Practical Integration of Data-Driven Models for Production Analysis and Inference of Reservoir Heterogeneities in SAGD Operations

by

Zhiwei Ma

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

PETROLEUM ENGINEERING

Department of Civil and Environmental Engineering
University of Alberta

Abstract

Steam-assisted gravity drainage (SAGD) technique has been widely adopted for heavy oil production. However, SAGD performance is strongly affected by reservoir heterogeneities, such as shale barriers and lean zones, as they are often detrimental to SAGD production efficiency. Therefore, it is necessary to characterize such reservoir heterogeneities for practical SAGD operations. Conventional characterization workflows, which entail construction of prior reservoir models, numerical flow simulation, and history-matching process, have some apparent limitations, such as high computational demands, as well as the involvement of numerous assumptions and simplifications regarding the underlying physical processes. In addition, the rapidly increasing volumes of SAGD field data from public domains provide fundamental information pertinent to reservoir properties and production characteristics. It is of great interest to propose a feasible SAGD analysis alternative that is capable of utilizing these field data for production analysis and heterogeneities characterization. Data-driven modeling techniques, which involve data analytics and implementation of artificial intelligence (AI) methods for capturing internal structures and non-linear relationships among data, are customized here to address this challenge.

This thesis will develop a set of workflows suitable for prediction of SAGD production performance and inference of heterogeneities by means of data analytics and production data analysis. First, through a comprehensive analysis of field data, a workflow is developed to forecast SAGD production. Data-driven models are built as a proxy model of reservoir simulation process to approximate the forward relationship between SAGD production and reservoir parameters. The forecast performances of these trained models are shown to be both

reliable and satisfactory. Next, a series of synthetic SAGD models based on typical Athabasca oil reservoir properties and operating conditions is constructed. Heterogeneities are modeled by randomly sampling distribution, volume, and orientation of shale barriers and lean zones from several probability distributions inferred from field data, and are superposed to the base homogenous models. Many parameterization schemes are investigated to extract input and output parameters from production time-series data and heterogeneous configurations, respectively. Data-driven models are constructed to approximate the inverse relationship between reservoir characteristics and production data, thus to infer the complex reservoir heterogeneities stemmed from shale barriers and lean zones. The developed models can reliably estimate the relevant shale and lean zone parameters and the associated uncertainties. The proposed methods facilitate the selection of an ensemble of reservoir models that are consistent with the production history of the true models.

In the thesis, data-driven models are constructed used artificial neural network (ANN). Techniques such as principal component analysis, clustering analysis, and wavelet transform are employed to process the data and to improve the model robustness. The outcomes would improve our ability to infer uncertain reservoir heterogeneities from SAGD production data. It offers a complementary tool for extracting additional information from field data and incorporating data-driven models into existing simulation and history-matching workflows. The developed workflow can potentially be extended to analyze other engineering datasets derived from various sources and integrated directly into existing reservoir management and decision-making routines.

Preface

A version of chapter 3 has been published as Ma, Z., Leung, J. Y., Zanon, S., & Dzurman, P. (2015), Practical implementation of knowledge-based approaches for steam-assisted gravity drainage production analysis*, Expert Systems with Applications, 42*(21), 7326-7343. I was responsible for the data collection and analysis as well as the manuscript composition. Zanon, S. and Dzurman, P. contributed to the manuscript edits. Leung, J. Y. was the supervisory author and was involved in concept formation and manuscript composition.

A version of chapter 4 has been published as Ma, Z., Leung, J. Y., & Zanon, S. (2017), Practical data mining and artificial neural network modeling for SAGD production analysis, *Journal of Energy Resources Technology, 139*(3), 032909. I was responsible for the data collection and analysis as well as the manuscript composition. Zanon, S. contributed to the manuscript edits. Leung, J. Y. was the supervisory author and was involved in concept formation and manuscript composition.

A version of chapter 5 has been published as Ma, Z., Leung, J. Y., & Zanon, S. (2018). Integration of artificial intelligence and production data analysis for shale heterogeneity characterization in steam-assisted gravity-drainage reservoirs, *Journal of Petroleum Science and Engineering, 163*, 139-155. I was responsible for the data collection, model construction, and analysis as well as the manuscript composition. Zanon, S. contributed to the manuscript edits. Leung, J. Y. was the supervisory author and was involved in concept formation and manuscript composition.

A version of chapter 6 will be submitted for publication as Ma, Z. & Leung, J., Y., Integration of data-driven modeling techniques for lean zone and shale barrier characterization in SAGD reservoirs. I was responsible for the data collection, model construction, and analysis as well as the manuscript composition. Leung, J. Y. was the supervisory author and was involved in concept formation and manuscript composition.

Chapter 2 summarizes the important materials, model setups and methodologies from chapters 3 to 6. Chapters 1, 7, and 8 are originally written by Zhiwei Ma and have never been published before.

Dedicated to my family
for their love, endless support and encouragement!

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Juliana Leung, for giving me the opportunity to work on this thesis work under her guidance. It is her excellent teaching, enlightening guidance, continuous support and encouragement that helped me throughout the course of my Ph.D. study and finished this thesis.

I also want to thank my Ph.D. final defense committee, Dr. Andy Li, Dr. Jeff Boisvert, Dr. Vinay Prasad and Dr. Farshid Torabi for attending my final defense and providing insightful suggestions and comments regarding my thesis.

I thank all my fellow group members of Dr. Leung's research group for the stimulating discussions and all my friends for the support and encouragement.

I would like to thank my mother Qinghua Meng, my father Jun Ma, and my brother Zhiqiang Ma for their continuous encouragement.

Most importantly, I want to thank my wife Chunling Long and my son Leo Ma for the support, help, sacrifice, and love throughout my studies.

This financial support from the Nexen Energy ULC, and the collaborative research and development grant program administered by Natural Sciences and Engineering Research Council of Canada (NSERC) is appreciated. I also want to thank the University of Alberta for granting access to the Numerical and Statistical Server and the Computer Modelling Group Ltd. for the academic license of CMG software.

Table of Contents

List of Tables

List of Figures

List of Symbols

b	=	bias term for ANN
b_i	=	bias term of the neuron i
B_i	=	bias term of neuron i
B_b	=	bootstrap sample size
cA	=	approximation coefficients of DWT
cD	=	detail coefficients of DWT
d	=	shortest distance between the horizontal well pair and the vertical well
$\mathbf{d}$	=	data vector
d_{sh_inj}	=	distance from the shale layer to the injection well
D	=	the shortest distance between the left bottom corner of the heterogeneity and the injector
D_d	=	dimensionless form of D
D_{max}	=	maximum value of D
e	=	error term in linear and nonlinear regression model
f	=	activation function of neural network
$\hat{f}$	=	predicted value from a model
g	=	number of groups for k-means algorithm
h	=	pay zone thickness
h_{sh}	=	shale thickness
H	=	shortest horizontal distance between a shale barrier and the left boundary of the modeling domain
H_d	=	dimensionless form of H
I_{ij}	=	data sample group indicator for k-means algorithm
$\Delta iSOR$	=	*iSOR* difference between a heterogeneous case and the homogenous base case
k	=	number of test samples for neural network
L	=	length of shale barrier

L_d = dimensionless form of L

m = total number of neurons in the preceding layer of network

m_{inj} = number of production wells in the well pair

m_{prod} = number of production wells in the well pair

n = number of folds in n-fold cross-validation routine

N = total number of data records

N_h = number of heterogeneity

N_e^{inj} = effective number of injection wells in a certain well pair

N_e^{prod} = effective number of production wells in a certain well pair

N_i^{inj} = normalized number of injector i

N_i^{prod} = normalized number of producer i

N_l = number of lean zones in the reservoir

N_r = number datasets sampled from parametric bootstrapping

N_s = number of shale barriers in the reservoir

N/G = net-to-gross ratio

p = number of predictor variables in linear and nonlinear regression models

P = probability

PC = principal components

PS = principal scores

q = oil production rate

Δq = oil rate difference between a heterogeneous case and the homogenous base case

Q = cumulative oil production

R^2 = coefficient of determination

RF = random forests

s = scaling parameter of a wavelet function

S = the surface distance between two adjacent well pairs

S_c = within-cluster sum of squares.

