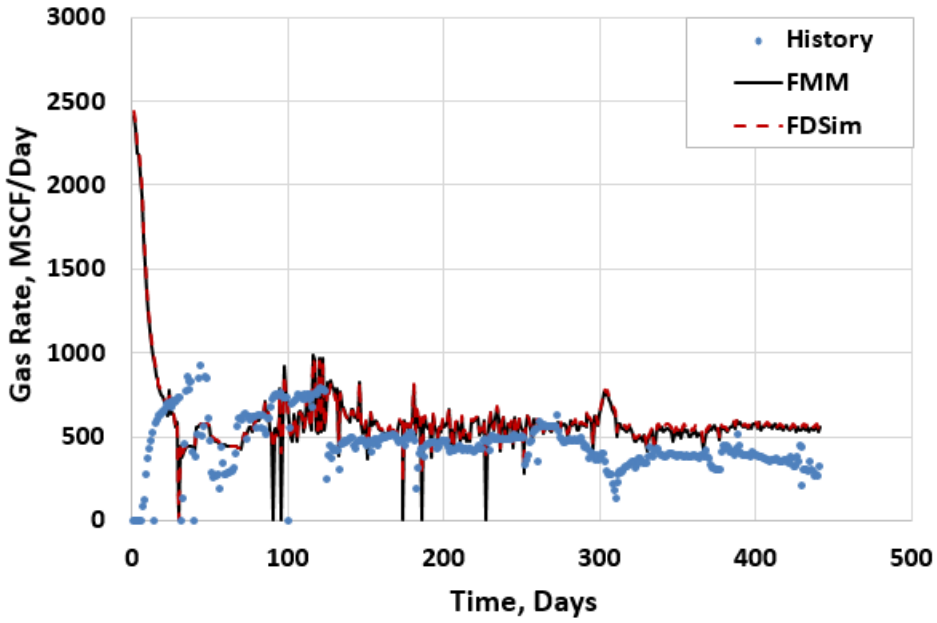


Figure 4.80 Gas Rate Production of FMM vs Eclipse as compared to History data with base case variables (compositional FMM)

As presented in the previous section of this chapter, a multi-stage GA approach has been utilized for this study. In stage 1, sensitivity analysis is done and relative importance of variaous variables are checked. Heavy hitter variables or the variables making relatively larger impact on the objective error functions are identified and rest of the variables are discarded for this stage. **Fig 4.81** shows the results of sensitivity analysis. Parameters not included for this stage GA are shown in green boxes.

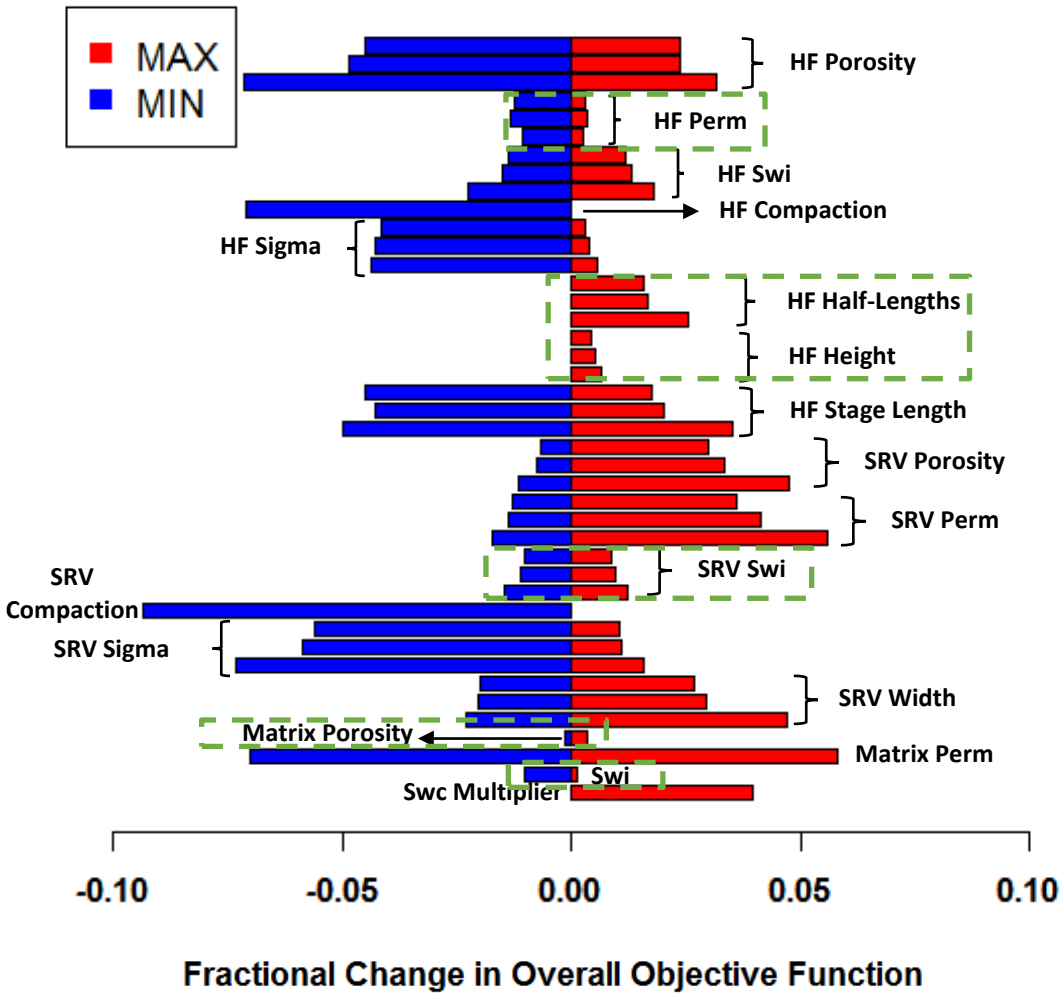


Figure 4.81 Sensitivity analysis at the beginning of Stage 1 (compositional FMM)

Fig. 4.82 shows the results of GA in stage 1. As can be observed from this figure, after multiple generations, improvement in objective error function reduces. Also, since variables in this GA operation show large shrinkage in their ranges from generation 1 to generation 12 **(Figs. 4.83 to 4.88)**, GA was stopped at this point and a collection of best models was selected **(Fig. 4.82).** These best models are chosen to derive new variable ranges of the variables included for this GA stage. **Figs. 4.89 to 4.94** show the variable distribution in generation 1 of this stage while **Figs. 4.95 to 4.100** show the variable ranges in the best models selected at the end of this GA stage. It may be observed that a relatively uniform variable distribution transforms into a narrower and close to normal distribution.

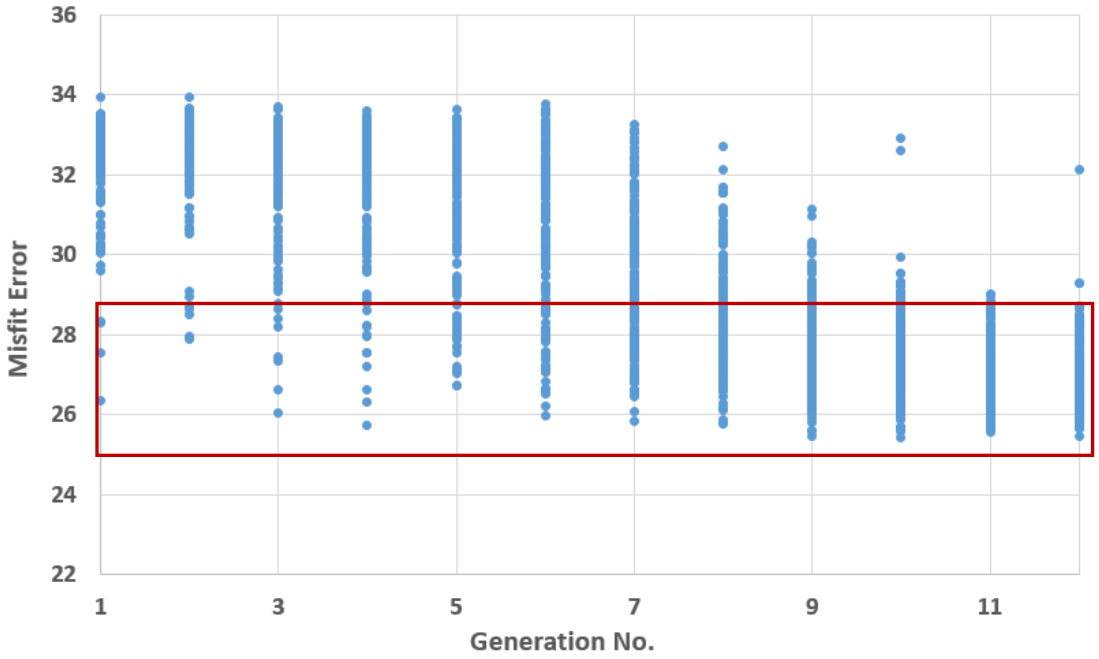


Figure 4.82 GA results for Stage 1 (compositional FMM)

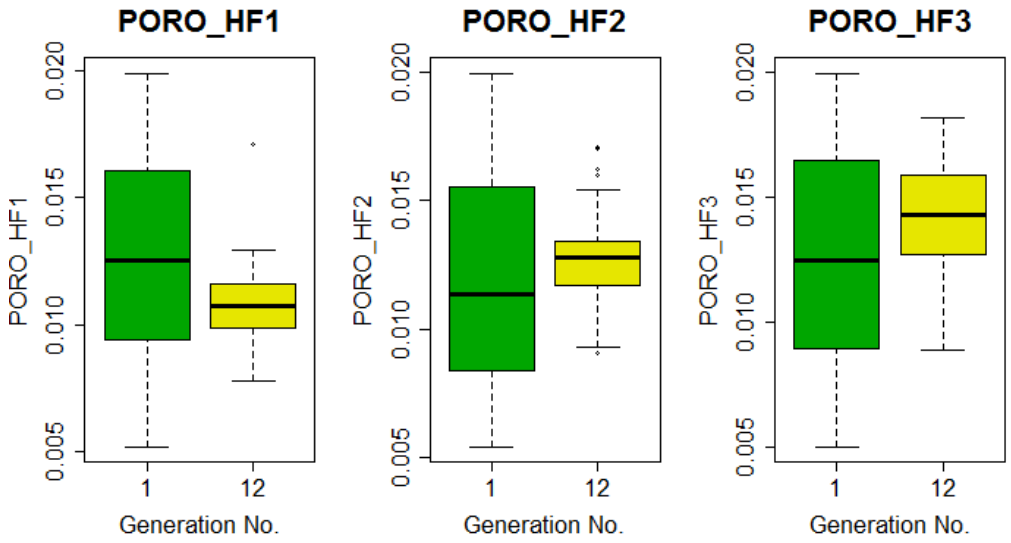


Figure 4.83 Uncertainty reduction in hydraulic fracture porosity during GA - Stage 1 (compositional FMM)

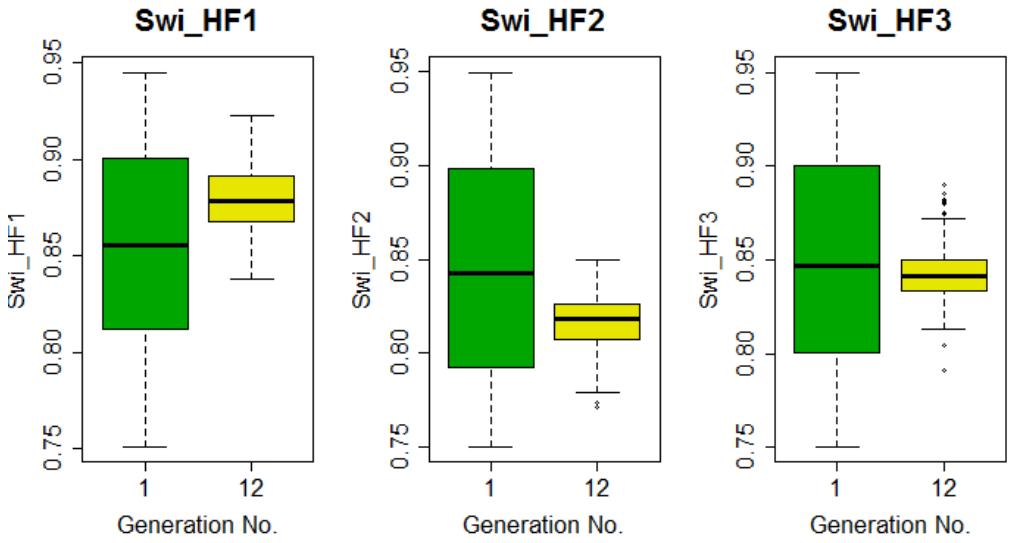


Figure 4.84 Uncertainty reduction in hydraulic fracture initial water saturation during GA - Stage 1 (compositional FMM)

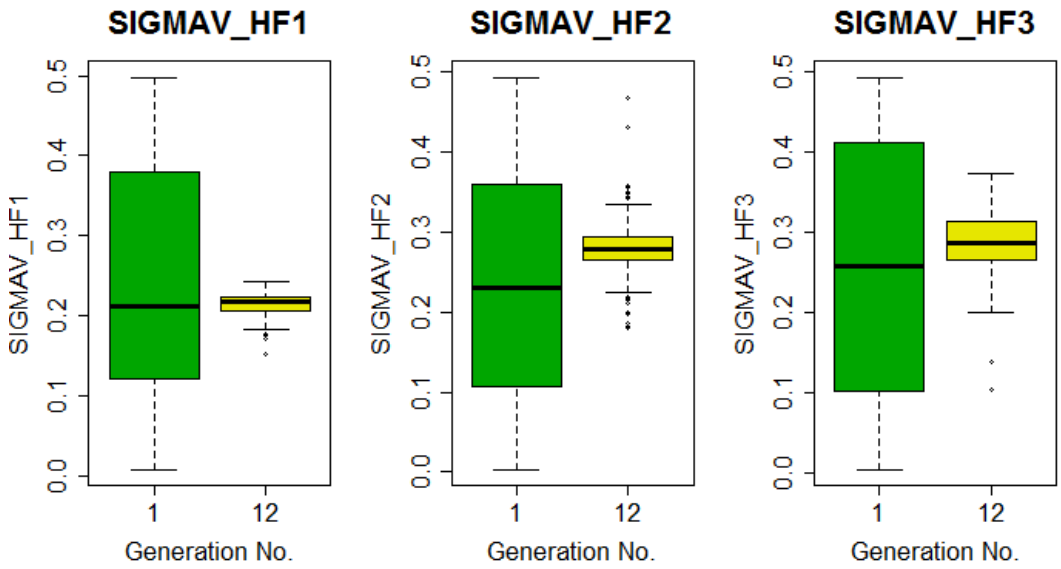


Figure 4.85 Uncertainty reduction in hydraulic fracture shape factor during GA - Stage 1 (compositional FMM)

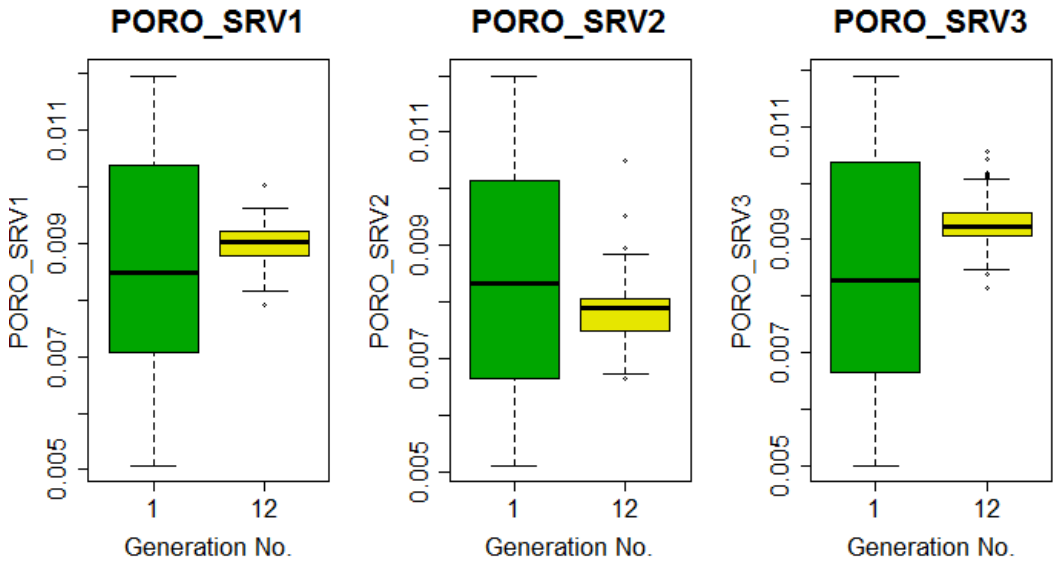


Figure 4.86 Uncertainty reduction in SRV porosity during GA - Stage 1 (compositional FMM)

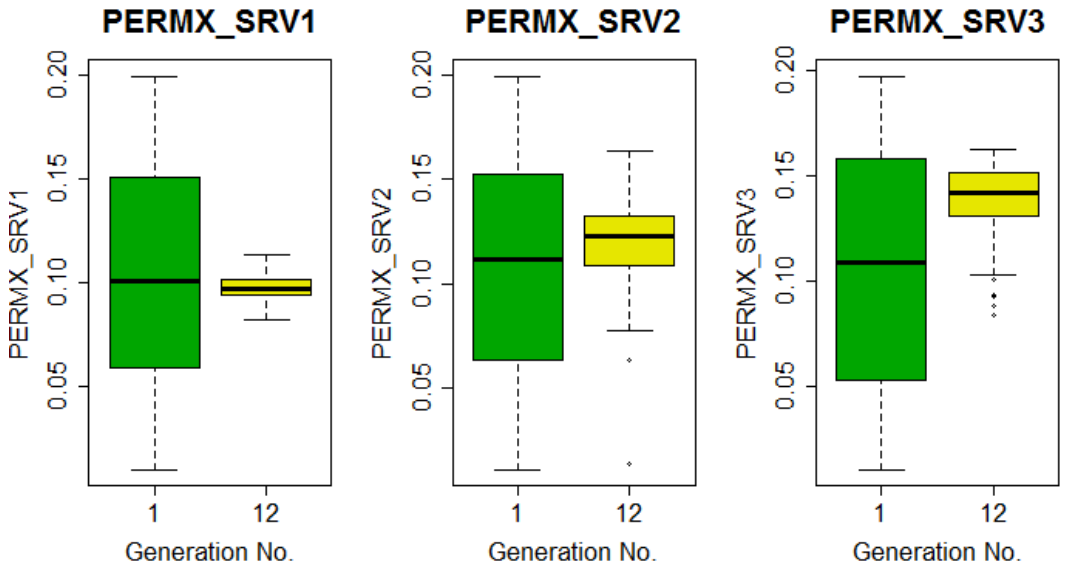


Figure 4.87 Uncertainty reduction in SRV permeability during GA - Stage 1 (compositional FMM)