S_o = oil saturation

S_w = water saturation

t = cumulative production time

$\bar{T}$ = average of all targets

T_i^{inj} = effective injection time of the i_{th} injection well during the total production period of the well pair

T_i^{prod} = effective production time of the i_{th} production well during the total production period of the well pair

T_{total} = total production period of a given well pair

$\mathbf{w}$ = model parameter vector

$w_{i,j}$ = weight between node i in current layer and the nodes j in previous layer

$\bar{X}_j$ = average value of the j_{th} dimension of the original dataset

$\mathbf{x_n}$ = input vector

x^N = normalized data

x_i = input value of node i in the current layer of ANN

x_{max} = maximum value of original data before normalization

x_{min} = minimum value of original data before normalization

X_{ij} = variable in the original dataset

$\mathbf{y_n}$ = output vector

y_i = output signal from neuron i in current layer of network

Y = response or target of a linear and nonlinear regression model

Z_{ij} = mean-adjusted variable

Greek letters:

α = coefficient associated with a nonlinear regression model

β = coefficient associated with a linear regression model

ε = Gaussian white noise

λ = number of targets or response variables in linear and nonlinear model

σ = variance

τ = translation parameter of a wavelet function

φ = porosity

Ψ = wavelet function

Acronyms

AI	=	artificial intelligence
ANN	=	artificial neural network
BPNN	=	back-propagation neural network
COP	=	cumulative oil production.
CSI	=	cumulative steam injection
cSOR	=	cumulative steam-oil-ratio
CSS	=	cyclic steam stimulation
CWT	=	continuous wavelet transform
DWT	=	distract wavelet transform
EDA	=	exploratory data analysis
EnKF	=	ensemble Kalman filter
EOR	=	enhanced oil recovery
GA	=	genetic algorithm
GR	=	gamma ray logging
iSOR	=	instantaneous steam-to-oil ratio
LM-BP	=	Levenberg–Marquardt backpropagation
MLP	=	multi-layer perceptron
MSE	=	mean squared error
PCA	=	principal component analysis
PDF	=	probability density function
PSO	=	particle swarm optimization
RF	=	random forest
RT	=	true resistivity logging
SAGD	=	steam-assisted gravity drainage
SF	=	steam flooding
SI	=	shale index
SLP	=	single-layer perceptron
SOR	=	steam-oil-ratio
SP	=	spontaneous potential logging

Chapter 1 Introduction

1.1 Background

Heavy oil and bitumen, with extremely high viscosity and density, are important fuel resources for global energy consumption. According to Hein (2017), bitumen and heavy oil resources are distributed in over 70 countries and the estimated reserve is around 5.6 trillion barrels. Canada has large heavy oil and bitumen resources, most of which are located in Alberta. According to the Alberta Energy Regular (2016), the bitumen reserve in Alberta is approximately 165 billion barrels, as of 2015. However, the total extracted volume is less than 10% (only 11 billion barrels) since the late 1960s.

Unlike other conventional oil, the viscosity of bitumen is extremely high (larger than 10,000 cP), which inhibits its flow under reservoir conditions. Therefore, traditional oil recovery technologies, e.g., water flooding, cannot be employed for effective bitumen or heavy oil production. To conquer this difficulty, many thermal enhanced oil recovery (EOR) techniques have been invented and applied in heavy oil development, such as steam flooding (SF), cyclic steam stimulation (CSS), and steam-assisted gravity drainage (SAGD).

SAGD originally invented by Dr. Butler and his colleagues in Imperial Oil in the 1970s is one of the most established and important thermal EOR techniques to produce bitumen in Alberta. As shown in **Fig. 1-1**, two parallel horizontal wells (one injection well and one production well that is located a few meters beneath the former) are drilled into the bottom of a target formation. Pressurized high-temperature steam is continuously injected to form a steam chamber and to heat the bitumen; the viscosity of the heated oil is reduced and would drain to the corresponding producer along the edge of the formed steam chamber by the force of gravity, as depicted in **Fig. 1-2**. Since its inception (Butler et al., 1981), SAGD has been extensively employed for commercial bitumen production.

Although SAGD technique has achieved great success in oil and gas industry, many geologic challenges limit its practical application, such as reservoir heterogeneities due to shale barriers and high water zones, also known as lean zones. These two types of heterogeneities are commonly-observed in many SAGD projects in Alberta, especially in Nexen's Long Lake and Suncor's Firebag projects. A heterogeneous SAGD reservoir containing three regions of lean zones and two regions of lean zones is presented in **Fig. 1-3**. Shale barriers or lean zones are often detrimental for conventional bitumen recovery using SAGD techniques. It has been widely documented that the presence of shale barriers could pose significant adverse effects on SAGD production: impeding of the vertical growth and lateral spread of the steam chamber, hampering inter-well communication, and reducing instantaneous oil rate (Sheng, 2013; Yang and Butler, 1992); while the existence of high water saturation zone, would significantly impact the performance of SAGD by reducing steam utilization and increasing steam-oil-ratio (SOR). Therefore, it is crucial to determine the presence of shale barriers and lean zones and characterize their geological properties for reliable production forecast, effective reservoir management, and optimization of development strategies.

1.2 Conventional Heterogeneities Characterization Methods

Traditionally, to characterize reservoir heterogeneities, the first step is extracting static (geologic) and dynamic (flow) data. Static data, by definition, don't change with time; while dynamic data often refer to time-series data, defined as a sequence of data points which are measured and recorded in a time interval. Log interpretation, core analysis, and pressure transient analysis from certain wells are commonly-adopted methods for measuring such data. Given that these data are typically sampled only at a few locations, some stochastic modeling approaches have to be applied to create multiple reservoir realizations that honor the available data in order to calculate uncertainty. Among these approaches, geostatistical methods, e.g., sequential simulation can be used to infer inter-well reservoir properties (Chopra et al., 1990) and to generate multiple realizations.

The next step is combining static and dynamic data for characterization. Conventional workflows construct prior reservoir models using only static data, which are subjected to numerical flow simulation and history-matching processes, during which dynamic data is

integrated. It is an important technique to model multiphase flow in porous media and to characterize reservoir heterogeneities, and has been widely used in reservoir engineering. The goal of history-matching is to assign appropriate values to uncertain parameters of geological models through integrating of actual past production measurements, such that these tuned geological models are capable of reproducing the past production performances and making reliable production predictions. Most history-matching techniques can be classified into five categories: manual adjustment (Williams et al., 1998), gradient-based approach (Zhang et al., 2003), evolutionary method (Romero and Carter, 2001), ensemble filtering (Gu and Oliver, 2005) and gradual deformation (Hu, 2000).

The conventional heterogeneities characterization workflows offer certain advantages. First of all, they are capable of providing a detailed production match by adjusting reservoir variables, if a large number of grid blocks are employed. This is very useful to capture the miniature characteristics of heterogeneity. Secondly, parallel computation techniques can be applied to some of the automatic history-matching algorithms, e.g., evolutionary approach. This would significantly speed up the history-matching workflows and incur substantial savings in computational time. Thirdly, it is possible to quantify the uncertainty in future production performance by considering multiple realizations of the unknown parameters. Given the dimensionality of the unknown model parameter is usually larger than independent data, the history-matching procedure would result in an ill-posed inverse problem: more than one solution of model parameters that are consistent with the production history may be obtained (Oliver and Chen, 2011).