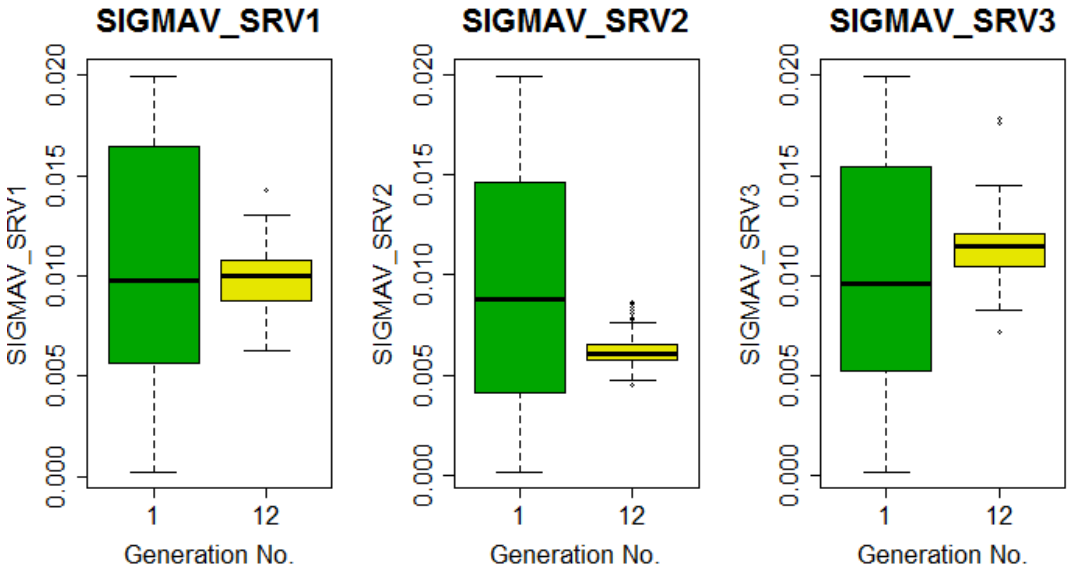


Figure 4.88 Uncertainty reduction in SRV shape factor during GA - Stage 1 (compositional FMM)

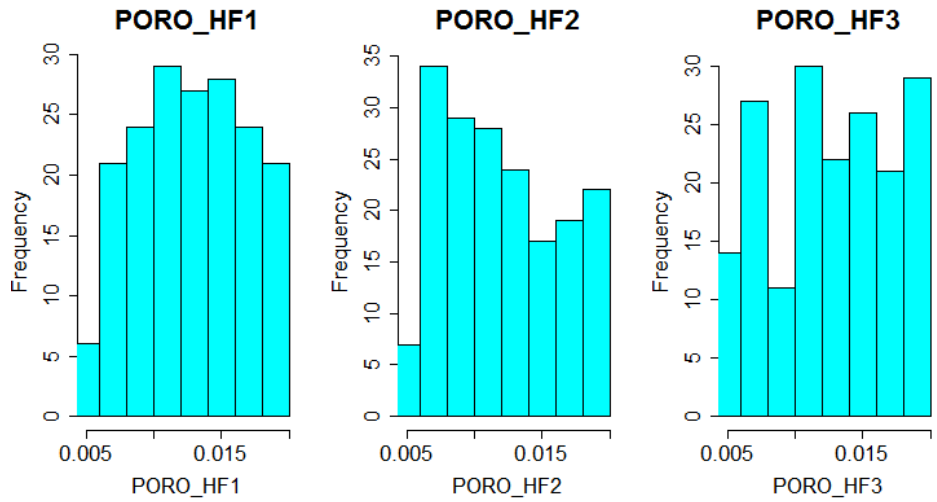


Figure 4.89 Variable distribution of hydraulic fracture porosity in the first generation of GA - Stage 1 (compositional FMM)

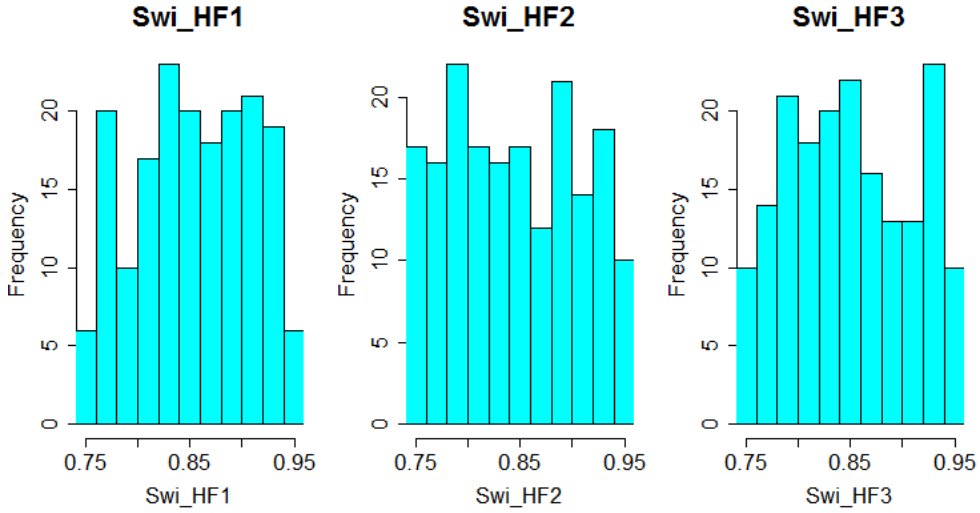


Figure 4.90 Variable distribution of hydraulic fracture initial water saturation in the first generation of GA - Stage 1 (compositional FMM)

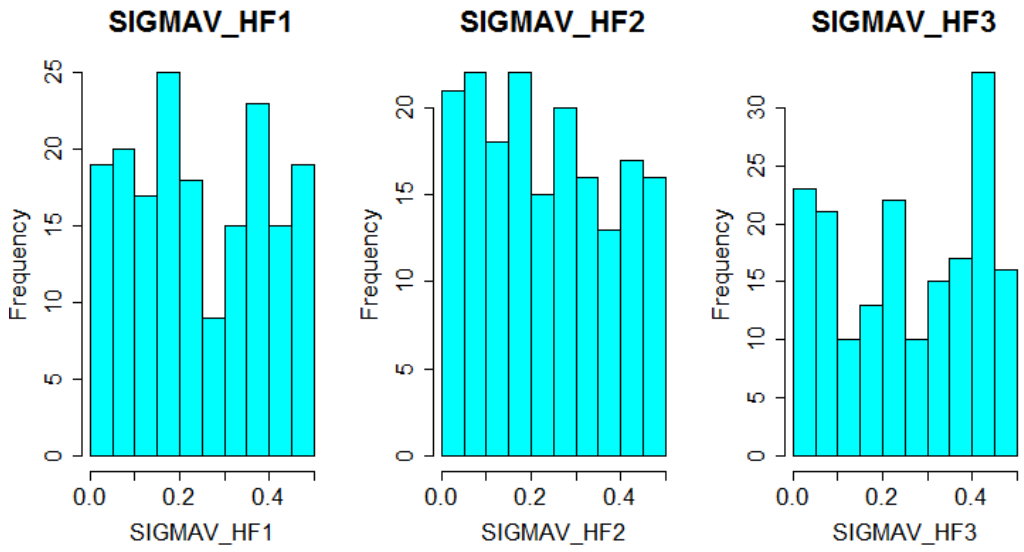


Figure 4.91 Variable distribution of hydraulic fracture shape factor in the first generation of GA - Stage 1 (compositional FMM)

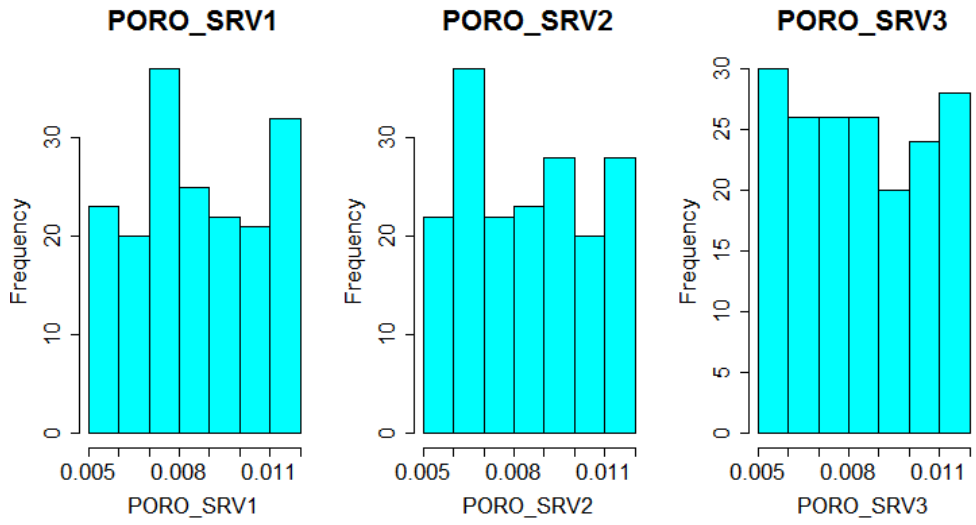


Figure 4.92 Variable distribution of SRV porosity in the first generation of GA - Stage 1 (compositional FMM)

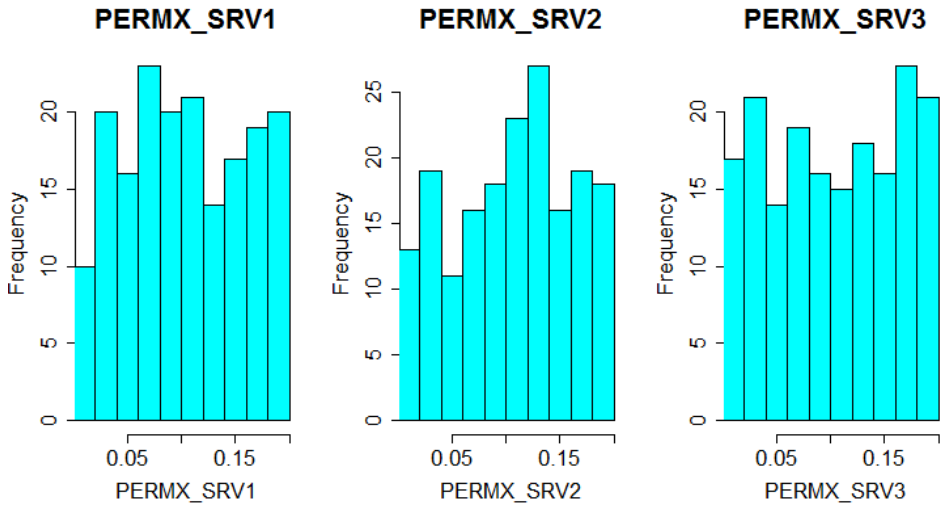


Figure 4.93 Variable distribution of SRV permeability in the first generation of GA - Stage 1 (compositional FMM)

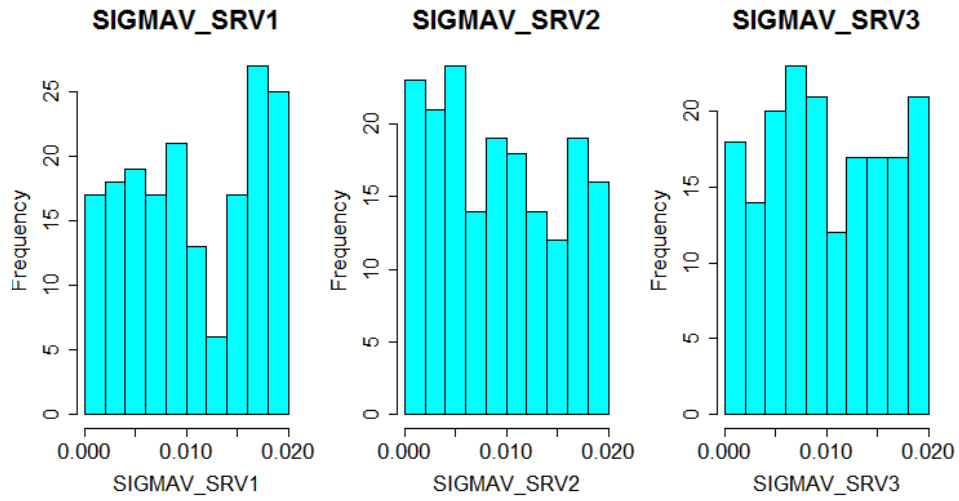


Figure 4.94 Variable distribution of SRV shape factor in the first generation of GA - Stage 1 (compositional FMM)

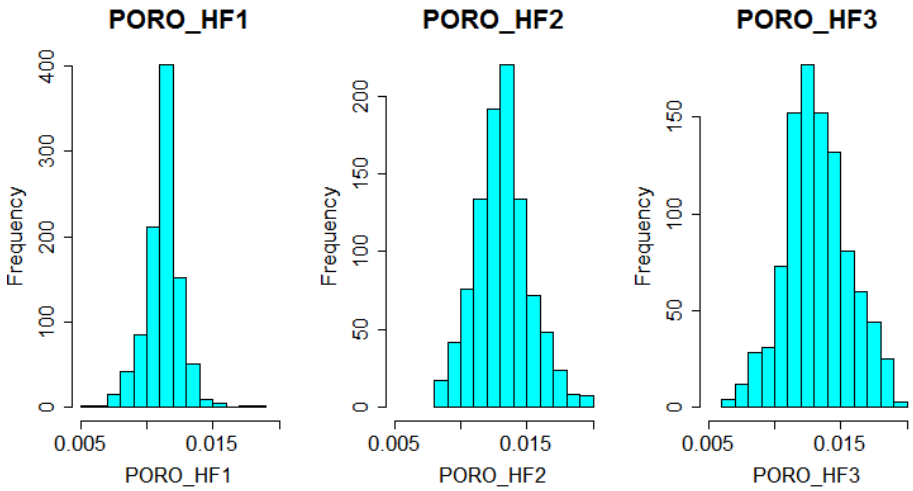


Figure 4.95 Variable distribution of hydraulic fracture porosity in the best selected models of GA - Stage 1 (compositional FMM)

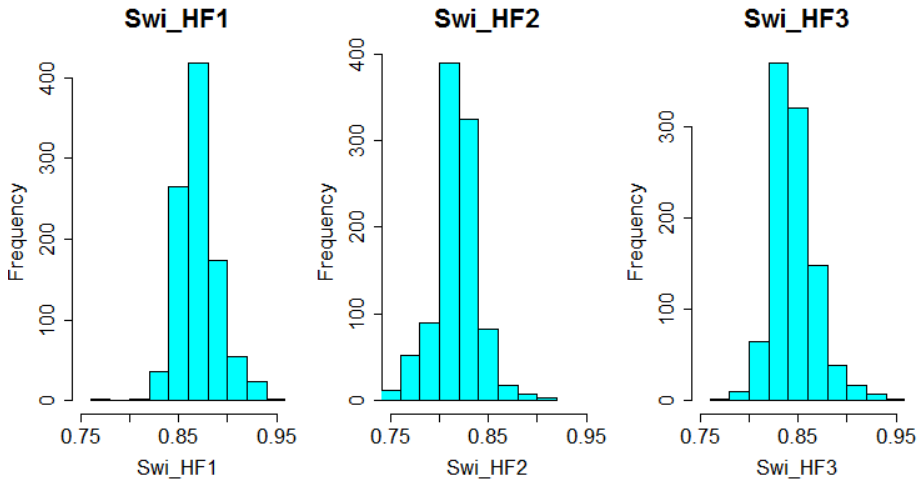


Figure 4.96 Variable distribution of hydraulic fracture initial water saturation in the best selected models of GA - Stage 1 (compositional FMM)

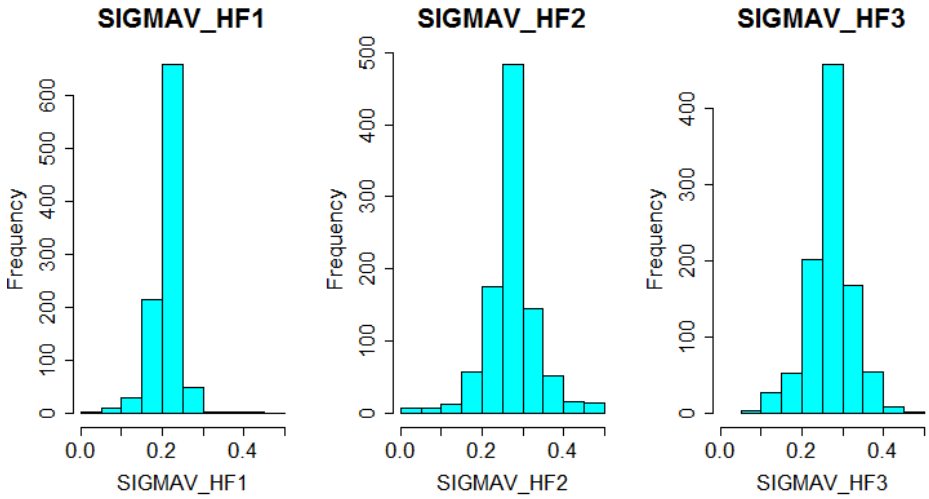


Figure 4.97 Variable distribution of hydraulic fracture shape factor in the best selected models of GA - Stage 1 (compositional FMM)