Since SAGD was first implemented for practical heavy oil production, numerous studies have been carried out to characterize reservoir heterogeneities using history-matching methods. A stochastic optimization approach was implemented to estimate horizontal permeability, initial oil situation, and porosity for a homogeneous synthetic model in Jia et al. (2009). Mirzabozorg et al. (2013) applied an evolutionary algorithm to estimate reservoir properties including the length of a single shale layer located at a fixed depth. To represent the complex shale barrier configurations in reservoir models, various model parameterization schemes have been adopted. In Hiebert et al., (2013), steam chamber inferred from seismic data was utilized in history-matching to identify geo-bodies. In Panwar et al. (2012), formation facies were perturbed during history-matching of well temperature data, during which the Ensemble Kalman Filter (EnKF)

and discrete cosine transform were implemented. Although these assisted history-match techniques are quite robust for integration of a wide range of production data (rate, temperature, saturation, etc.), they are usually computationally intensive (Mirzabozorg et al., 2013; Zhang et al., 2014). Additionally, certain assumptions and simplifications regarding the operational constraints, governing equations, and process physics must be invoked in the simulation model. Previous works that focus specifically on the inference of stochastic heterogeneous distributions of shale barrier and lean zone are rare in literature.

1.3 Issues in Conventional Heterogeneities Characterization Methods

As mentioned above, many issues and challenges limit conventional methods for practical heterogeneities characterization in SAGD reservoirs. These issues and challenges are summarized as:

a) High cost in computational time: Conventional characterization approaches for dynamic data integration (i.e., production history-matching) are usually computationally-expensive (Oliver and Chen, 2011; Shahkarami et al., 2015). The computational time includes the time required for data extraction and analysis, construction of complex reservoir model, and numerical flow simulation. For instance, simulations that involve high-resolution heterogeneity details and sharp transition in rock/transport properties would generally require an increased number of reservoir grids and time steps, which in turn, require a longer time for data preparation, model simulation, and result interpretation (Mattax and Dalton, 1990).

b) Challenges in reparameterization of unknown variables: The dimensionality of the unknown model vector is generally quite large. Various reparametrization schemes are often employed to reduce the dimensionality of the model vectors. Besides the improvement in computational efficiency, dimensionality reduction also reduce the size of the model null space (Oliver and Chen, 2011). However, formulating appropriate reparameterization of the unknown parameters remains challenging: it is highly problem dependent since there is no universal guideline for selecting the number or types of variables. Using a small number of variables would underestimate the

uncertainties in the final predictions while selecting too many parameters would not significantly reduce the computational time (Oliver and Chen, 2011).

c) Involvement of many simplifications and assumptions in numerical flow simulation: To solve the governing partial differential equations (PDE) of numerical models, many simplifications and assumptions regarding the physical phenomena must be invoked. For instance, finite volume/difference/element methods need to be applied to discretize the governing PDEs. In addition, some simplified well models have to be employed to simulate well performance. Therefore, these models usually provide only approximate solutions to recovery responses. In other words, any relationship between production history patterns and unknown model parameters is captured only to the extent of the physical phenomena represented in the numerical simulation. It is not possible to calibrate this relationship directly from the data.

d) Difficulty in controlling convergence and avoiding the local minima: Many of the conventional automatic history-matching workflows can be regarded as an optimization process. Therefore, one should consider the issue of convergence and how to find the global minima when applying these algorithms.

e) Failure in the integration of actual field data: Since SAGD technique has been adopted in the heavy oil industry for many years, there are extensive field data become available. Because such data usually are noisy and uncertain, they are not easily included in conventional characterization methods.

1.4 SAGD Field Data

Field data in oil and gas industries provide important insights about subsurface reservoirs and underlying flowing systems and can be utilized to analyze reservoir properties and to predict production performance. As SAGD has been adopted in the heavy oil industry for more than two decades, a set of field data that is descriptive of production performance and reservoir characteristics can be gathered from the public domain. Similarly, field data in SAGD operation can be grouped into static and dynamic data. Static data include core analysis, well log measurements, well locations, well trajectories, etc. Dynamic data may contain well production/injection data, temperature profiles, pressure data, SOR, etc. It is of great interest to

propose strategies for analysis and utilization of such data in an effort to enhance the existing capabilities for characterizing heterogeneities in SAGD reservoirs.

In this thesis, field data from over 15 SAGD projects are collected, which contain information from above 2,000 injection/production and delineation wells. Analysis of raw SAGD field data of this magnitude is a challenging task. Examples of common challenges for field data analysis include data redundancy, noise, data incompletion, and large size. Another difficulty is construction of appropriate models to correlate the reservoir properties and SAGD production performances. It is because the reservoir properties can only be obtained from the nearby observation vertical wells; while operational data are only available at the horizontal production well pair, as shown in **Fig. 1-4**. Therefore, special techniques need to be proposed to deal with these issues, such as the definition of a search domain around the well pair. Because of the complicated nature of time-series data, e.g., high dimensionality and continuous update (Fu, 2011), in this thesis, many time-series data analysis techniques such as piecewise linear approximation, wavelet transform, and cubic spline interpolation are applied to analyze production time-series for fast and automatic features extraction.

1.5 Data-Driven Modeling Techniques

A definition of data-driven modeling is given by Solomatine et al. (2008) as "*data-driven modeling is based on analyzing the data about a system, in particular finding connections between the system state variables (input and output variables) without explicit knowledge of the physical behavior of the system*". Data-driven modeling techniques, combining data-mining, computational intelligence, machine learning, statistical data analysis, soft computing, and pattern recognition (Solomatine and Ostfeld, 2008), are capable of constructing proxy models that describe the behaviors of corresponding physical processes via the analysis of relevant data characterizing the systems of interest (Kjræulff and Madsen, 2008). For a given dataset containing many observable cases generated by an underlying process, the main advantage of data-driven modeling techniques is inferring the dependencies between system inputs and outputs using certain learning algorithms without building the complex physical models (Kjræulff and Madsen, 2008). For instances, instead of construction of construction of complex reservoir models, Cao et al. (2016) built proxy data-driven modes for production forecast.

Although data-driven modeling techniques have been widely employed in the oil and gas industry, its application in assessing heavy oil production in heterogeneous reservoirs and characterizing reservoir heterogeneities are lacking; in particular, the applications of data-driven models for heterogeneities characterization involving actual field data from the McMurray formation are rare.

One of the widely adopted techniques to construct the data-driven models is artificial neural network (ANN), which mimics the human neuronal structure and thinking. Since ANN was first proposed by McCulloch and Pitts (1943), it has achieved high popularity in the areas of prediction and pattern recognition (Haykin, 2008). To employ ANN for building the data-driven models, the most important step is assembling a large dataset that contains many data records or samples. Each data record should have an input vector and an output vector. For a given dataset containing a collection of data records, ANN can identify and approximate the non-linear, complex, and uncertain relationships between system inputs and outputs.

The prediction quality of ANN is often compromised due to the high-dimensional input vector, probable inter-correlation between predicting variables, and limited records in training dataset. Therefore, additional data-mining techniques need to be incorporated to process the data to increase the model robustness and improve the estimation accuracy. Examples of commonly-adopted data-mining algorithms include dimensionality reduction techniques (e.g., principal component analysis or called PCA) and clustering analysis. PCA facilitates the reduction of dimensionality of an original dataset while retaining much of its information (variation) by orthogonally transforming the original dataset into a new set of uncorrelated variables or principal components. Clustering analysis is a type of unsupervised learning methods (there is no pre-defined class label for each data sample) to partition a large dataset into many smaller subsets that share similar characteristics. The similarity among the data samples is assessed based on the measurements of distance, such as the Euclidean distance. Clustering analysis provides a useful tool to understand the hidden patterns and characteristics of data objects.

1.6 Problem Statement

Reservoir heterogeneities, including shale barrier and lean zone, have significant impacts on SAGD operations. It is necessary to determine the presence of heterogeneities and estimate their

geological properties in SAGD for effective reservoir management and operation strategy optimization. Due to various limitations and drawbacks, the inference of shale and lean zone heterogeneities in SAGD reservoir using conventional reservoir simulation and characterization methods can be challenging. The increasing volumes of SAGD field data available from the public domain may offer important information describing the relationship between reservoir characteristics and production performances.

The problem statement of this thesis is the following: "*Can the inference of reservoir heterogeneities from production data be facilitated by the integration of machine learning techniques in a series of data-driven modeling workflows?*"

1.7 Research Objectives

The main research objective is *to develop a novel workflow for reservoir heterogeneities characterization in SAGD through the integration of production data analysis and data-driven modeling techniques and to demonstrate the feasibility of data-driven modeling approaches as practical tools for heterogeneities characterization in SAGD reservoirs*. The specific objectives are listed here:

a) Explore the possibility of relating production performance to reservoir heterogeneities measures using data-driven models. Both the forward and inverse relationship between production parameters and reservoir characteristics will be investigated.
b) First, develop a forward workflow for SAGD production prediction using data-driven models, in which, SAGD field data are integrated. Given the challenges associated with field data, a fast field data analysis procedure must be proposed to analyze large amounts of raw field data and to assemble training datasets with sufficient data samples. Considering reservoir properties/heterogeneities and production performance are highly correlated, it supposed that SAGD production performance can be predicted from the trained data-driven models effectively and quickly. Multiple types of uncertainties stemmed from model parameters and data will be studied during the process of building data-driven models.
c) Next, develop workflows for inference of reservoir heterogeneities by integrating production data analysis and data-driven models. The primary goal is to characterize

reservoir heterogeneities directly from field data. Due to the difficulties and limitations of analysis of field data, such as missing important parameters (e.g., production constraints), only the forward relationship between SAGD production and reservoir parameters (conventional numerical simulation) can be inferred via data-driven models; while the inverse relationship between reservoir characteristics and production data (traditional history-matching) cannot be constructed. To deal with this challenge, an idea comes up: "*how about constructing synthetic models instead, and use the clean data obtained from synthetic models to develop a characterization workflow?*" The synthetic models contain all the important information and their parameters are easy to control. Therefore, many synthetic SAGD models are built based on the extracted properties from field data. Parameterizations of production profiles and heterogeneities based on their geological features are then carried out from the synthetic models to assemble the training dataset. Consequently, data-driven models employed to construct the inverse workflow to efficiently estimate heterogeneities characteristics. To accomplish this sub-objective, techniques of formulating inputs and outputs need to be investigated.

d) Demonstrate the feasibility of applying the proposed characterization workflows in estimating shale barriers and lean zones for numerous cases, in which, the true reservoir heterogeneity distributions are unknown.

1.8 Thesis Structure

It should be noted that this thesis is written in a paper-based manner by combining 4 articles (chapters 3 to 6); detailed and specific introduction, literature review, methodology, and conclusion can be found in each chapter. In addition, the corresponding references are listed at the end of each chapter. The aggregated bibliography for the whole thesis is also provided at the end of the thesis.

There are totally 8 chapters in this thesis including the introduction (chapter 1), the main topics (chapters 2 to 7) and the concluding remarks (chapter 8). The thesis is organized as follows:

Chapter 1 presents a general introduction of this thesis study, including the background of heavy oil resources, SAGD technology, conventional and proposed characterization methods, motivations of this research, problem statement, research objectives, etc.

Chapter 2 summarizes the main materials, methodologies, and algorithms employed in the 4 papers presented in chapter 3 to 6 in this thesis. For more detailed explanations, please refer to the consequent chapters.

Chapter 3 describes a novel and practical methodology for SAGD production performance prediction by the implementation of knowledge-based techniques. In this chapter, field data from a small number of SAGD projects are assembled to extract important parameters describing reservoir properties and operational conditions. Data-driven models are applied to forecast SAGD production.

Chapter 4 employs and extends the workflow presented in chapter 3 to a larger field dataset to correlate SAGD production performance and reservoir heterogeneities. Additional input and output parameters are integrated to 1) consider uncertainty results from the extraction of input feature from nearby wells and 2) describe SAGD production performance (steam injection efficiency). Other data-mining algorithms are also included to improve prediction accuracy and robustness of the proposed workflow.

Chapter 5 investigates the correlation between some important production patterns extracted from production time-series data and reservoir heterogeneities due to the presence of randomized shale barriers in SAGD reservoirs. Considering the difficulties associated with inferring shale heterogeneities from analysis of actual field data, a series of synthetic models are generated to test the methodology. A practical shale heterogeneities characterization workflow is proposed to estimate shale distribution in the formation by combining production data analysis and artificial intelligence techniques. Many different shale distribution scenarios are studied to model realistic shale heterogeneities in SAGD reservoirs.

Chapter 6 explores the effect of the presence of two types of heterogeneities, i.e., lean zones and shale barriers on SAGD production. Similar to chapter 5, data-driven models are constructed to characterize the complex reservoir heterogeneities. A total number of 2800 heterogeneous cases for 15 different heterogeneities scenarios are generated. To improve the characterize performance, a novel two-level data-driven model is proposed to estimate the

number of each heterogeneity and then to characterize their corresponding geological characteristics.

Chapter 7 provides a critical analysis of the proposed data-driven models including 1) comparing the results obtained from the data-driven models with other conventional linear and nonlinear algorithms; 2) comparing the proposed methods with conventional numerical simulation and history-matching workflows; 3) explaining the benefits of application of data-driven models for reservoir management routines; 4) illustrating the limitations of the proposed data-driven models.

Chapter 8 summarizes the main findings, conclusions, and original contributions of this thesis. The recommendations for the future work are also provided at the end of this chapter.

1.9 Reference

Alberta Energy Regulator (2016), *Alberta energy regulator 2015/16 annual report*. Report prepared by Alberta Energy Regulator, Alberta, Canada.

Butler, R., McNab, G., & Lo, H. (1981). Theoretical studies on the gravity drainage of heavy oil during in-situ steam heating. *The Canadian Journal of Chemical Engineering, 59*(4), 455-460.

Cao, Q., Banerjee, R., Gupta, S., Li, J., Zhou, W., & Jeyachandra, B. (2016). Data driven production forecasting using machine learning. Paper presented at the *SPE Argentina Exploration and Production of Unconventional Resources Symposium*, Buenos Aires, Argentina.

Chopra, A., Severson, C., & Carhart, S. (1990). Evaluation of geostatistical techniques for reservoir characterization. Paper presented at the *SPE Annual Technical Conference and Exhibition*, New Orleans, LA, USA.

Francis, L. (2001). The basics of neural networks demystified. *Contingencies (11/12 2001)*, 56-61.

Fu, T. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence, 24*(1), 164-181.

Gu, Y., & Oliver, D. S. (2005). History matching of the PUNQ-S3 reservoir model using the ensemble kalman filter. *SPE Journal, 10*(02), 217-224.

Haykin, S.S. (2008). *Neural networks and learning machines* (3rd ed.), Upper Saddle River, NJ, USA: Pearson.

Hein, F. J. (2017). Geology of bitumen and heavy oil: An overview. *Journal of Petroleum Science and Engineering, 154*, 551-563.