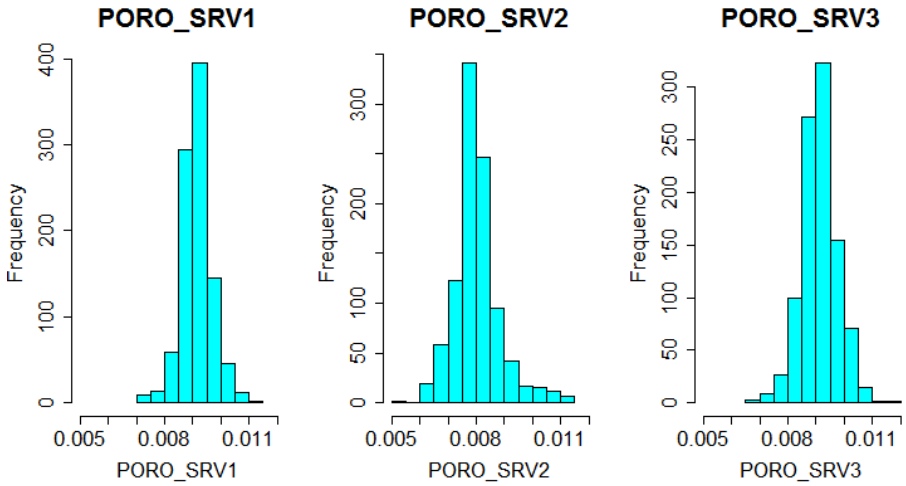


Figure 4.98 Variable distribution of SRV porosity in the best selected models of GA - Stage 1 (compositional FMM)

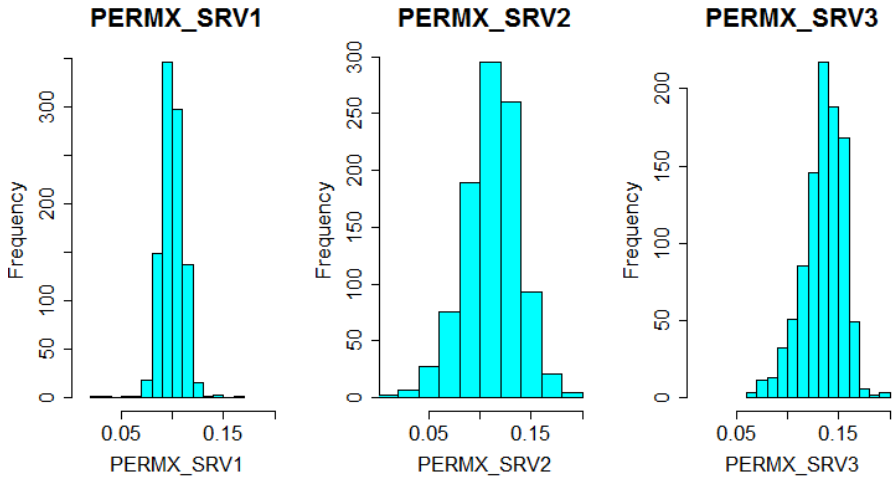


Figure 4.99 Variable distribution of SRV permeability in the best selected models of GA - Stage 1 (compositional FMM)

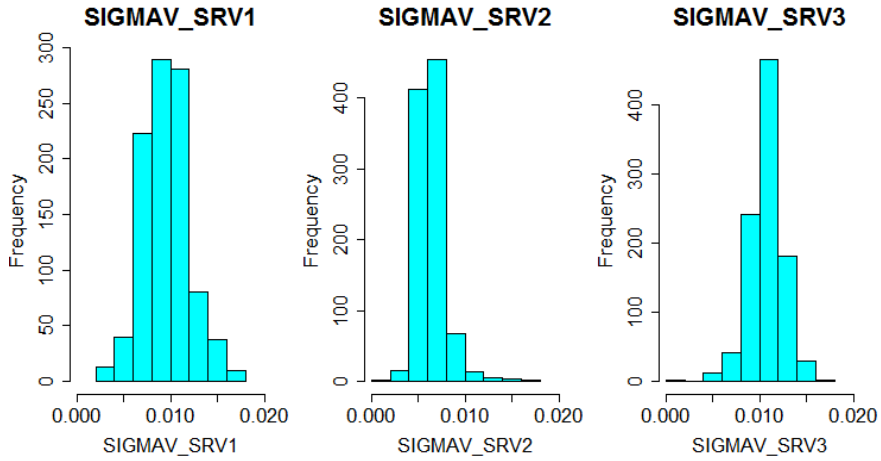


Figure 4.100 Variable distribution of SRV shape factor in the best selected models of GA - Stage 1 (compositional FMM)

In the next GA stage, the variables of the previous stage are kept with updated ranges based on best models selected previously. **Fig. 4.101** shows the new sensitivity plot. It can be observed that this time, some of the variables are not making big impact

due to shrinkage of their ranges in the previous GA stages. However, all the variables are included in this GA stage.

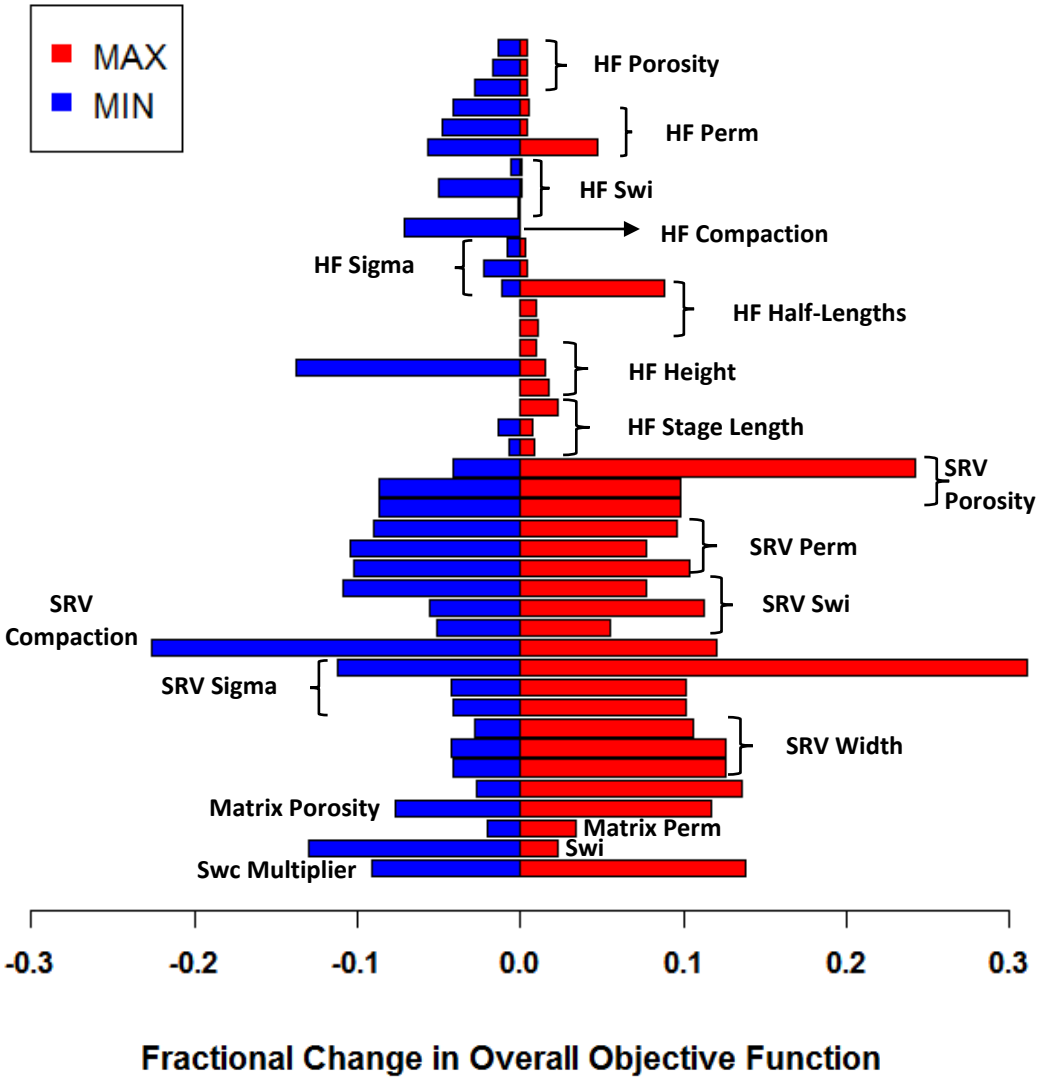


Figure 4.101 Sensitivity analysis at the beginning of Stage 2 (compositional FMM)

Fig. 4.102 shows the results of GA in stage 2. As can be observed from this figure, after multiple generations, improvement in objective error function reduces. Also, since

variables in this GA operation show large shrinkage in their ranges from generation 1 to generation 10 **(Figs. 4.103 to 4.110)**, GA was stopped at this point and a collection of best models was selected **(Fig. 4.102).** These best models are chosen to derive new variable ranges of the variables included for this GA stage. **Figs. 4.111 to 4.118** show the variable ranges in the best models selected at the end of this GA stage. It may be observed that distributions of the variables common with previous stage have become narrower showing further reduction in uncertainty.

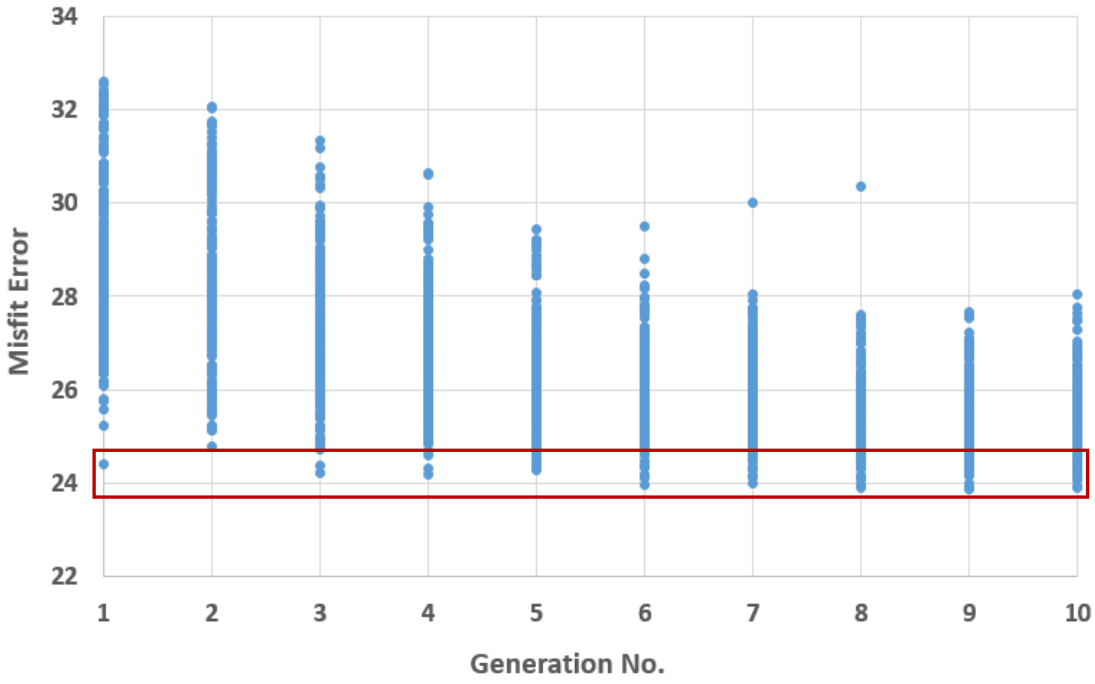


Figure 4.102 GA results for Stage 2 (compositional FMM)

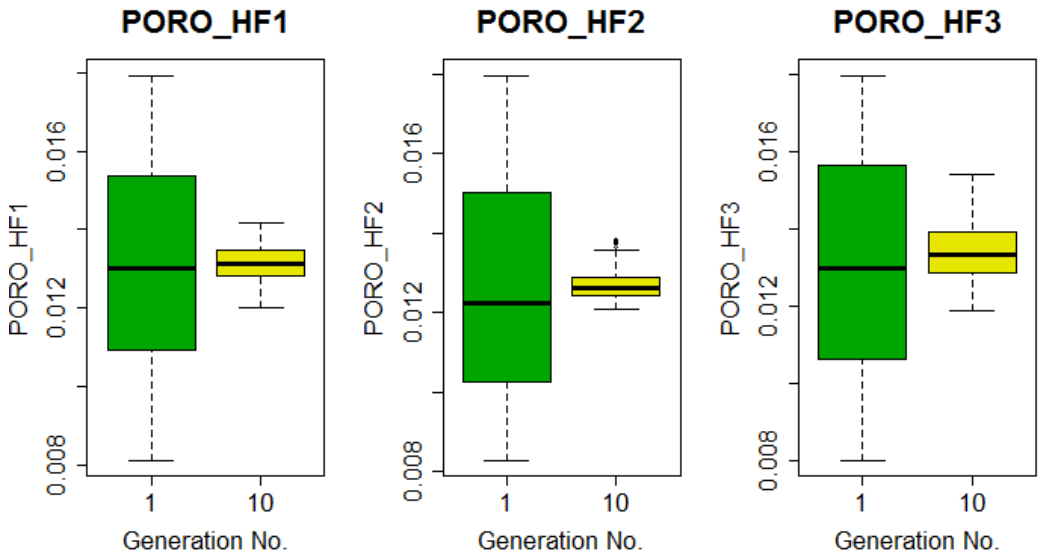


Figure 4.103 Uncertainty reduction in hydraulic fracture porosity during GA - Stage 2 (compositional FMM)

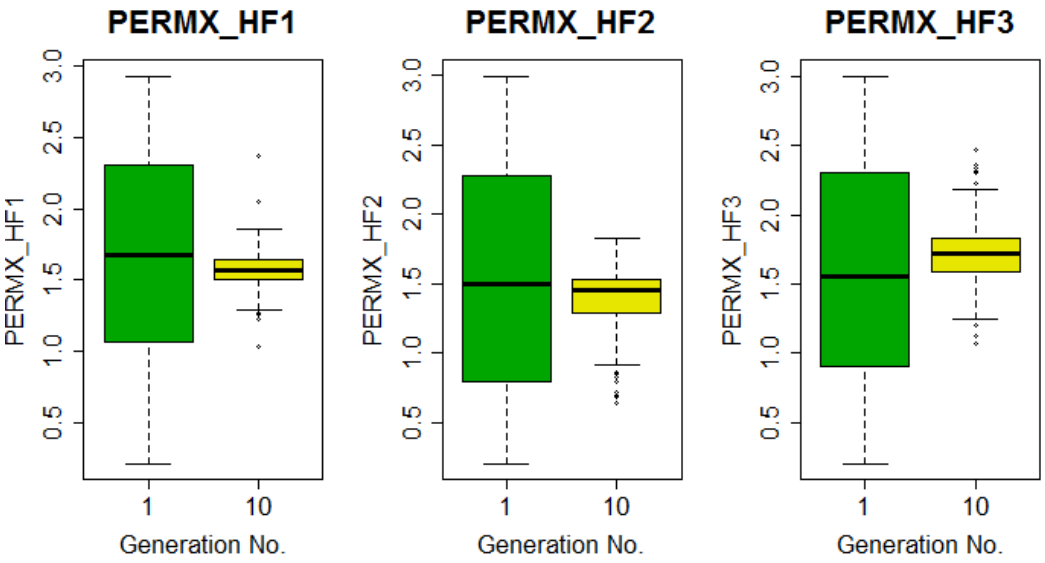


Figure 4.104 Uncertainty reduction in hydraulic fracture permeability during GA - Stage 2 (compositional FMM)

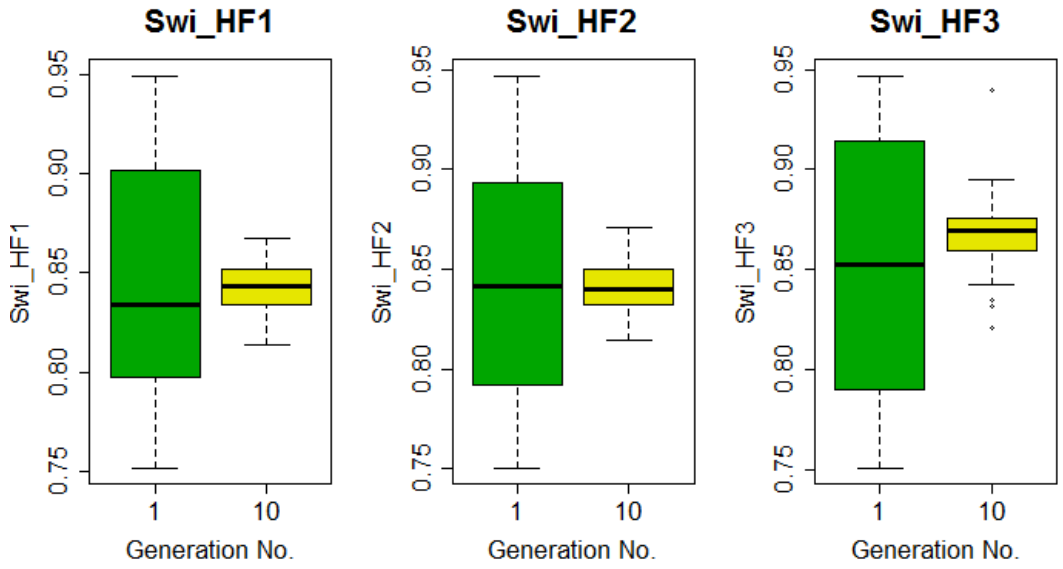


Figure 4.105 Uncertainty reduction in hydraulic fracture initial water saturation during GA - Stage 2 (compositional FMM)

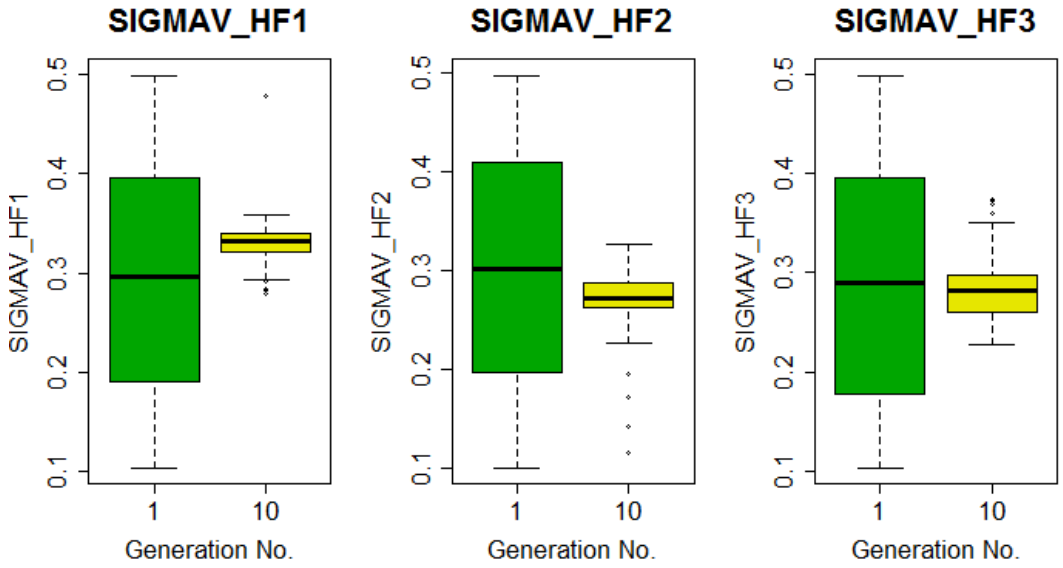


Figure 4.106 Uncertainty reduction in hydraulic fracture shape factor during GA - Stage 2 (compositional FMM)