Hiebert, A. D., Morrish, I. C., Card, C., Ha, H., Porter, S., Kumar, A., . . . , & Close, J. C. (2013). Incorporating 4D seismic steam chamber location information into assisted history matching for A SAGD simulation. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Canada.

Hu, L. Y. (2000). Gradual deformation and iterative calibration of Gaussian-related stochastic models. *Mathematical Geology, 32*(1), 87-108.

Jia, X., Cunha, L., & Deutsch, C. (2009). Investigation of a stochastic optimization method for automatic history matching of SAGD processes. *Journal of Canadian Petroleum Technology, 48*(01), 14-18.

Kjræulff, U. B., & Madsen, A. L. (2008). *Bayesian networks and influence diagrams: A guide to construction and analysis*, Springer Science+Business Media.

Mattax, C. C., & Dalton, R. L. (1990). Reservoir simulation (includes associated papers 21606 and 21620*). Journal of Petroleum Technology, 42*(06), 692-695.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

Mirzabozorg, A., Nghiem, L., Chen, Z., & Yang, C. (2013). Differential evolution for assisted history matching process: SAGD case study. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Oliver, D. S., & Chen, Y. (2011). Recent progress on reservoir history matching: A review. *Computational Geosciences, 15*(1), 185-221.

Panwar, A., Trivedi, J. J., & Nejadi, S. (2012). Importance of distributed temperature sensor (DTS) placement for SAGD reservoir characterization and history matching within

ensemble kalman filter (EnKF) framework. Paper presented at the *SPE Latin America and Caribbean Petroleum Engineering Conference*, Mexico City, Mexico.

Romero, C., & Carter, J. (2001). Using genetic algorithms for reservoir characterization. *Journal of Petroleum Science and Engineering, 31*(2), 113-123.

Shahkarami, A., Mohaghegh, S. D., & Hajizadeh, Y. (2015). Assisted history matching using pattern recognition technology. Paper presented at the *SPE Digital Energy Conference and Exhibition*, The Woodlands, Texas, USA.

Solomatine, D., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics, 10*(1), 3-22.

Solomatine, D., See, L. M., & Abrahart, R. (2009). Data-driven modelling: concepts, approaches and experiences. *Practical hydroinformatics* (pp. 17-30), Springer.

Williams, M., Keating, J., & Barghouty, M. (1998). The stratigraphic method: A structured approach to history matching complex simulation models. *SPE Reservoir Evaluation & Engineering, 1*(02), 169-176.

Zhang, F., Reynolds, A. C., & Oliver, D. S. (2003). An initial guess for the Levenberg–Marquardt algorithm for conditioning a stochastic channel to pressure data. *Mathematical Geology, 35*(1), 67-88.

Zhang, X. K., Feizabadi, S. A., & Yang, P. (2014). An integrated approach to building history-matched geomodels to understand complex long lake oil sands reservoirs, part 1: Geomodeling. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Figures

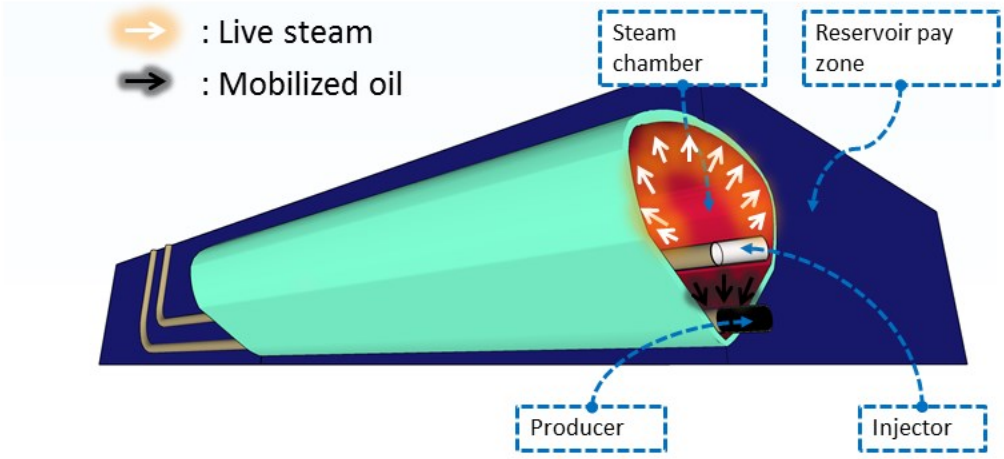


Fig. 1-1 Schematic of a typical single well pair SAGD in 3D view.

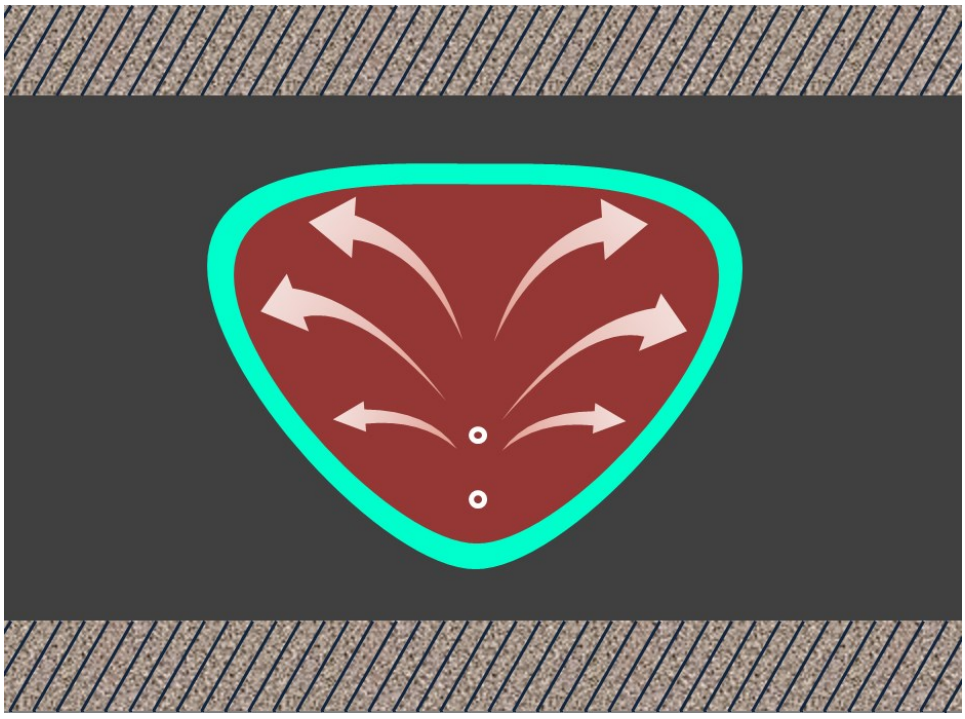


Fig. 1-2 Schematic of a typical SAGD steam chamber in 2D view.

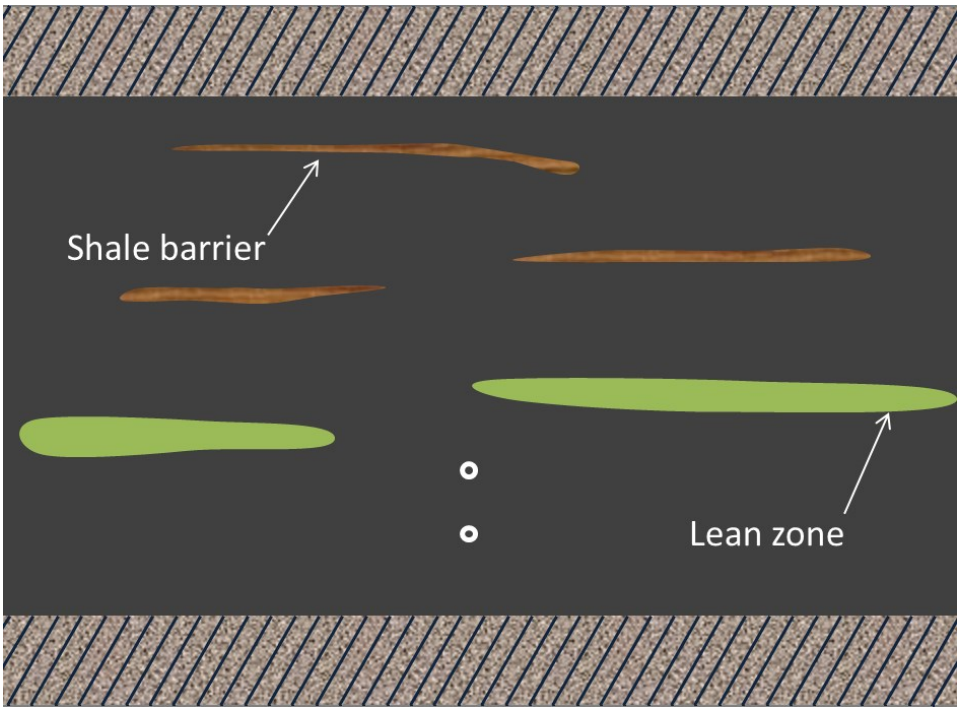


Fig. 1-3 Schematic of heterogeneity distributions in SAGD reservoir.

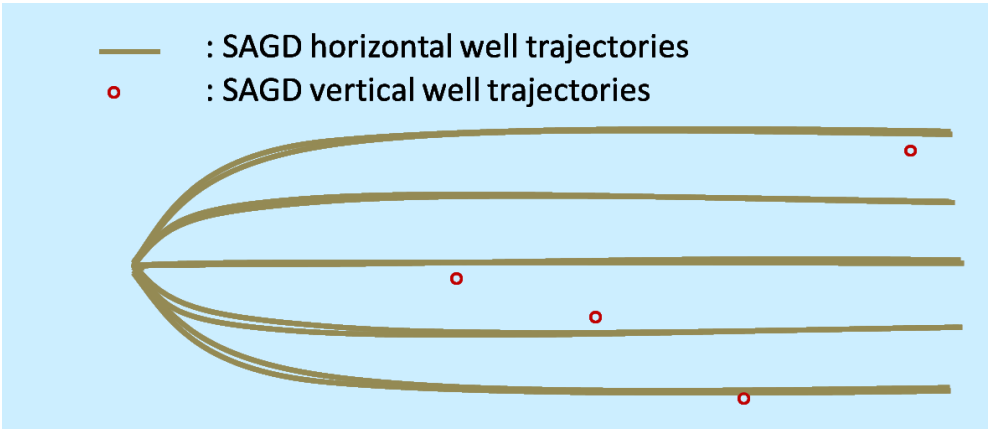


Fig. 1-4 Schematic of a SAGD well pad consisting of 5 horizontal production well pairs and 4 vertical observation wells in top view.

Chapter 2 Materials, Model Setups, and Methodologies

Chapter Overview

This thesis is written in a paper-based structure by summarizing a continuous project. Therefore it is inevitable to have some similarities among the 4 included publications, such as techniques used to analyze field data and methodologies used to train the models, etc. A consolidated summary of the materials, model setups, and methodologies used in this research are explained and illustrated in this chapter. Further details and their specific applications in this work are presented in the subsequent chapters.

2.1 SAGD Field Data Pre-process

In this thesis, a set of SAGD field data assembled from the public domain is studied. The available data can be sub-divided into the following categories: Logging, Production, Injection, Well Header, Well Pair, and Deviation Survey:

- Logging: well log data including original petrophysical logging measurements such as spontaneous potential (SP), gamma ray (GR), true resistivity (RT), and a number of interpreted logs such as reservoir porosity and water saturation;
- Production: production rate, production time, and cumulative production of oil, gas, and water;
- Injection: steam and water injection volumes;
- Well Header: surface and bottom locations in different geographic coordinate systems, well depths, Kelly bushing elevation, and other relevant drilling and well completion information;
- Well Pair: producing field, pad, and associated horizontal well pairs;

- Deviation Survey: well trajectories of horizontal wells.

Well depth is measured from the elevation of Kelly bushing, which varies from well to well. Therefore, it must be normalized against a reference surface. This pre-processing step is illustrated in **Fig. 2-1**(**a**). It should be mentioned that vertical wells are drilled as delineation wells, while horizontal wells are drilled as producers and injectors. Logging information is available only at the vertical wells, while rests of the data are associated with the horizontal wells.

In chapter 3 and 4, a data record is assembled by combining logging information from the nearby vertical wells and operating/production information from the associated horizontal wells. The input attributes related to reservoir properties are extracted directly from logging interpretation, while for the inputs pertinent to operation conditions are extracted from analyzing the injection and production datasets. This is facilitated by assigning a rectangular search domain around the well pair, as shown in **Fig. 2-1**(**b**). The blue rectangle outlines the boundaries of the search domain for this well pair example. The red and black solid lines denote the well trajectories of the producer and the injector in this well pair, respectively. The logging wells that are located in the rectangular domain are selected for interpretation. If no logging wells can be found within the domain, or if a significant number of logs are missing such that no reliable interpretation is possible, this well pair would be excluded from the dataset. If a particular logging well is located nearby multiple horizontal well pairs, its interpreted values would be assigned to only the closest horizontal well pair.

2.2 Data-Driven Models Construction

2.2.1 Background of Data-Driven Modeling Techniques

Data-driven modeling involves analysis of data characterizing the system of interest and focuses on application of the machine learning methods to understand and build models that describe the behavior of the corresponding physical processes using experience, knowledge, and observed data generated from the processes of interest (Kjræulff and Madsen, 2008). Learning the dependencies between inputs and corresponding outputs is the primary focus in data-driven modeling, and it is often accomplished using various supervised learning techniques (Solomatine and Ostfeld, 2008). Examples of some popular methods used in data-driven modeling are

statistical methods, ANN (Joo et al., 2014), and fuzzy logic (Petrović et al., 2014). The methods used nowadays have advanced significantly beyond the ones used in the conventional empirical regression. They are used for solving numerical prediction problems, reconstructing highly non-linear relationships, performing data classification, and building rule-based expert systems.

ANN is a widely adopted data-driven modeling technique used for identifying or approximating a complex non-linear relationship between input and target variables with only a limited number of assumptions about the "physical" behavior of the system. The neural network is a kind of machine learning algorithm; it is widely used for pattern recognition and prediction by mimicking the information transfer in the central nervous system of human (Haykin, 2008). Once a model is trained, it can be used to describe the behaviors and properties of this physical process. Compared to other function approximation techniques (e.g., response surface and Taylor expansion), ANN offers certain advantages including its capacity of inferring highly complex, nonlinear, and possibly uncertain relationships between system variables, requiring essentially zero prior knowledge regarding the unknown function (Hasani and Emami (2008). Many different learning tasks such as classification and non-linear function approximation can be well suited for ANN modeling. The first neural network model was introduced by McCulloch and Pitts (1943). After some major improvements and developments of ANN in recent decades, many formulations of neural network utilizing different transfer functions, learning algorithms, and network architectures (including hybridized fuzzy neural network) have been proposed and applied in various fields.