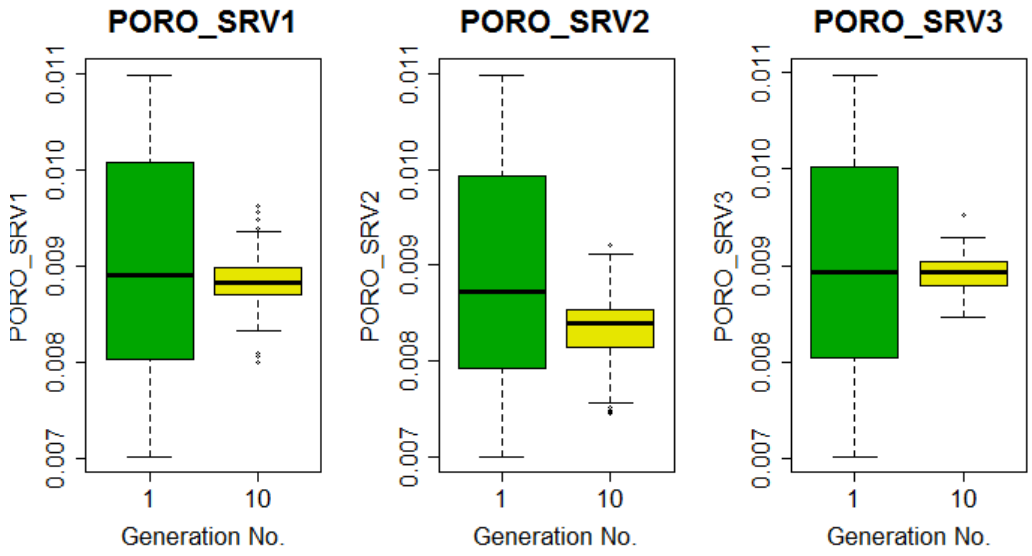


Figure 4.107 Uncertainty reduction in SRV porosity during GA - Stage 2 (compositional FMM)

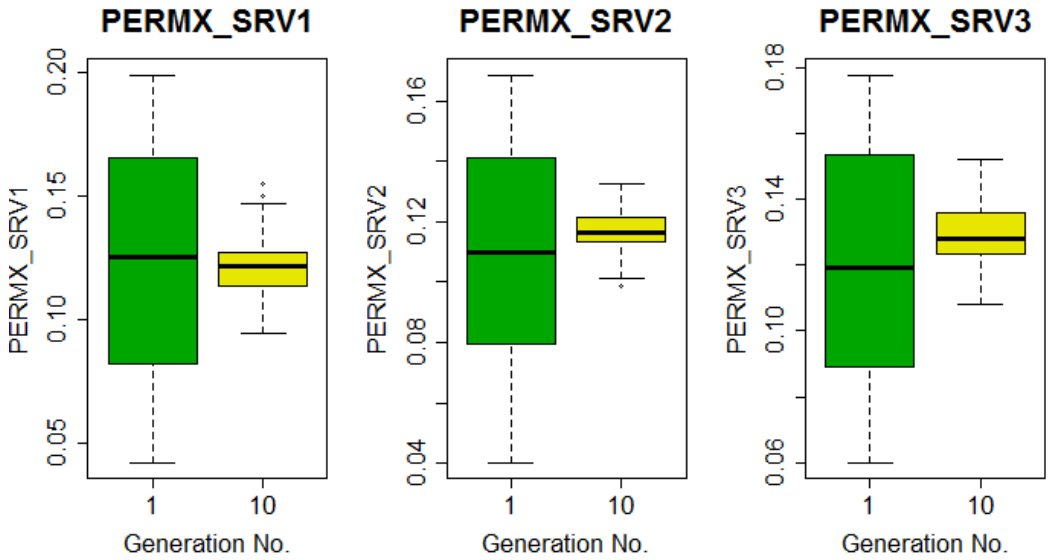


Figure 4.108 Uncertainty reduction in SRV permeability during GA - Stage 2 (compositional FMM)

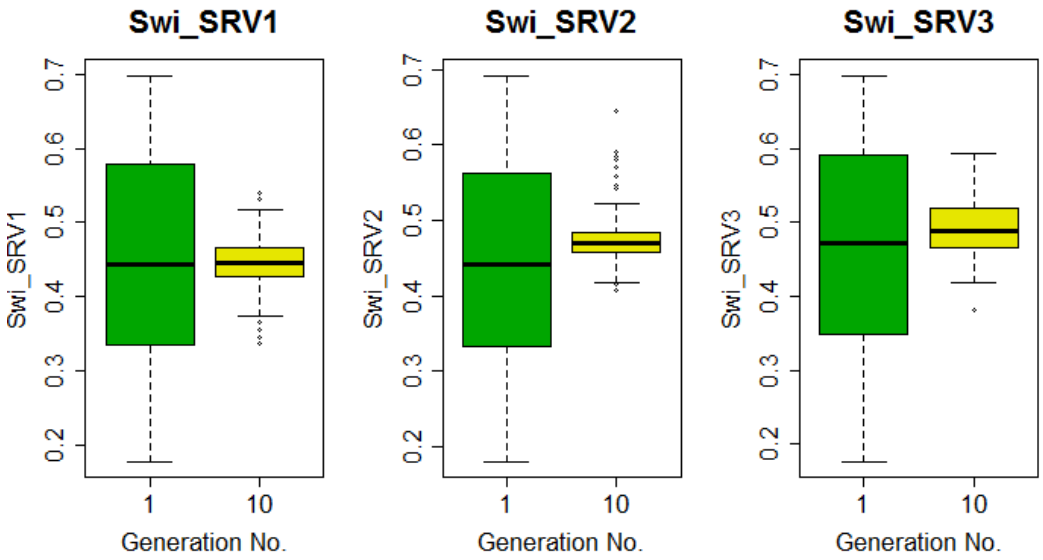


Figure 4.109 Uncertainty reduction in SRV initial water saturation during GA - Stage 2 (compositional FMM)

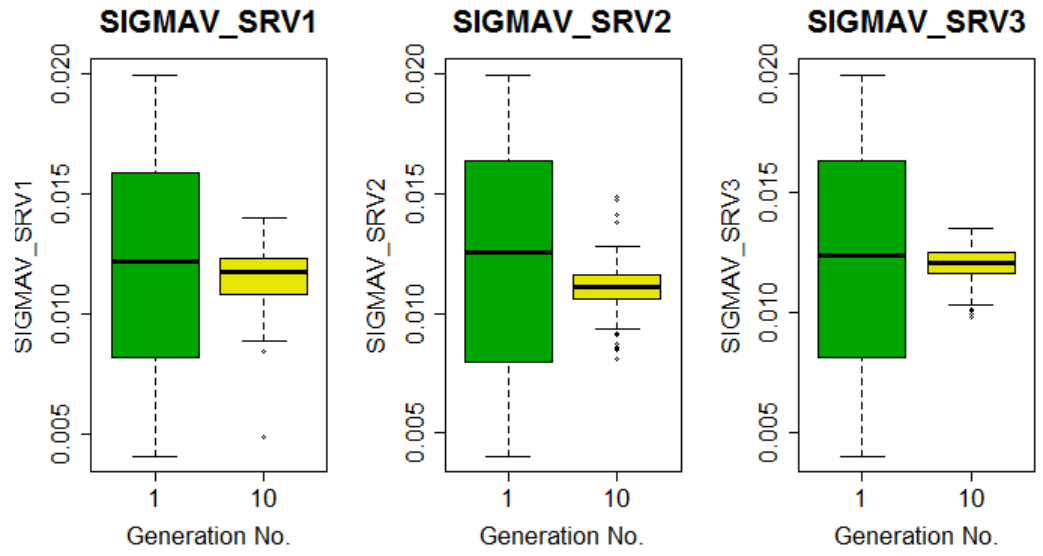


Figure 4.110 Uncertainty reduction in SRV shape factor during GA - Stage 2 (compositional FMM)

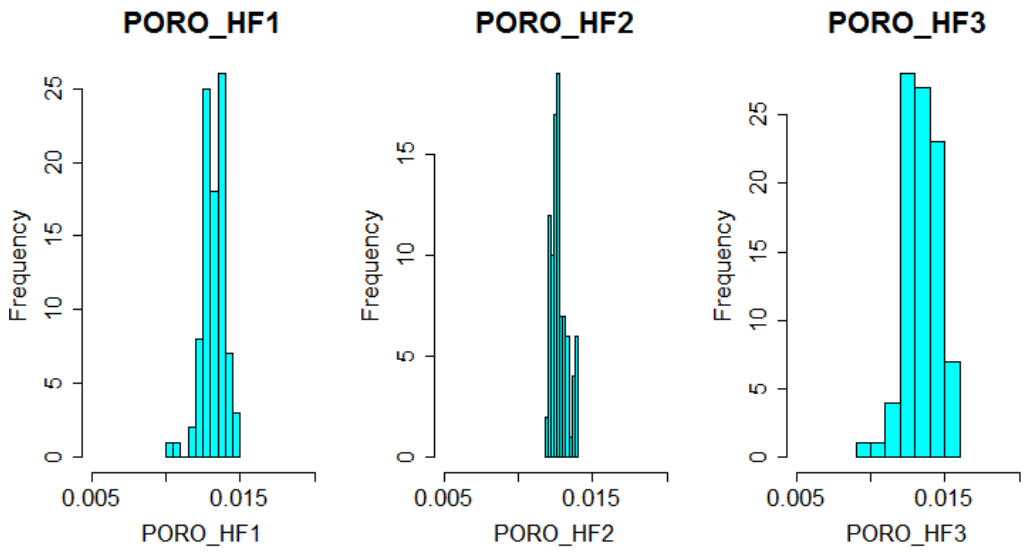


Figure 4.111 Variable distribution of hydraulic fracture porosity in the best selected models of GA - Stage 2 (compositional FMM)

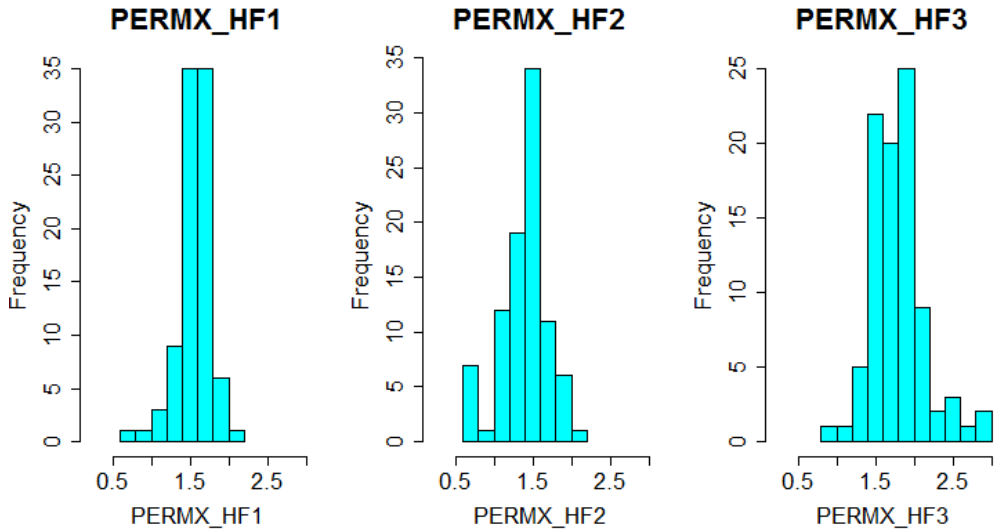


Figure 4.112 Variable distribution of hydraulic fracture permeability in the best selected models of GA - Stage 2 (compositional FMM)

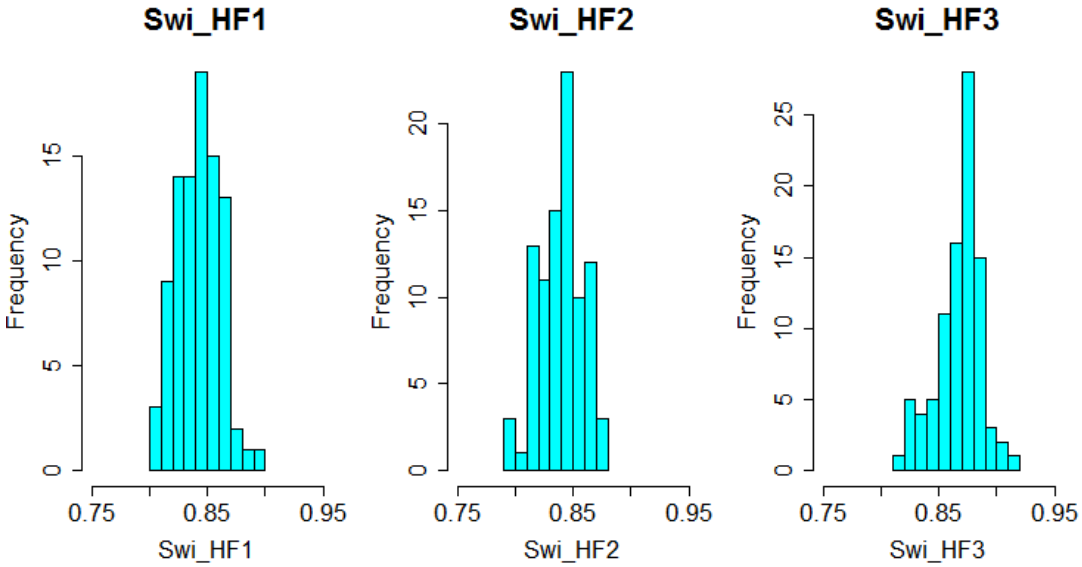


Figure 4.113 Variable distribution of hydraulic fracture initial water saturation in the best selected models of GA - Stage 2 (compositional FMM)

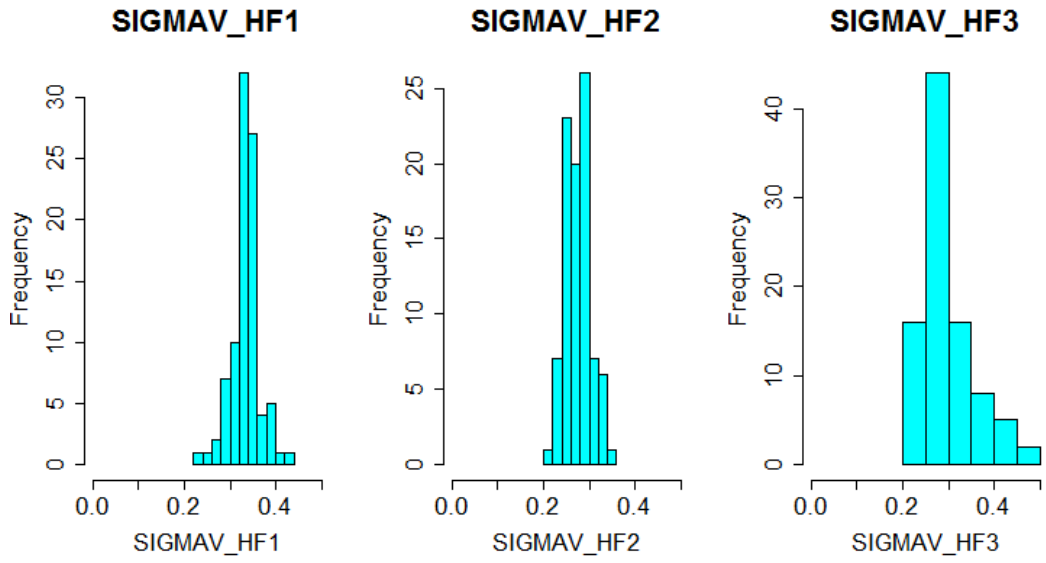


Figure 4.114 Variable distribution of hydraulic fracture shape factor in the best selected models of GA - Stage 2 (compositional FMM)

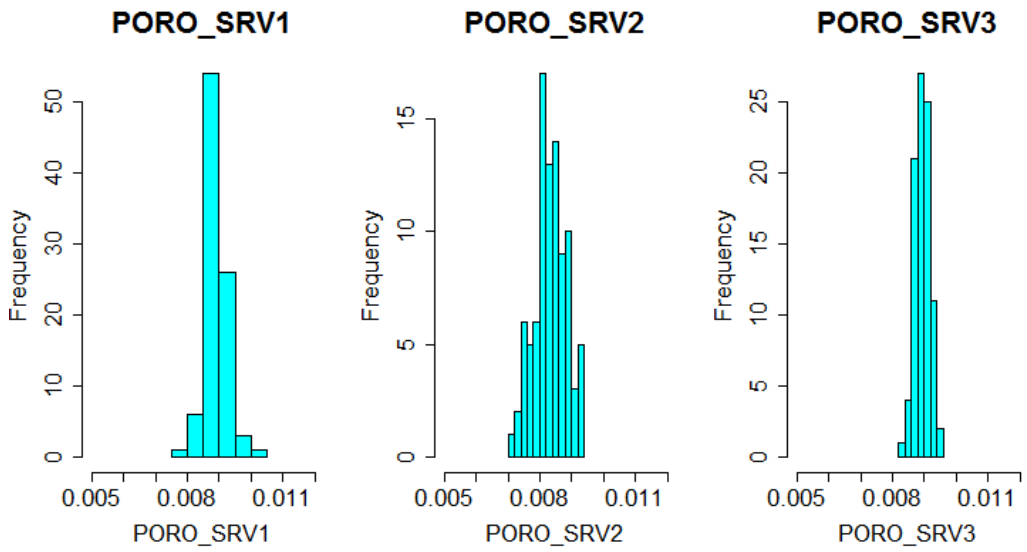


Figure 4.115 Variable distribution of SRV porosity in the best selected models of GA - Stage 2 (compositional FMM)