Applications of neural network can also be found in petroleum engineering. ANN has achieved significant popularity in areas such as production prediction (Al-Fattah and Startzman, 2003), reservoir characterization or properties prediction (An and Moon, 1993; Gharbi and Elsharkawy, 1999; Tang et al., 2011), history-matching (Ramgulam, 2006), classification (Stundner and Al-Thuwaini, 2001), proxy for prediction of recovery performance (Lechner and Zangl, 2005), production operation optimization and well design (Yeten et al., 2002). In recent years, the neural network has also been utilized to evaluate enhanced oil recovery projects (Parada and Ertekin, 2012; Zerafat et al., 2011) and assess CO_2 sequestration process (Mohammadpoor et al., 2012). In this thesis, the data-driven models used in chapters 3 to 6 are constructed using ANN. A brief explanation of ANN technique is explained here.

2.2.2 Artificial Neural Network Modeling

The basic neural network architecture is composed of an input layer, an output layer, and any number of hidden layers. The output layer is made of the target variables, while the input layer consists of attributes that are related to the target variables. A neural network with only the input and output layer is called a single-layer perceptron (SLP), which can be applied to problems that are linearly separable. In other cases, the multi-layer perceptron (MLP) neuron network can be implemented. The MLP may contain any number of hidden layers, which serve to transform the original input data space into new spaces, where it is easy to perform the classification or regression process. The MLP is the most widely adopted perceptron in solving problems with real data. **Fig. 2-2**(**a**) shows an example of the basic architecture of the MLP neuron network, which has one hidden layer; the blue circles, red squares, and green triangles denote neurons in the input layer, hidden layer, and the output layer, respectively. It is a fully connected neural network since every neuron in the network is connected to the nodes in its adherent layers.

The feedforward backpropagation neural network is implemented in this thesis, where the error is back propagated to train the network parameters (weights and biases assigned to each connection) in a supervised learning algorithm. The nodes between neighboring layers are connected by weights, as shown in **Fig. 2-2**(**b**). Two kinds of signals are transferred in the feedforward neural network. The first one is function signal (or input signal), which comes from the input neurons and propagates forward through the hidden layers to the output layer; another one is error signal, which is generated from the output neurons and propagates backward through the hidden layers to the input layer (Haykin, 2008). A schematic of signal transfer is shown in **Fig. 2-2**(**c**), where black solid arrows and red dashed arrows denote the function signals and the error signals, respectively. Values of weights and biases are updated using a training dataset such that the mismatch between network predictions and known values of the target variables is minimized (Francis, 2001).

The input signal x of a certain node in the hidden or output layer is the weighted summation of output signals from previous layer, as **Eq. 2-1** shows:

$$x_i = b_i + \sum_{j=1}^{m} w_{ij} \cdot y_j \quad \text{..... (2-1)}$$

where x_i is the weighted sum of input signals at node i in the current layer; y_j denotes the output value of node j in preceding layer; m represents the total number of nodes in the preceding layer; b_i is the threshold (bias) value; w_{ij} is the weight associated with the connection between node i and node j. To calculate the output value of this neuron, a transfer function (or activation function) is applied to the weighted sum. Various transfer functions such as pure line function, threshold function, and sigmoid function (e.g., hyperbolic tangent function and logistic function) can be used. In this thesis, the hyperbolic tangent function, as shown in its general form in **Eq. 2-2**, is used:

$$f(x) = \tanh(x) \quad \text{(2-2)}$$

where x is an independent variable. The outputs of the hyperbolic tangent function are in the range of (-1, 1). It scales large positive and negative input values to 1 and -1, respectively. The output of node i or y_i can be computed by **Eq. 2-3**:

$$y_i = f(b_i + \sum_{j=1}^{m} w_{ij} \cdot y_j) \quad \text{(2-3)}$$

The value calculated from **Eq. 2-3** is output signal from the node i, which can be considered as the input signal to the next layer. **Eqs. 2-1** to **2-3** are repeated until the final output layer is reached and predicted value for the output variable is calculated. These equations allow the function signals to be propagated from the input layer to the output layer.

The backpropagation algorithm is applied to update the weight and biases during the learning stage. The classical backpropagation algorithm is a gradient descent supervised learning algorithm. An error signal is computed as the mismatch between target and prediction at the output layer. The weights of the output layer are updated based on estimated derivatives of error with respect to weight. This step is repeated to transmit the error signal in the reverse direction until all weights are updated. This entire updating procedure must be repeated for many epochs

until a certain stop criterion is reached. Since backpropagation is a gradient descent method, the transfer functions must be differentiable.

Even though the gradient descent backpropagation is feasible for most problems, if the step-size is properly selected, it bears some limitations such as slow convergence. In order to enhance computational efficiency, various modified schemes have been developed including the Levenberg–Marquardt backpropagation (LM-BP) algorithm. The Levenberg–Marquardt backpropagation algorithm was proved as an efficient method to update weights and biases of neural network (Ampazis and Perantonis, 2000; Hagan and Menhaj, 1994). The details of the LM-BP algorithm can be found in a number of references (Hagan and Menhaj, 1994; Haykin, 2008), and it is applied in this thesis.

Due to the large disparity in scales of different data sources, normalization or standardization procedure is often performed (Francis, 2001). Normalization is an important pre-processing step for ANN modeling, with all data values being transformed to vary between a certain range such as [0, 1] or [-1, 1]. This step can help to reduce bias in the minimized solution as a result of the overwhelmingly large data values (Al-Fattah and Startzman, 2003). A data point x can be normalized by **Eq. 2-4**:

$$x^N = 2 \times \frac{x - x_{\min}}{x_{\max} - x_{\min}} - 1 \quad \text{..... (2-4)}$$

where x^N is the normalized data value ranging between -1 and 1; x_{max} and x_{min} represent the maximum and minimum value of this data vector, respectively.

In order apply ANN in building data-driven models, a dataset consisting of a set of records composed of both input (predicting) attributes and the corresponding output (target) attributes need to assembled first. The entire dataset is usually divided into three portions to be used during the three stages of ANN model construction. These stages include (1) training (applying certain learning algorithms to calibrate the network parameters including weights and biases; (2) testing (evaluating the performance of trained network and optimizing the network architecture); and (3) validating (verifying the network performance using data that has not been previously presented to the network during the first two stages). In this study, the neural network modeling is implemented using the Neural Network Toolbox (Demuth et al., 2008) in Matlab®.

2.2.3 Performance Evaluation for Data-Driven Models

In thesis three approaches are applied to evaluate the performance of data-driven models, including cross-plot, coefficient of determination (R^2) and mean squared error (*MSE*).

- Cross-plot: it is the most intuitive visualization method to examine the performance of data-driven models. A schematic of a cross-plot is shown in **Fig. 2-3**. Usually, one axis (x-axis in this figure) represents the target values, while another axis (y-axis in this figure) denotes the predicted values from models. For a given sample whose target is known, the location of the scatter point in the 2D figure depicts its corresponding prediction from the model. The 45-degree splitting line indicates a perfect model prediction, i.e., the predicted values are equal to their targets. However, it is not always easy to obtain such perfect predictions in reality. For many cases, if scatter points follow the splitting line, then we can conclude that the model provides a reasonable estimation.
- Coefficient of determination (R^2): it is an important statistical measure of how well a model fits their targets. According to Barrett (1974), R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(T_i - \hat{f}_i)^2}{\sum_{i=1}^{n}(T_i - \overline{T})^2} \quad \text{..... (2-5)}$$

  where T_i is the observable variable or target; $\hat{f}$ represents the predicted value from the model; $\overline{T}$ denotes the average of all targets; n is total number of samples; For a set of data samples, a larger value of R^2 indicates an increasing prediction precision from a regression model (Barrett, 1974). In this thesis, a perfect prediction means the value of R^2 is 1.
- Mean squared error (*MSE*): it is another variable to quantitatively evaluate the model prediction. A small value of *MSE* indicates that the estimation of the model is good. In this thesis, *MSE* can be defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - T_i)^2 \quad \text{..... (2-6)}$$

2.2.4 Determination of ANN Structure

The performance of ANN modeling would be affected by the number of hidden layer nodes (Tan and Smeins, 1996). Too many neurons (or connections) may lead to an overfitting problem, while the prediction performance of ANN is compromised with insufficient nodes. There are no concrete guidelines to determine the number of free parameters in the hidden layer. In order to enhance computational efficiency and network predictability, the network architecture including the number of hidden layers, the number of neurons in hidden layers should be optimized by balancing between prediction accuracy and overfitting. This number of hidden nodes is typically considered to vary as a function of input vector dimension and the amount of training data.