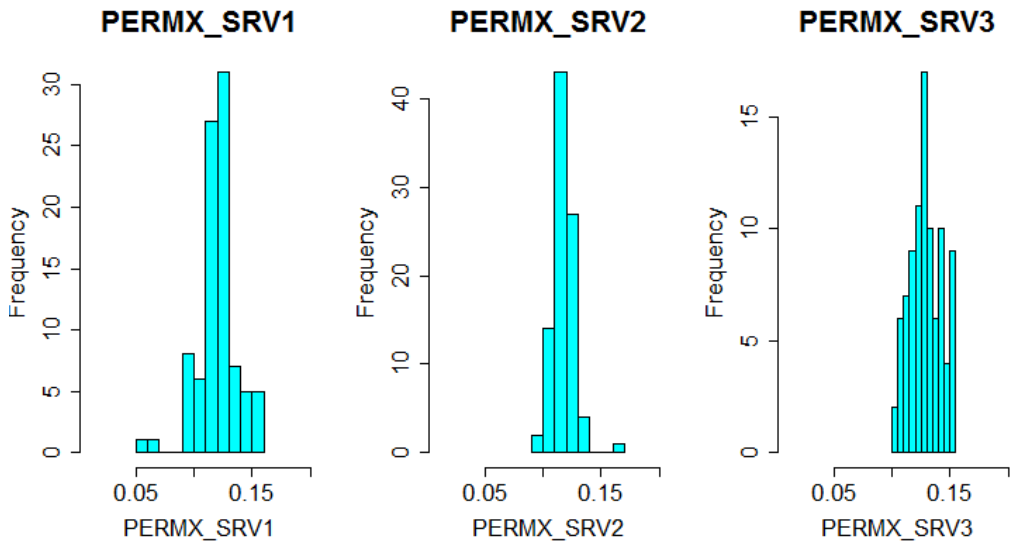


Figure 4.116 Variable distribution of SRV permeability in the best selected models of GA - Stage 2 (compositional FMM)

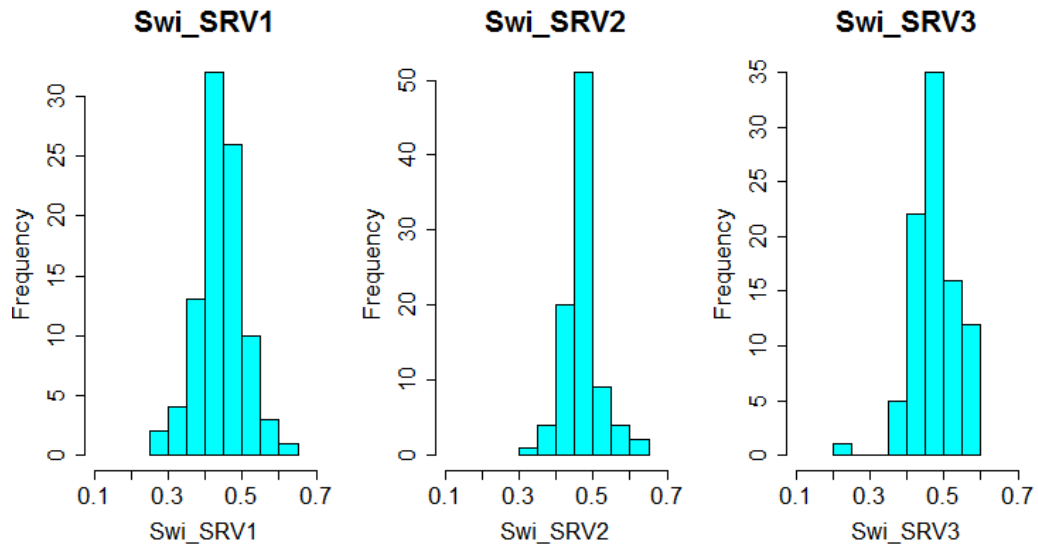


Figure 4.117 Variable distribution of SRV initial water saturation in the best selected models of GA - Stage 2 (compositional FMM)

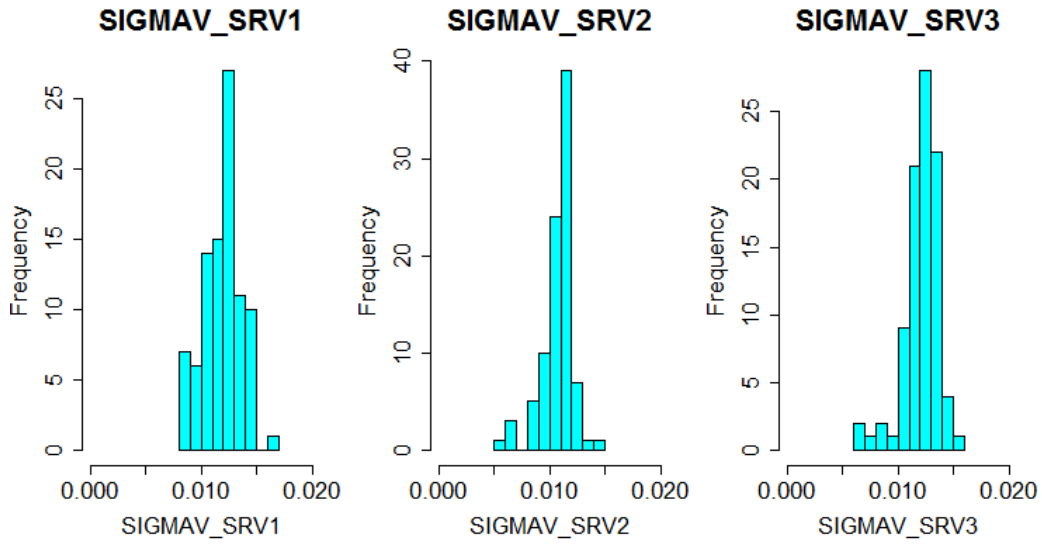


Figure 4.118 Variable distribution of SRV shape factor in the best selected models of GA - Stage 2 (compositional FMM)

Fig. 4.119 shows the combined plot showing all GA stages. It may be observed that there is significant improvement from one GA stage to the next one. At this point the best models are selected as mentioned previously and plotted against history data **(Figs. 4.120 to 4.125)**.

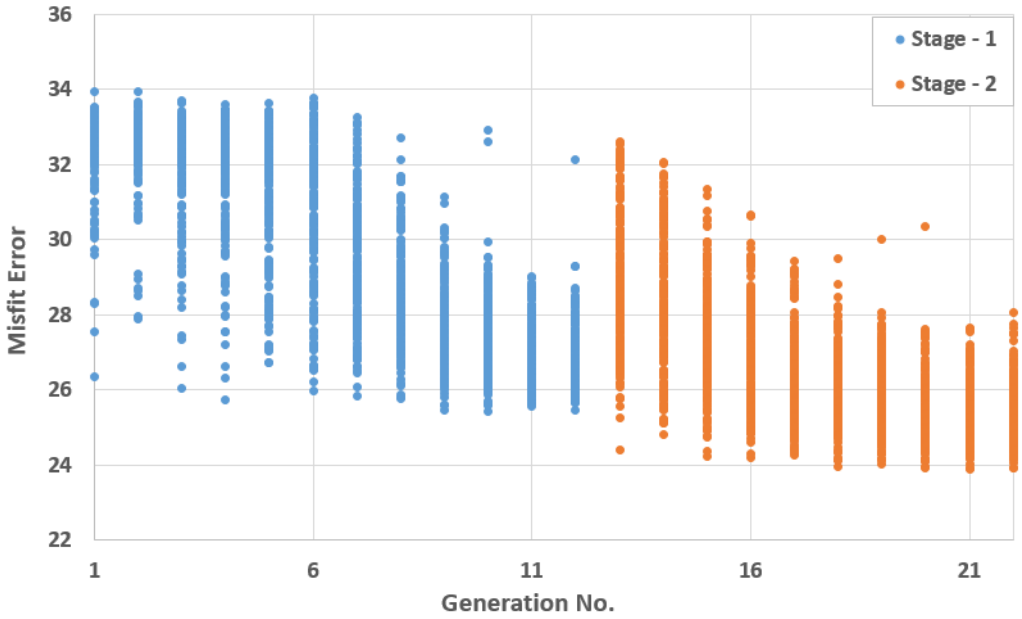
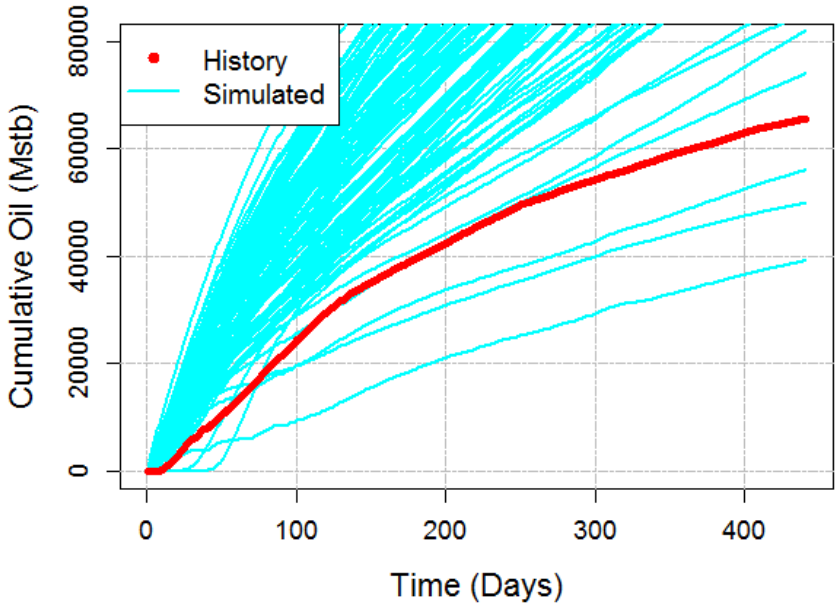
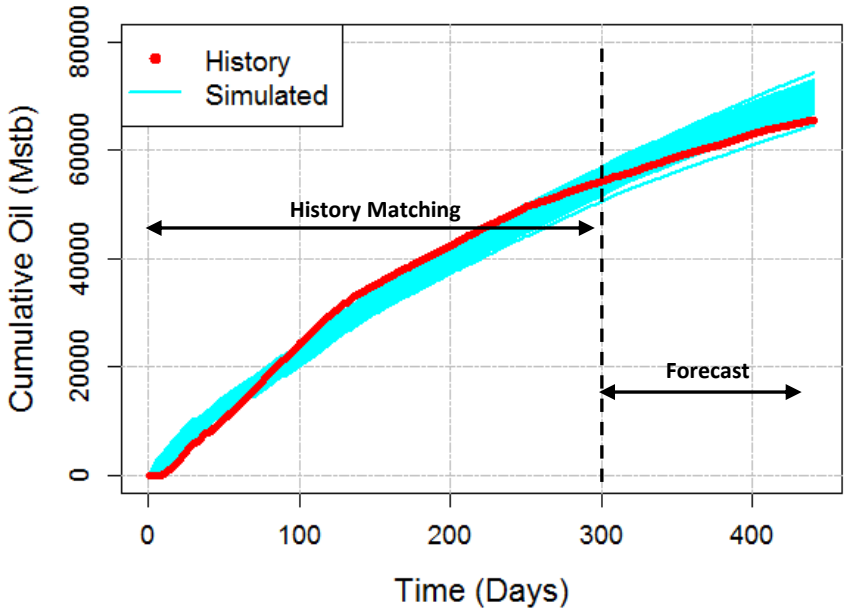


Figure 4.119 Combined GA results of all stages (compositional FMM)

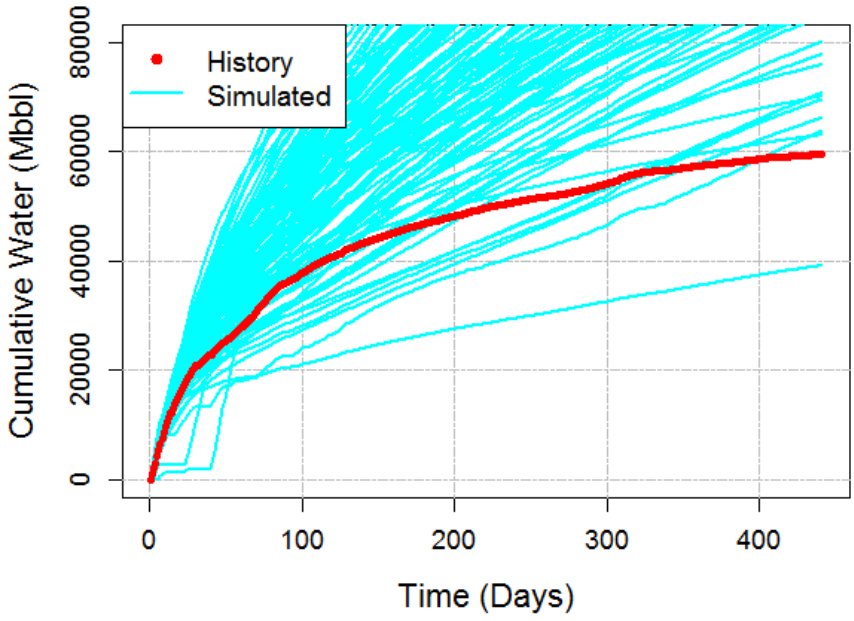


(a)

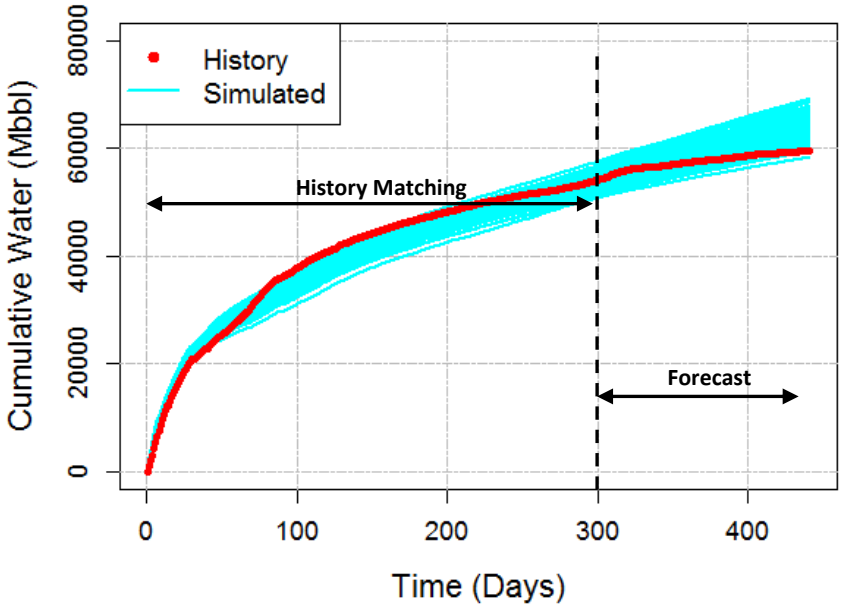


(b)

Figure 4.120 Cumulative oil history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (compositional FMM)

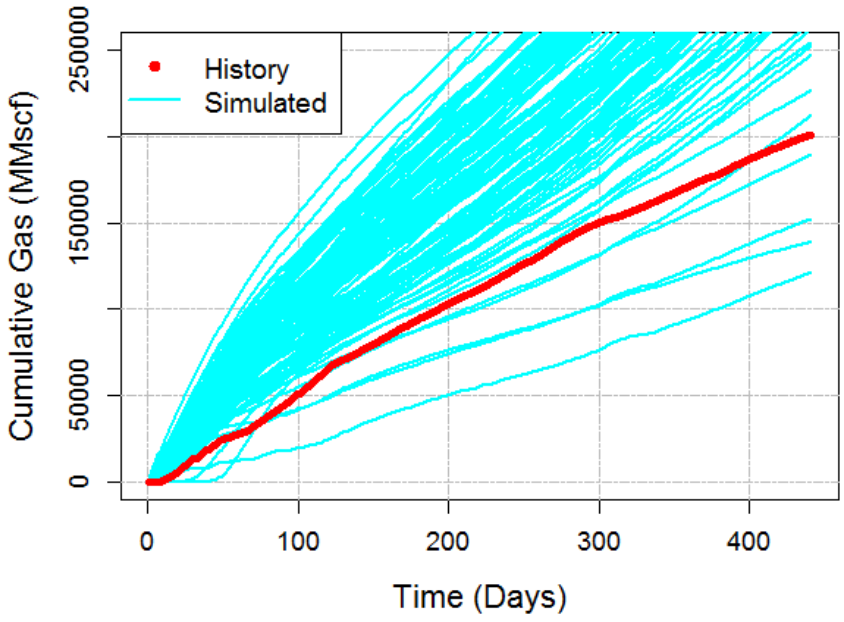


(a)

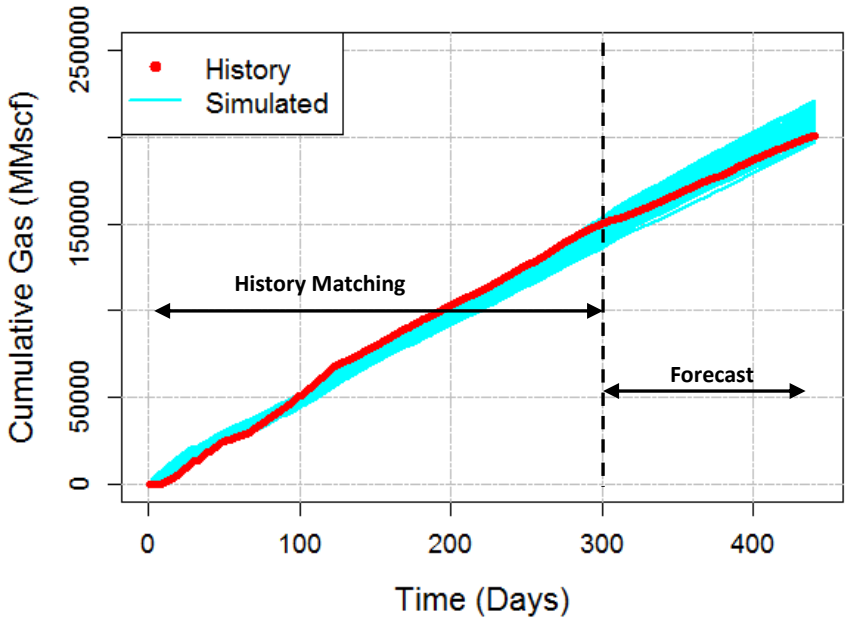


(b)

Figure 4.121 Cumulative Water history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (compositional FMM)

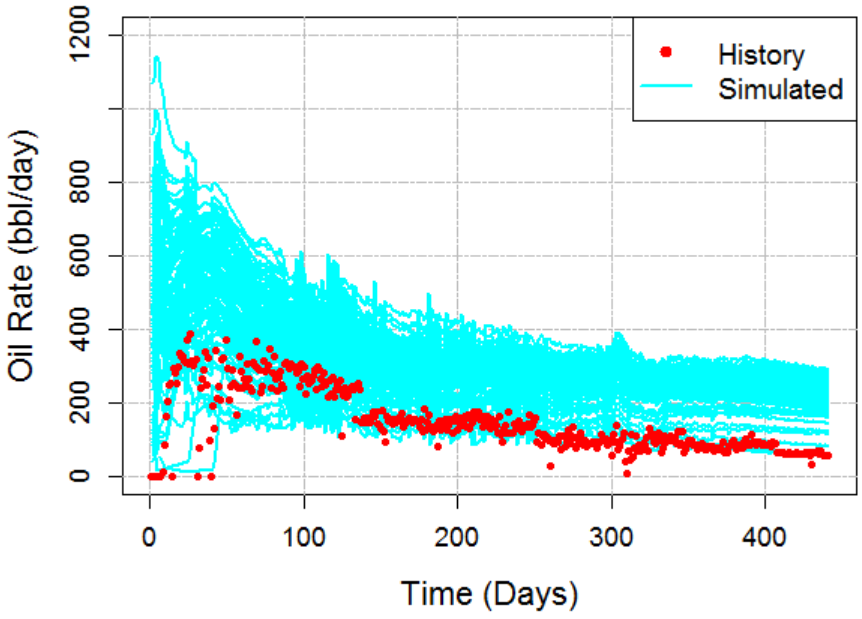


(a)

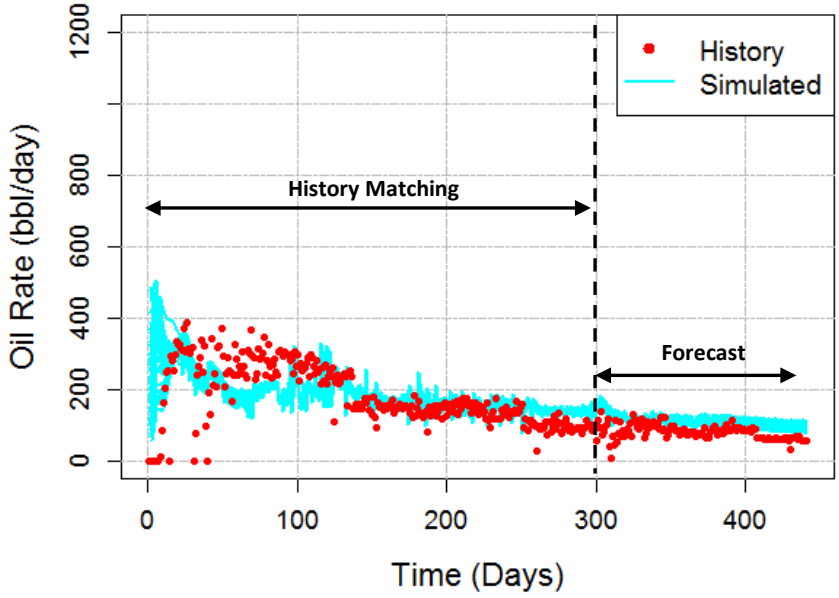


(b)

Figure 4.122 Cumulative Gas history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (compositional FMM)

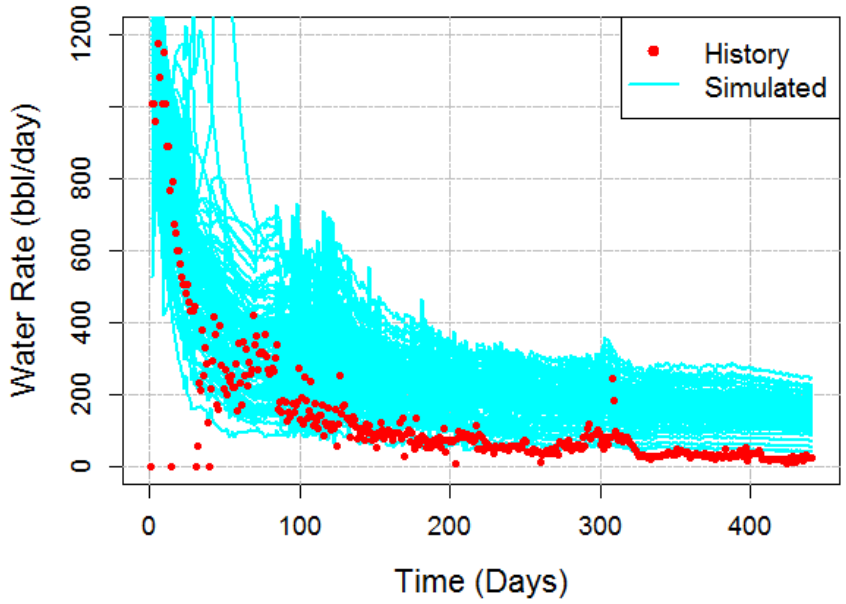


(a)

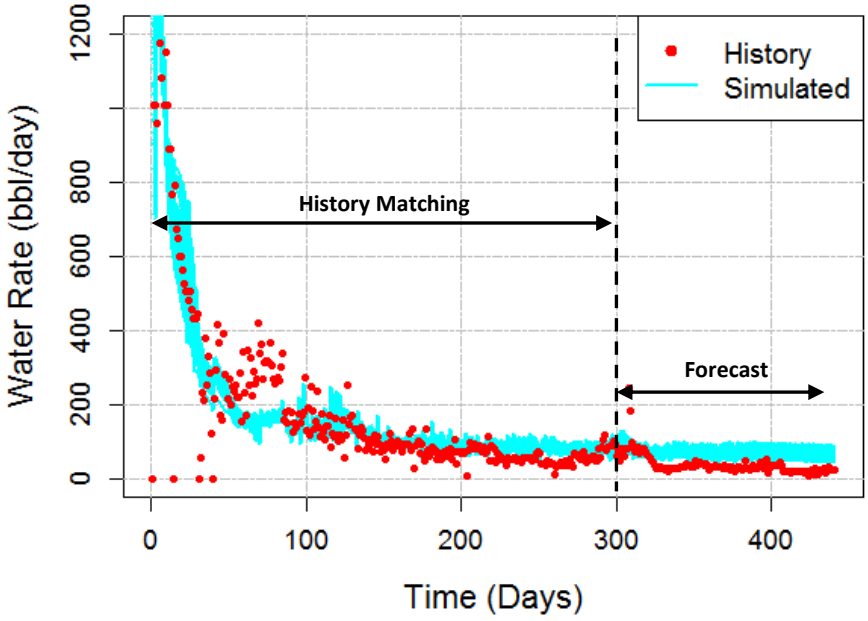


(b)

Figure 4.123 Oil rate history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (compositional FMM)

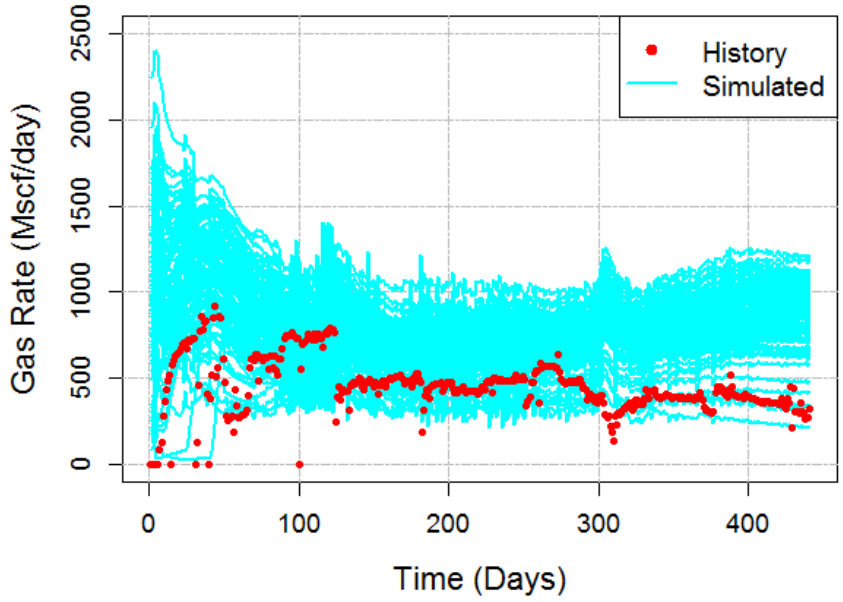


(a)

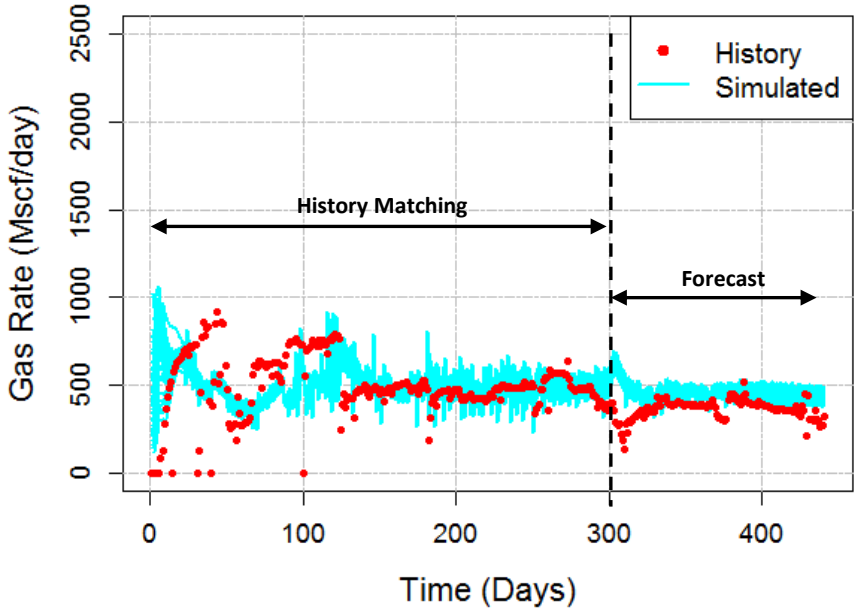


(b)

Figure 4.124 Water rate history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (compositional FMM)



(a)



(b)

Figure 4.125 Gas rate history production data vs simulated production data (a) in the first stage first generation and (b) including only the best selected models from the last stage (compositional FMM)

4.4 Summary

1. History matching using GA can be an effective tool in reducing model variable uncertainty. Results show that variable uncertainty can be significantly reduced from first generation to the final generation.
2. In a scenario with unknown variable sensitivities and ranges, taking a larger initial variable range is common. This study shows how heavy-hitter variables can be separated out from other variables and GA can then be conducted only using heavy-hitter variables. GA can be repeated in the next stage(s) by including previously eliminated variables and refined ranges of heavy hitters.
3. Best models can be selected from a GA stage to repeat workflow for the new stage thus converging to the solution faster. Variable distribution plots presented for best selected models explain how a uniform distribution in the beginning of stage 1 can be reduced to a smaller range of normal distribution in a stage. This refined variable range can then be carried over to the next GA stage.
4. History matching and forecast results for the field case has been presented using a multi-stage GA approach. It has been shown that multi-stage GA can be a faster alternative to single stage GA to get reasonable history matching results.
5. FMM based simulator has been proven to be an accurate and faster alternative to commercial simulator for an optimization study requiring large number of forward simulations. However, this study can be repeated using any commercial finite difference simulator.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary and Conclusions

This dissertation study has presented different applications of machine learning algorithms including GA. Following conclusions can be drawn from this dissertation:

1. In the second chapter, Eagle Ford well data was collected from public website and fitted with various decline curve models to get best fit decline curve parameters and expected EUR for each well. Several machine learning algorithms such as Random Forest, Support Vector Machine and Gradient Boosting Machines are then applied to correlate well decline curve parameters and EUR to well completion and well location variables. The models thus developed have been utilized to predict well rate production as a function of time and also well EUR with reasonable accuracy. Also, variables making most impact on the EUR have been identified in this study.
2. In the third chapter, Genetic Algorithm (GA) based workflow has been presented to optimize the Net Present Value (NPV) during well production period. It has been found in this chapter that NPV cannot be optimized simply by increasing the number of stages in a horizontal well. A GA based workflow which involves various fracturing variables such as proppant amount and fracturing fluid amount has been presented and applied to a synthetic unconventional shale gas reservoir model. The most optimum design variable set has been compared to the uniformly spaced design to compare the difference between the two cases. Also, this chapter presents the effects of uncertainty in reservoir permeability on NPV if the presented workflow is

used to optimize the hydraulic fracture design.

3. In the fourth chapter, a multistage GA approach has been presented to match history data in a shale oil field case. In this method only the most significant history matching variables are utilized in the first stage of GA. Once first stage converges based on criteria mentioned in this chapter, next stage including updated variables and their ranges are utilized. The updated variable ranges are based upon the best models in the previous stage. This method can further fine tune variable ranges with better history matching error as compared to single stage GA.

5.2 Recommendations

Following points are recommended as an extension/improvement to current dissertation work:

1. In the second chapter study, more variables can be included that impact well rates such as well head pressure/bottom hole pressure. Also, in case of major changes in the well constraint variables, fitting a single decline curve may not be suitable for a given well. In that case multiple decline curves may be fitted and predicted.
2. In the third chapter, ways to predict natural fracture distribution in larger uncertainty is needed in case this workflow is applied to reservoirs with little or no knowledge of natural fracture density distribution.

NOMENCLATURE

a	=	Intercept Constant (Duong Model)
α or alpha	=	Scale parameter (Weibull Model)
b	=	Decline coefficient (Arps)
ACE	=	Alternating Conditional Expectation
BMA	=	Bayesian Model Averaging
BHP	=	Bottom Hole Pressure
CART	=	Classification and Regression Trees
CLENGTH	=	Completed Length
DCA	=	Decline Curve Analysis
DFN	=	Discrete Fracture Network
DOE	=	Design of Experiments
D_i	=	Initial Decline Rate (Arps)
DTOF	=	Diffusive Time of Flight
EUR	=	Estimated Ultimate Recovery
FRAC_FLUID_TOTAL	=	Total Fracturing Fluid used for a well
FMM	=	Fast Marching Method
GA	=	Genetic Algorithm
γ or gamma	=	Shape parameter (Weibull Model)
GAM	=	Generalized Additive Model
GBM	=	Gradient Boosting Machine

GCV	=	Generalized Cross-Validation
GLUE	=	Generalized Likelihood Uncertainty Estimation
GOR	=	Gas-Oil ratio
LATITUDE	=	Latitude of a well's location
LHS	=	Latin Hypercube Sampling
LONGITUDE	=	Longitude of a well's location
LEFM	=	Linear Elastic Fracture Mechanics
m	=	Slope parameter (Duong Model)
M	=	Carrying capacity (Weibull)
MARS	=	Multivariate Adaptive Regression Splines
MD	=	Measured Depth
MSE	=	Mean Squared Error
n	=	Exponent parameter (SEDM)
NNET	=	Neural Networks
NPV	=	Net Present Value
OLS	=	Ordinary Least Squares
PKN	=	Perkins-Kern-Nordgren
PROP_TOTAL	=	Total proppant amount used for a well
q_i	=	Initial flow rate or Maximum Flow Rate
q_1	=	Flow rate during first month (Duong Model)
R^2	=	Coefficient of Determination

RI_p	=	Relative Variable Importance
${R^2}_p$	=	R^2 of a model using all predictors
${R^2}_{-p}$	=	R^2 of a model using all predictors except p^{th} predictor
RF	=	Random Forest
RMSE	=	Root Mean Squared Error
RSS	=	Residual Sum of Squares
SEDM	=	Stretched Exponential Decline Model
SVM	=	Support Vector Machine
SVR	=	Support Vector Regression
SRV	=	Stimulated Reservoir Volume
STAGE	=	Number of hydraulic fracture stages in a well
t	=	Time elapsed during well production
τ	=	Characteristic time (SEDM)
TOC	=	Total Organic Content
TVD	=	Total Vertical Depth
TVD_HEEL	=	Total Vertical Depth of horizontal well heel
TVD_HEEL_TOE_DIFF	=	Difference between TVDs of Heel and Toe
UFD	=	Unified Fracture Design

SUBSCRIPTS

f	=	fracture
i	=	initial condition
m	=	matrix
p	=	proppant
up	=	upstream

REFERENCES

Aizerman M.A., Braverman E.M., and Rozonoer L.I. 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25: 821–837.

Arps, J.J. 1945. Analysis of Decline Curves. Trans. AIME: 160: 228-247

Beven, K.J., and A. Binley. 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes 6, 279–298

Biswas, P., & Ley, S. B. (2015). Seismic Methodologies Adapted For Use In Acoustic Logging. Society of Petroleum Engineers. doi:10.2118/175995-MS

Breiman, L., 1996. Technical note: Some properties of splitting criteria. Machine Learning, 24(1), pp.41-47.

Breiman, L. 2001. "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32

Centurion, S.M., Cade, R. and Luo, X.L., 2012, January. Eagle Ford Shale: Hydraulic Fracturing, Completion, and Production Trends: Part II. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Centurion, S., Cade, R., Luo, X.L. and Junca-Laplace, J.P. 2013, September. Eagle Ford Shale: Hydraulic Fracturing, Completion and Production Trends, Part III. In SPE Annual Technical Conference and Exhibition.

Centurion, S., Junca-Laplace, J.P., Cade, R. and Presley, G., 2014, January. Lessons Learned From an Eagle Ford Shale Completion Evaluation. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Cheng, H., Dehghani, K., & Billiter, T. C. (2008). A Structured Approach for Probabilistic-Assisted History Matching Using Evolutionary Algorithms: Tengiz Field Applications. Society of Petroleum Engineers. doi:10.2118/116212-MS

Cipolla, C. L., Lolon, E., Erdle, J., & Tathed, V. S. (2009). Modeling Well Performance in Shale-Gas Reservoirs. Society of Petroleum Engineers. doi:10.2118/125532-MS

Cortes, C. and Vapnik, V. 1995. Support vector networks. Machine Learning 20: 273–297.

Cosma Shalizi. 2006. Statistics 36-350: Data Mining, Fall 2006 online lecture notes.

Daal, J. A., & Economides, M. J. (2006). Optimization of Hydraulically Fractured Wells in Irregularly Shaped Drainage Areas. Society of Petroleum Engineers. doi:10.2118/98047-MS

Datta-Gupta, A. and King, M. J., Streamline Simulation: Theory and Practice, Textbook Series #11, Society of Petroleum Engineers, Richardson, TX, ISBN 978-1-55563-111-6 (2007)

Datta-Gupta, A., Xie, J., Gupta, N. et al. 2011. Radius of Investigation and its Generalization to Unconventional Reservoirs. Journal of Petroleum Technology 63 (7): 52-55.

Dershowitz, B., LaPointe, P., Eiben, T., Wei, L. 2000. Integration of Discrete Feature Network Methods with Conventional Simulator Approaches. SPE Reservoir Eval. & Eng., 3 (2).

Draper, D. 1995. Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society: Series B 57, no. 1: 45–97.

Duong, A. N. 2011. "Rate-Decline Analysis for Fracture-Dominated Shale Reservoirs." SPEREE 14 (3): 377-387. http://dx.doi.org/10.2118/137748-PA.

Economides, M.J., Oligney, R.E. and Valko, P.P. "Unified Fracture Design". Orsa Press, Alvin TX, May 2002.

Economides, M.J., Hill, A.D., Ehlig-Economides, C. and Zhu, D. "Petroleum Production Systems". Second Edition. Prentice Hall, 2012.

Fan, L., Thompson, J. W., & Robinson, J. R. (2010). Understanding Gas Production Mechanism and Effectiveness of Well Stimulation in the Haynesville Shale through Reservoir Simulation. Society of Petroleum Engineers. doi:10.2118/136696-MS

Fetkovich, M.J. 1980. Decline Curve Analysis Using Type Curves. J PetTechnol 32 (6): 1065–1077.

Fisher, M.K., Wright, C.A., Davidson, B.M., Goodwin, A.K., Fielder, E.O., Buckler, W.S. and Steinsberger, N.P., 2005, January. Integrating fracture mapping technologies to improve stimulations in the Barnett Shale. SPE Productions and Facilities 20 (2): 85-93. doi: 10.2118/77441-PA

Friedman, J. H. 1991. Multivariate Adaptive Regression Splines. The Annals of Statistics. Vol. 19. No. 1: 1-141.

Friedman, J. H. 1993. Fast MARS Stanford University Department of Statistics, Technical Report 110.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, pp.1189-1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), pp.367-378.

Fujita, Y., Datta-Gupta, A. and King, M., 2016. A Comprehensive Reservoir Simulator for Unconventional Reservoirs That Is Based on the Fast-Marching Method and Diffusive Time of Flight. SPE Journal.

Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), pp.100-108.

Helgesen, T. B., Fulda, C., Meyer, W. H., Thorsen, A. K., Baule, A., Ronning, K. J., & Iversen, M. (2005). Accurate Wellbore Placement using a Novel Extra Deep Resistivity Service. Society of Petroleum Engineers. doi:10.2118/94378-MS

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical science, pp.382-401.

Holcomb, W.D., Lafollette, R.F. and Zhong, M., 2015, February. The Third Dimension: Productivity Effects From Spatial Placement and Well Architecture in Eagle Ford Shale

Horizontal Wells. In SPE Hydraulic Fracturing Technology Conference. Society of Petroleum Engineers.

Holditch, S. A. 2010. Shale Gas Holds Global Opportunities. The American Oil & Gas Reporter, August 2010 Editor's Choice.

Holland, J.H. 1992. Genetic Algorithms. Scientific American July 1992: 66-72.

Iino, A., Vyas, A., Huang, J., Datta-Gupta, A., Fujita, Y., Bansal, N. and Sankaran, S., April, 2017. Efficient Modeling and History Matching of Shale Oil Reservoirs Using the Fast Marching Method: Field Application and Validation. SPE Western Regional Meeting held in Bakersfield, California, USA

Iino, A., Vyas, A., Huang, J., Datta-Gupta, A., Fujita, Y. and Sankaran, S., July, 2017. Rapid Compositional Simulation and History Matching of Shale Oil Reservoirs Using the Fast Marching Method. Unconventional Resources Technology Conference held in Austin, Texas, USA

Ilk, D., Anderson, D. M., Stotts, G. W. J., Mattar, L., & Blasingame, T. (2010). Production Data Analysis--Challenges, Pitfalls, Diagnostics. Society of Petroleum Engineers. doi:10.2118/102048-PA

Johnston, D.C. 2006. Stretched Exponential Relaxation Arising From a Continuous Sum of Exponential Decays. Phys. Rev. B 74: 184430

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), pp.881-892.

Kaplan, S., 1981. On the method of discrete probability distributions in risk and reliability calculations–application to seismic risk assessment. Risk Analysis, 1(3), pp.189-196.

Kass, R.E., and A.E. Raftery. 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Kennedy, R. L., Gupta, R., Kotov, S. V., Burton, W. A., Knecht, W. N., & Ahmed, U. (2012). Optimized Shale Resource Development: Proper Placement of Wells and Hydraulic Fracture Stages. Society of Petroleum Engineers. doi:10.2118/162534-MS

Kim, J. U., Datta-Gupta, A., Brouwer, R., & Haynes, B. (2009). Calibration of High-Resolution Reservoir Models Using Transient Pressure Data. Society of Petroleum Engineers. doi:10.2118/124834-MS

Kulkarni, K. N., Datta-Gupta, A., & Vasco, D. W. (2000). A Streamline Approach for Integrating Transient Pressure Data into High Resolution Reservoir Models. Society of Petroleum Engineers. doi:10.2118/65120-MS

LaFollette, R.F. and Holcomb, W.D., 2011, January. Practical Data Mining: Lessons-Learned From the Barnett Shale of North Texas. Paper SPE 140524 presented at the Hydraulic Fracturing Technology Conference and Exhibition held in the Woodlands, Texas, USA, 24-26 January.

Lafollette, R., Holcomb, W.D. and Aragon, J., 2012, January. Impact of completion system, staging, and hydraulic fracturing trends in the Bakken Formation of the Eastern Williston Basin. In SPE Hydraulic Fracturing Technology Conference. Society of Petroleum Engineers.

Lafollette, R., Holcomb, W.D. and Aragon, J., 2012. Practical Data Mining: Analysis of Barnett Shale Production Results with Emphasis on Well Completion and Fracture Stimulation. Paper SPE 152531 presented at the SPE Hydraulic Fracturing Technology Conference, The Woodlands, Texas, USA, 6–8 February.

LaFollette, R.F. 2013. Shale Gas and Light Tight Oil Reservoir Production Results: What Matters?. *Proceedings of the Twenty-third (2013) International Offshore and Polar Engineering Conference.* International Society of Offshore and Polar Engineers, Anchorage, Alaska, USA, June 30 – July 5.

LaFollette, R.F., Izadi, G. and Zhong, M. 2013, February. Application of Multivariate Analysis and Geographic Information Systems Pattern-Recognition Analysis to Production Results in the Bakken Light Tight Oil Play. In SPE Hydraulic Fracturing Technology Conference. Society of Petroleum Engineers.

LaFollette, R.F., Izadi, G. and Zhong, M., 2014, February. Application of Multivariate Statistical Modeling and Geographic Information Systems Pattern-Recognition Analysis to Production Results in the Eagle Ford Formation of South Texas. In SPE Hydraulic Fracturing Technology Conference. Society of Petroleum Engineers.

Lee S.H., Kharghoria, A. and Datta-Gupta, A. 2002. Electrofacies Characterization and Permeability Predictions in Complex Reservoirs. *SPE Reservoir Evaluation and Engineering*, 5 (03), pp. 237-248.

Lee, W. J. Well Testing, Society of Petroleum Engineers, Richardson, TX (1982)

Ma, X., Plaksina, T., & Gildin, E. (2013). Optimization of Placement of Hydraulic Fracture Stages in Horizontal Wells Drilled in Shale Gas Reservoirs. Society of Petroleum Engineers. doi:10.1190/URTEC2013-151

Maxwell, S. C., Urbancic, T. I., Steinsberger, N., & Zinno, R. (2002). Microseismic Imaging of Hydraulic Fracture Complexity in the Barnett Shale. Society of Petroleum Engineers. doi:10.2118/77440-MS

Mishra, S. A New Approach to Reserves Estimation in Shale Gas Reservoirs Using Multiple Decline Curve Analysis Models. Paper SPE 161092 presented at the SPE Eastern Regional Meeting held in Lexington, Kentucky, USA, 3-5 October 2012.

Mishra, S., Choudhary, M.K. and Datta-Gupta, A., 2002. A novel approach for reservoir forecasting under uncertainty. SPE Reservoir Evaluation & Engineering, 5(01), pp.42-48. Mitchell, M. 1999.

An Introduction to Genetic Algorithms. The MIT Press, Cambridge, Massachusetts.

Morales, A. N., Nasrabadi, H., & Zhu, D. (2010). A Modified Genetic Algorithm for Horizontal Well Placement Optimization in Gas Condensate Reservoirs. Society of Petroleum Engineers. doi:10.2118/135182-MS

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stochastic Environmental Research and Risk Assessment, 17(5), pp.291-305.

Nilsson, N.J. 1965. Learning machines: Foundations of Trainable Pattern Classifying Systems. McGraw-Hill.

Oda, M. 1985. Permeability Tensor for Discontinuous Rock Masses. Geotechnique, 35 (4), pp. 483-495.

Perez, H.H., Datta-Gupta, A. and Mishra, S., 2005. The Role of Electrofacies, Lithofacies, and Hydraulic Flow Units in Permeability Prediction from Well Logs: A Comparative

Analysis Using Classification Trees. SPE Reservoir Evaluation & Engineering, 8(02), pp.143-155

Pitakbunkate, T., Yang, M., Valko, P. P., & Economides, M. J. (2011). Hydraulic Fracture Optimization with a p-3D Model. Society of Petroleum Engineers. doi:10.2118/142303-MS

Rankin, R.R., Thibodeau, M., Vincent, M.C. and Palisch, T., 2010, January. Improved production and profitability achieved with superior completions in horizontal wells: a bakken/three forks case history. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.

Riahi, A and Damjanac, B. (2013). Numerical Study of Interaction Between Hydraulic Fracture and Discrete Fracture Network. Proceedings of the International Conference for Effective and Sustainable Hydraulic Fracturing, Brisbane, Australia.

Saldungaray, P. M., Palisch, T., & Shelley, R. (2013). Hydraulic Fracturing Critical Design Parameters in Unconventional Reservoirs. Society of Petroleum Engineers. doi:10.2118/164043-MS

Savitski, A. A., Lin, M., Riahi, A., Damjanac, B., & Nagel, N. B. (2013). Explicit Modeling of Hydraulic Fracture Propagation in Fractured Shales. International Petroleum Technology Conference. doi:10.2523/17073-MS

Schuetter J., Mishra S., Zhong M. and LaFollette R. 2015. Data Analytics for Production Optimization in Unconventional Reservoirs. Paper SPE 178653-MS/URTeC:2167005 presented at the Unconventional Resources Technology Conference held in San Antonio, Texas. USA, 20-22 July.

Sehbi, B. S., Kang, S., Datta-Gupta, A., & Lee, W. J. (2011). Optimizing Fracture Stages and Completions in Horizontal Wells in Tight Gas Reservoirs Using Drainage Volume Calculations. Society of Petroleum Engineers. doi:10.2118/144365-MS

Sethian, J. A. 1996. A Fast Marching Level Set Method for Monotonically Advancing Fronts. Proceedings of the National Academy of Science 93:1591-1595.

Sethian, J. A., Level Set Methods and Fast Marching Methods, Cambridge University Press, New York City (1999).

Sierra, L., Mayerhofer, M., & Jin, C. J. (2013). Production Forecasting of Hydraulically Fractured Conventional Low-Permeability and Unconventional Reservoirs Linking the More Detailed Fracture and Reservoir Parameters. Society of Petroleum Engineers. doi:10.2118/163833-MS

Singh, A., Mishra, S. and Ruskauff, G., 2010. Model averaging techniques for quantifying conceptual model uncertainty. Ground Water, 48(5), pp.701-715.

Smola, A., J. and Schölkopf, B. 2004. "A tutorial on support vector regression." Statistics and Computing, vol.14, no. 3, pp. 199-222.

Song, B., & Ehlig-Economides, C. A. (2011). Rate-Normalized Pressure Analysis for Determination of Shale Gas Well Performance. Society of Petroleum Engineers. doi:10.2118/144031-MS

Valko, P.P. and Lee, J.W. 2010. A Better Way To Forecast Production From Unconventional Gas Wells. Paper SPE 134231 presented at the SPE Annual Technical Conference and Exhibition, Florence, Italy, 19-22 September.

Virieux, J., Flores-Luna, C. and Gibert, D., 1994. Asymptotic theory for diffusive electromagnetic imaging. Geophysical Journal International, 119(3), pp.857-868.

Vasco, D. W., Keers, H., and Karasaki, K. 2000. Estimation of Reservoir Properties Using Transient Pressure Data: An Asymptotic Approach. Water Resources Research 36 (12): 3447-3465.

Warpinski, N.R., Branagan, P.T., Peterson, R.E., Wolhart, S.L. and Uhl, J.E., 1998, January. Mapping hydraulic fracture growth and geometry using microseismic events detected by a wireline retrievable accelerometer array. In SPE Gas Technology Symposium. Society of Petroleum Engineers.

Warpinski, N.R., Kramm, R.C., Heinze, J.R. and Waltman, C.K., 2005. Comparison of Single-and Dual-Array Microseismic Mapping Techniques in the Barnett Shale. Paper SPE 95568 presented at the SPE Annual Technology Conference and Exhibition, Dallas, 9–12 October.

Weibull, W. 1951. A Statistical Distribution Function of Wide Applicability. J. Appl. Mech. 18: 293-297.

Xie, J., Yang, C., Gupta, N., King, M. J., & Datta-Gupta, A. (2015a). Depth of Investigation and Depletion in Unconventional Reservoirs With Fast-Marching Methods. Society of Petroleum Engineers. doi:10.2118/154532-PA

Xie, J., Yang, C., Gupta, N., King, M. J., Datta-Gupta, A. (2015b). Integration of Shale-Gas-Production Data and Microseismic for Fracture and Reservoir Properties With the Fast Marching Method. Society of Petroleum Engineers. doi:10.2118/161357-PA

Yang, C., Vyas, A., Datta-Gupta, A., Ley, S.B. and Biswas, P., 2017. Rapid multistage hydraulic fracture design and optimization in unconventional reservoirs using a novel Fast Marching Method. Journal of Petroleum Science and Engineering.

Yang, M., Valko, P.P. and Economides, M.J., 2012, March. Hydraulic Fracture Production Optimization with a Pseudo-3D Model in Multi-layered Lithology. In SPE/EAGE European Unconventional Resources Conference & Exhibition

Yin, J., Park, H., Datta-Gupta, A., & Choudhary, M. K. (2010). A Hierarchical Streamline-Assisted History Matching Approach With Global and Local Parameter Updates. Society of Petroleum Engineers. doi:10.2118/132642-MS

Yin, J., Xie, J., Datta-Gupta, A., & Hill, A. D. (2011). Improved Characterization and Performance Assessment of Shale Gas Wells by Integrating Stimulated Reservoir Volume and Production Data. Society of Petroleum Engineers. doi:10.2118/148969-MS

Zhang, Y., Yang, C., TETKing, M. J., & Datta-Gupta, A. (2013). Fast-Marching Methods for Complex Grids and Anisotropic Permeabilities: Application to Unconventional Reservoirs. Society of Petroleum Engineers. doi:10.2118/163637-MS

Zhang, Y., Bansal, N., Fujita, Y., Datta-gupta, A., King, M. J., & Sankaran, S. (2014). From Streamlines to Fast Marching: Rapid Simulation and Performance Assessment of Shale Gas Reservoirs Using Diffusive Time of Flight as a Spatial Coordinate. Society of Petroleum Engineers. doi:10.2118/168997-MS

Zhang, Y., Neha., B., Fujita, Y., Datta-Gupta, A., King., M. and Sankaran, S. 2016. "From Streamlines to Fast Marching: Rapid Simulation and Performance Assessment of Shale-Gas Reservoirs by Use of Diffusivity Time of Flight as a Spatial Coordinate." SPEJ 21 (5): 1-16. http://dx.doi.org/10.2118/168997-PA.

Zhong M., Schuetter J., Mishra and S. LaFollette. 2015. Do Data Mining Methods Matter? : A Wolfcamp “Shale” Case Study. Paper SPE 173334-MS presented at the SPE Hydraulic Fracturing Technology Conference held in The Woodlands, Texas, USA, 3-5 February.

APPENDIX A

This appendix describes how to regenerate figures and results presented in Chapter 2. This is a standalone R application code. A new user needs to copy the R code folder named as 'ML' in C drive keeping the names of this folder and all the subfolders unchanged.

Prerequisites: As a prerequisite R needs to be installed on the user computer. R Studio should be installed in order to edit code if needed. Also, some of the libraries needs to be installed before running code.

Following list of libraries need to be downloaded/installed: 'xlsx', 'GA', 'Metrics', 'randomForest', 'earth', 'e1071', 'MASS', 'glmnet', 'gbm', 'acepack', 'ggplot2', 'cvTools', 'neuralnet', 'class', 'maps', 'devtools', 'rpart.plot', 'FNN', 'reshape2'.

In order to install a library, go to R Studio menu bar and press Tools → Install Packages. A window should be opened up where the needed library can be installed.

Another way to install more than one package is through R commands. An R script file named as Install_Packages.R is provided with other R files. This file can be run in order to install all packages needed.

In case a library is needed but not installed, R Studio should generate error in console.

The contents of ML folder and their main job are:

1. **DCA_Well_Data**: This folder contains several excel sheets e.g., 'DCA_100.xlsx'. Each excel sheet belongs to a well and contains monthly rate data. The corresponding well API number is also provided in each file. This folder also contains well completion data of all wells in H_VAR_EXPORT_DCA.xlsx.

2. **Output_Files**: Output files of all the R script files are saved in this folder.

3. **DCA_FIT_ARPS.R**: This R script file reads the monthly rate data and completion

data for each of the study wells in DCA_Well_Data folder and fits Arp's decline curves. It fits the best decline model parameters ('Di' and 'b') and predicts the Estimated Ultimate Recovery (EUR) based on them. EUR is calculated for each well based on 30 years of production using decline curve extrapolation. Each well's initial flow rate (taken as maximum flow rate) is also identified for monthly rate data and referred to as 'qi' or initial flow rate in this study. Finally, the fitted decline model parameters and the corresponding completion data e.g., no. of stages, proppant amount, etc. (pulled from H_VAR_EXPORT_DCA.xlsx) for each well are stored in an excel sheet named as 'Model_data_ARPS.xlsx'. In this excel sheet, each row corresponds to a well identified by a serial number. Wells are identified by their unique serial number or well number. If needed, API number corresponding to a well serial number can be retrieved from a well's corresponding excel file in DCA_Well_Data folder. It should be noted here that those wells with less than 12 months of production history are not included in this study.

4. **DCA_FIT_SEDM.R**: This R script file has similar job to do as DCA_FIT_ARPS.R except that it is trying to fit SEDM parameters ('tau' and 'n') instead of Arp's parameters. EUR is also calculated based on extrapolated SEDM curve.

5. **DCA_FIT_DUONG.R**: This R script file has similar job to do as DCA_FIT_ARPS.R except that it is trying to fit Duong's model parameters ('a' and 'm') instead of Arp's parameters. EUR is also calculated based on extrapolated DUONG curve.

6. **DCA_FIT_WEIBULL.R**: This R script file has similar job to do as DCA_FIT_ARPS.R except that it is trying to fit Weibull's model parameters ('gamma', 'alpha' and 'M') instead of Arp's parameters. EUR is also calculated based on extrapolated WEIBULL curve.

7. **DCA_Data_Clean.R**: This R script file combines the output files of DCA_FIT_ARPS.R, DCA_FIT_SEDM.R, DCA_FIT_DUONG.R and

DCA_FIT_WEIBULL.R and generates a single file (Model_data.xlsx) which contains decline curve parameters for each well. This file also generates boxplots for distribution of various predictor variables in each of the 4 clusters clustered with respect to Initial flow rate, qi. This file also generates bubble plots for various predictor variables on Texas map. This file is also used to filter out outlier wells which have unrealistic predictor/response values.

8. **ML_Algorithms.R**: This R script file fits one or more machine learning algorithms selected by the user and builds models to predict decline model parameters. A user can change some of the parameters as discussed below:

Figs. A.1 and **A.2** shows the snapshots from R script file ML_Algorithms.R. These snapshots show where exactly a user can change inputs.

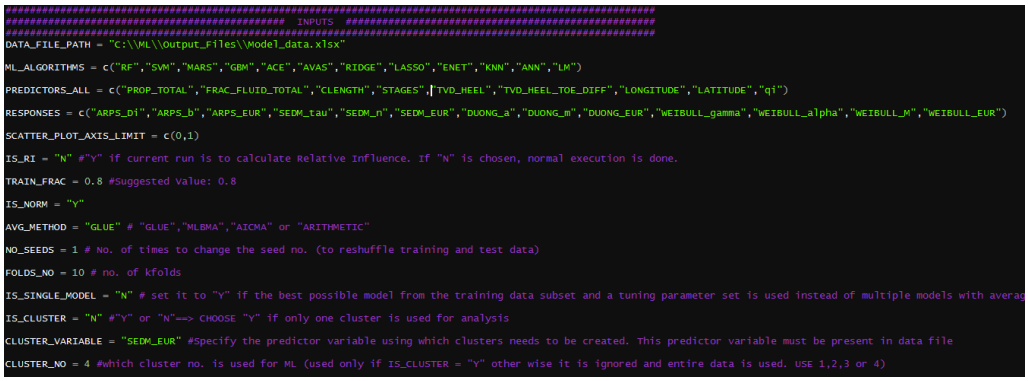


Figure A.1 Input parameters in ML_Algorithms.R script – Part 1

```
# TUNING PARAMETERS FOR RF
NTREE = 300
MTRY_SEQ_VALUES = seq(from = 1,to = 9,by = 1)

# TUNING PARAMETERS FOR SVM
KERNEL_TYPES = c("linear","radial","polynomial")
COST_VALUES = seq(from = 0.1,to = 3,by = 0.1)

# TUNING PARAMETERS FOR MARS
DEGREE_VALUES = seq(from = 1,to = 3,by = 1)

# TUNING PARAMETERS FOR RIDGE/LASSO
LAMBDA_VALUES = seq(from = 0,to = 0.01,by = 0.0001)

# TUNING PARAMETERS FOR ENET
ALPHA_VALUES = seq(from = 0.1,to = 0.9,by = 0.1)

# TUNING PARAMETERS FOR GBM
NTREES_VALUES = seq(from = 1e+4,to = 3e+4,by = 1e+4)

# TUNING PARAMETERS FOR ANN
HIDDEN_VALUES = seq(from = 9,to = 20,by = 3) #seq(from = 3,to = 10,by = 1)
HIDDEN_LAYERS_VALUES = seq(from = 1,to = 3,by = 1) #no. of hidden layers
THRESHOLD_VALUES = seq(from = 0.1,to = 10,by = 1) #seq(from = 0.1,to = 10,by = 0.3)

# TUNING PARAMETERS FOR KNN
KNN_VALUES = seq(from = 1,to = 10,by = 1)

#TUNING PARAMETERS FOR LM
MAX_TERMS_VALUES = c(20,30,1) #max. no. of terms allowed in the linear model

################################################################################
########################################## END OF INPUT SECTION ################
################################################################################
```

Figure A.2 Input parameters in ML_Algorithms.R script – Part 2

The explanation of various variables and their possible values are provided below:

DATA_FILE_PATH

This variable assigns the path of excel sheet containing all predictors and responses that are needed various for machine learning algorithms. For e.g., for the current settings, it is set to "C:\\ML\\Output_Files\\Model_data.xlsx".

ML_ALGORITHMS

This variable assigns type of machine learning algorithm used. One or more algorithms can be run at a time. E.g., c("RF", "SVM", "MARS") would run code for RF, SVM and MARS in that order. In total, 12 machine learning algorithms are allowed.

Suggested Values: one or more of "RF", "SVM", "MARS", "GBM", "ACE", "AVAS", "RIDGE", "LASSO", "ENET", "KNN", "ANN", "LM"

Above acronyms stand for following machine learning algorithms:

RF: Random Forest

SVM: Support Vector Machine

MARS: Multivariate Adaptive Regression Splines

GBM: Gradient Boosting Machine

ACE: Alternative Conditional Expectations

AVAS: Additivity Variance Stabilization

RIDGE: Ridge Regression

LASSO: Least Absolute Shrinkage and Selection Operator

ENET: Elastic Net regression

KNN: K-Nearest Neighbors

ANN: Artificial Neural Network

LM: Linear Model

PREDICTORS_ALL

This variable assigns the list of predictor variables. These variables must be present in the data file – "Model_data.xlsx".

Suggested Values: For Chapter 1 study it is set to c("PROP_TOTAL", "FRAC_FLUID_TOTAL", "CLENGTH", "STAGES", "TVD_HEEL", "TVD_HEEL_TOE_DIFF", "LONGITUDE", "LATITUDE", "qi")

RESPONSES

Response variable to be predicted. Can be one or more variables.

Suggested values:

For ARPS, it can be set to "ARPS_Di", "ARPS_b" or "ARPS_EUR". Multiple response variables can be predicted as in c("ARPS_Di", "ARPS_b", "ARPS_EUR")

For SEDM, it can be set to "SEDM_tau", "SEDM_n" or "SEDM_EUR". Multiple responses can be predicted as in c("SEDM_tau", "SEDM_n", "SEDM_EUR")

For DUONG, it can be set to “DUONG_a”, “DUONG_m” or “DUONG_EUR”. Multiple responses can be predicted as in c(“DUONG_a”, “DUONG_m”, “DUONG_EUR”)

For WEIBULL, it can be set to “WEIBULL_gamma”, “WEIBULL_alpha”, “WEIBULL_M” or “WEIBULL_EUR”. Multiple responses can be predicted as in c(“WEIBULL_gamma”, “WEIBULL_alpha”, “WEIBULL_M”, “WEIBULL_EUR”)

SCATTER_PLOT_AXIS_LIMIT

This variable sets minimum and maximum limits for the response variable for which model training is being done.

Suggested Values: In the Eagle Ford study case, following values for this variable has been used.

Table A.1 Axis scale values used for Eagle Ford plots

DCA Model	Response	SCATTER_PLOT_AXIS_LIMIT
ARPS	ARPS_Di	c(0,1)
	ARPS_b	c(0,1)
	ARPS_EUR	c(0,300)
SEDM	SEDM_tau	c(0,20)
	SEDM_n	c(0,2)
	SEDM_EUR	c(0,300)
DUONG	DUONG_a	c(0,3)
	DUONG_m	c(1,3)
	DUONG_EUR	c(0,300)
WEIBULL	WEIBULL_gamma	c(0,1)
	WEIBULL_alpha	c(0,20)
	WEIBULL_M	c(0,2e+5)
	WEIBULL_EUR	c(0,300)

IS_RI

If this script needs to be run for calculation of relative importance of various predictor variables, this variable is set to “Y” otherwise “N”. In case the current run is for Relative Influence calculations, ACE and AVAS algorithms need to be removed from the set of ML_ALGORITHMS in input sections. For ML_ALGORITHMS, use one or more of “RF”, “SVM”, “MARS”, “GBM”, “RIDGE”, “LASSO”, “ENET”, “KNN”, “ANN”, “LM”

TRAIN_FRAC

The fraction of data points used for training purpose. The rest will be used for testing the machine learning model.

Suggested values: 0.8

IS_NORM

This variable decides whether the data needs to be normalized for learning or not. The final predictions are stored after de-normalizing the data.

Suggested Values: Choose “Y” if data needs to be normalized and choose “N” otherwise.

Note: If using “ANN” as the machine leaning algorithm, normalizing the data is necessary. Therefore, IS_NORM should be set to “Y” if one of the machine learning algorithms is “ANN”.

AVG_METHOD

This variable assigns the type of averaging algorithm used

Suggested values: “GLUE”, “MLBMA”, “AICMA” or “ARITHMETIC”

Each of these averaging keywords stand for different ways of assigning model weights to be used for model averaging:

GLUE: Generalized Likelihood Uncertainty Estimation

MLBMA: Maximum Likelihood Bayesian Model Averaging

AICMA: Akaike Information Criterion Model Averaging

ARITHMETIC: All models are assigned equal weights. The averaging is based on arithmetic average of all models.

NO_SEEDS

This variable assigns the number of seeds used to reshuffle the given training dataset. Reshuffling the data would give different data points in k-folds that are generated during model building and will generate extra set of models in model pool.

FOLDS_NO

This variable assigns the number of folds into which the training data is split. If this number is high, smaller sets of data would lie in each fold. On the other hand, smaller values would split training data into bigger sets of data in each fold.

Suggested Value: This is set to 5 or 10 most commonly. In the current settings, it is set to 10.

IS_SINGLE_MODEL

This variable indicates whether model averaging needs to be done or not. If it is set to "Y", then only the best model is used for final prediction for test data. If it is set to "N", then model averaging is done with corresponding weights of each model.

IS_CLUSTER

This variable decides if a machine learning is to be done for a particular cluster or not. If it is set to "Y" data is divided into 4 clusters based on the variable name specified.

Suggested Values: "Y" or "N".

CLUSTER_VARIABLE

The variable to be used to partition data into 4 clusters based on quartiles. This is useful only if IS_CLUSTER is set to "Y".

Suggested Values: This is assigned to one of the predictor variables, for e.g., "qi".

CLUSTER_NO

This variable assigns the cluster number to be used for machine learning. This variable is useful only if IS_CLUSTER is set to "Y" otherwise it is ignored and entire dataset is used.

Suggested Values: 1, 2, 3 or 4

NTREE

This variable is a tuning parameter for Random Forest model which is equal to the number of trees used.

Suggested values: Usually a large number will help dealing with overfitting. For Eagle Ford data, NTREE = 300 has been used.

MTRY_SEQ_VALUES

This variable is a tuning parameter for Random Forest and gives sequence of options for number of predictor variables to be considered to partition data at each node of a tree in Random Forest.

Suggested values: It is suggested to use all possible subsets of predictor variables. In Eagle Ford data, since there are 9 predictors, MTRY_SEQ_VALUES is set to seq(from = 1,to = 9,by = 1) giving all possible subsets of predictor variables to be used at each node.

KERNEL_TYPES

This variable is a tuning parameter for SVM and assigns the kernel type(s) to be used for SVM learning.

Suggested values: One or more of "linear", "radial" and "polynomial" kernels are suggested. In current settings all of them are assigned as a sequence - c("linear", "radial",

“polynomial”). Multiple kernel types can be used for building multiple models for model averaging.

COST_VALUES

This is a tuning parameter for SVM. It assigns the cost parameter for SVM. Changing cost value can reduce overfitting.

Suggested values: In current settings, a sequence of cost values are provided as seq(0.1,3,0.1) ranging between 0.1 and 3.0 in steps of 0.1.

DEGREE_VALUES

This variable is a tuning parameter for MARS. It sets possible degree values for MARS model. Degree in MARS model controls the maximum degree of interaction. If degree is set to 1, no interaction terms are included, i.e., an additive model is built.

Suggested values: In the current settings a range of degree values are given - seq(from = 1,to = 3,by = 1). Therefore degree can be 1, 2 or 3.

LAMBDA_VALUES

This is a tuning parameter for Ridge and LASSO regression models. This variable assigns values for lambda which controls model regularization term.

Suggested values: In the current settings, it is in the sequence from 0 to seq(from = 0,to = 0.01,by = 0.0001)

ALPHA_VALUES

This variable is a tuning parameter for Elastic Net (ENET) regression and assigns one or more values for the Elastic Net mixing parameter, alpha.

Suggested values: In case of Elastic Net regression, alpha should lie between 0 and 1. In current settings, it is within a range of 0.1 and 0.9 in the steps of 0.1, i.e., seq(from = 0.1,to = 0.9,by = 0.1). In case alpha is set to 0, the model becomes Ridge regression and if alpha

is set to 1, model becomes LASSO regression. In case of Ridge or LASSO regression, the corresponding alpha values are automatically used by the code and this variable is ignored.

NTREES_VALUES

This is a tuning parameter for GBM and it assigns the number of trees to fit. A single value or a sequence of values may be provided.

Suggested values: In the current settings, this variable is assigned to a range of values from 10000 to 30000 in steps of 10000, i.e., seq(from = 10000,to = 30000,by = 10000).

HIDDEN_VALUES

This is a tuning parameter for ANN model. This variable assigns number of neurons in a hidden layer. It may be a single value or a sequence of possible values.

Suggested Values: In the current settings, this variable is set to have a sequence of possible values ranging from 9 to 30 in steps of 3, i.e., seq(from = 9, to = 20, by = 3). A large number of neurons may lead to over fitting.

HIDDEN_LAYERS_VALUES

This variable assigns the number of hidden layers in ANN network. In current code settings, each hidden layer is set to be of equal number of neurons.

Suggested Values: In the current settings, it is set to a sequence from 1 to 3 in steps of 1s, i.e., seq(from = 1, to = 3, by = 1). Larger number of layers may lead to over fitting.

THRESHOLD_VALUES

This variable assigns the threshold value for the partial derivatives of the error function as stopping criteria. A small value may over fit model.

Suggested Values: In current settings, threshold is set to a range of values ranging from 0.1 to 10 in steps of 1.

KNN_VALUES

This is a tuning parameter for KNN regression. This variable assigns the number of nearest neighbors considered.

Suggested Values: In current settings, a range of values from 1 to 10 in steps of 1 are used, i.e., seq(from = 1, to = 10, by = 1).

MAX_TERMS_VALUES

This is a tuning parameter for LM (linear model) fitting. This variable sets maximum number of terms in a linear model including interaction terms. In current code, up to three way interactions are considered.

Suggested Values: More number of terms are likely to over fit model. In the current code settings, it is set to a range of values from 20 to 30 in steps of 1, i.e., seq(from = 20, to = 30, by = 1).

9. **DCA_Decline_Curves.R**: This R script file plots the test data well decline curves against actual rate data. In the input section of this R script file, user needs to specify values for following variables:

DCA_METHOD

This variable assigns the decline model for which plots need to be generated.

Suggested Values: One of the decline models - “ARPS”, “SEDM”, “DUONG” or “WEIBULL”

ML_ALGORITHM

The machine learning algorithm for which the decline model predictions have to be plotted.

Suggested Values: One of the following algorithms:

“RF”, “SVM”, “MARS”, “GBM”, “ACE”, “AVAS”, “RIDGE”, “LASSO”, “ENET”, “KNN”, “ANN”, “LM”

IS_CLUSTER

If the learning was done for each cluster, decline models would be plotted for each cluster separately.

Suggested Values: "Y" or "N"

10. **ERR_PLOTS_RELATIVE.R**: This R script file plots the error bar plots for training and test data predictions. Error plots are based on normalized RMSE, AAE or R^2 errors relative to the maximum value among all algorithms under investigation. Following input variables need to be set before running this script.

ML_ALGORITHMS

This variable assigns the list of machine learning algorithms that need to be included in error bar plots. Corresponding machine leaning algorithms need to be run before including them in this list.

Suggested Values: One or more of following machine learning algorithms:
"RF", "SVM", "MARS", "GBM", "ACE", "AVAS", "RIDGE", "LASSO", "ENET", "KNN", "ANN", "LM"

RESPONSE

This variable needs to be assigned to the response variable for which error bars need to be compared for different machine learning algorithms.

Suggested Values: E.g., "SEDM_EUR", "ARPS_EUR", etc.

11. **ERR_PLOTS.R:** This file does the same job as ERR_PLOTS_RELATIVE.R except that it creates bar plots based on un-normalized errors.

12. **RI_PLOTS**: This R script file needs to be executed in order to generate relative influence plots for the current study. Following variables need to be set before running this file.

ML_ALGORITHMS

This variable needs to be set to a list of machine learning algorithms which need to be included in relative influence.

Suggested Values: One or more of following machine learning algorithms:

“RF”, “SVM”, “MARS”, “GBM”, “ACE”, “AVAS”, “RIDGE”, “LASSO”, “ENET”, “KNN”, “ANN”, “LM”

RESPONSES

This variable is assigned to the list of variables that need to be included in relative influence plots.

Suggested Values: For. e.g., c(“ARPS_EUR”, “SEDM_EUR” ,”DUONG_EUR”)

RANKING_POLICY

This variable is assigned to the metric type used to calculate relative influence of a variable.

Suggested Values: One of “RMSE_Test”, “AAE_Test” or “R2_Test”.