Several relationships or rules of thumb exist in the literature relating the training-dataset size to some user-defined error parameters calculated for a given network configuration (Waszczyszyn, 1999; Xu and Chen, 2008). A recent review of a range of design issues related to ANN development in petroleum industry can be found in Al-Bulushi et al. (2012). It was demonstrated that a single hidden layer could approximate any function with a finite number of discontinuities (Kröse et al., 1993). Heaton (2008) showed the number of neurons should be less than two times of the number of input parameters. Although there are some rules of thumb to select the number of hidden nodes have been proposed in some previous studies, it is common to select the optimum number of hidden neurons by trial and error.

To design the optimum architecture of the neural network, the n-fold cross-validation method is implemented in this thesis. The training dataset is divided into the n equal size subsets randomly. From the n subsets, one subset is selected as the validation part while the remaining $n-1$ subsets are assigned as the training part. First, for a particular network structure, the training part is used to train the corresponding network parameters (weights and biases); the network performance is subsequently evaluated using the validation part. The mean squared error (*MSE*) between target and prediction is computed as a measure of network performance. This step is repeated numerous times with different random initial solution of the weights and biases, and the solution with the lowest *MSE* is selected. Next, another subset is selected as the validation part, and the remaining subsets are assigned as the training part. The training-validation process is repeated for n times to calculate an average performance (i.e., average *MSE*) of this network structure. This entire procedure is carried out again for another parameter

exploration (network structure). Finally, the optimum neural network architecture with the best average performance is determined. A simple flowchart of application of n-fold cross-validation technique in determination of optimum structure of ANN is shown in **Fig. 2-4**.

2.3 Improvement of Data-Driven Models' Performance

2.3.1 Principal Component Analysis

The prediction quality of ANN is often compromised due to the high-dimensional input vector, probable inter-correlation between predicting variables, and limited records in training data. Principal Component Analysis (PCA) can help to increase model robustness by reducing the dimensionality of the original dataset while retaining much of its information (variation) (Jolliffe, 2005). This is achieved by orthogonally transforming the dataset into a new set of uncorrelated variables or principal components, which are computed from an eigenvalue decomposition of the covariance matrix (Smith, 2002). PCA has been successfully applied in areas including history-matching (Sarma et al., 2007; Yadav, 2006), reservoir property estimation (Dadashpour et al., 2011; Lee et al., 2002; Scheevel and Payrazyan, 2001), and production data analysis (Bhattacharya and Nikolaou, 2013).

PCA is performed to reduce the dimensionality of the original dataset. First, the mean of each dimension is subtracted from the original data:

$$Z_{ij} = X_{ij} - \bar{X}_j \quad \text{..... (2-7)}$$

where X_{ij} is the j_{th} variable of i_{th} sample, $\bar{X}_j$ is the average of X_j over all N samples, while Z_{ij} is the new variable that represents deviation from the mean. The purpose of this step is to simplify the calculation for the covariance matrix and remove bias due to large disparity in mean values. Next, the covariance between two variables X_j and X_k is defined as:

$$COV(X_j, X_k) = \frac{\sum_{i=1}^{N} Z_{ij} Z_{ik}}{N-1} \quad \text{..... (2-8)}$$

where N is the number of data samples; i denotes sample index, j and k are dimension indices of two variables. Finally, eigenvalue decomposition of the covariance matrix (**Eq. 2-8**) is carried out. Individual eigenvalue represents the significance or contribution of the variance from the corresponding eigenvector to the total variance of the original data. The eigenvectors with highest eigenvalues are principal components (*PC*), which can be obtained by sorting the eigenvalues in a decreasing order. Once the principal components have been identified, the original dataset is transformed into the principal component space as principal scores (*PS*) according to the following equation, which are regarded as the inputs attributes in subsequent ANN modeling.

$$PS = PC \times Z^T \quad \text{..... (2-9)}$$

2.3.2 Clustering Analysis

When faced with a large amount of data, it is shown that robustness and accuracy of the prediction capability are greatly enhanced by performing clustering analysis to identify internal data structures and groupings prior to ANN modeling (Amirian et al., 2015). Clustering analysis facilitates the identification of internal structures among the data by partitioning a large dataset into numerous sub-datasets with similar characteristics. This is a necessary step in chapter 4 since the original dataset encompasses data collected from ten different SAGD fields in Canada, with wide-ranging reservoir properties and production characteristics. K-means (MacQueen, 1967), which is a widely-adopted partitioning clustering algorithm, is applied due to its simplicity and computational efficiency (Bahrololoum et al., 2015).

For a specified number of subgroups, the k-means algorithm assigns data points to individual groups by minimizing the within-cluster sum of squares according to **Eq. 2-10**:

$$S_{\mathrm{c}} = \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} I_{ij}(\mathbf{X}_i - \overline{\mathbf{X}}_j)(\mathbf{X}_i - \overline{\mathbf{X}}_j)^T \quad \text{..... (2-10)}$$

where S_c is the within-cluster sum of squares; g stands for the number of groups, I_{ij} equals to 1 if a data sample $\mathbf{X_i}$ belongs to cluster j and 0 otherwise; $\overline{\mathbf{X_j}}$ represents the center of cluster j, which is the arithmetic average of all data objects in this cluster. The within-cluster sum of squares is the objective function to be minimized through iteration.

Details about the theory and formulation of the k-means clustering algorithm were explained by Hammouda and Karray (2000). One shortcoming with this algorithm is that the final clustering results are dependent on the initialization of cluster centers; therefore, it is necessary to repeat the minimization procedure with numerous initial guesses in order to identify the optimal groupings.

2.4 Uncertainty Analysis

The analysis of uncertainty has been received much attention in recent years (Mezić and Runolfsson, 2008). Walker et al. (2003) explained various sources of uncertainty including input data uncertainty, model parameter uncertainty, and model outcome uncertainty, which is the accumulated uncertainty in the predicted values. In this application, data uncertainty primarily derives from inaccurate and incorrect data, limited number of records in the dataset and imprecise (indefinite) analysis criteria. Model parameter uncertainty is common with most data-driven modeling techniques like ANN, whose training can be posed as an under-determined inverse problem with non-unique solutions. In addition, model parameter uncertainty could also stem from random initializations

Three main groups of techniques are commonly adopted for uncertainty assessment: the Gaussian approach, the Monte Carlo method, and bootstrap method. Gaussian methods assume that distributions of uncertainty, including those exhibited by the input data and model parameters, are Gaussian. Specific applications of the Gaussian approaches for uncertainty quantification can be divided into several categories: analytical error propagation equation method (Refsgaard et al., 2007; Verga et al., 2002); Bayesian approach (Nigrin, 1993; Wright, 1999; Zhang et al., 2011); and uncertain neural network method (Ge et al., 2010). The Monte Carlo method refers to a general stochastic approach for approximating the probability of a certain outcome by random sampling of a large number of realizations. The application of the Monte Carlo simulation in uncertainty analysis and error propagation can be found in a number

of works (Guan et al., 1997; Hanna et al., 1998; Norman, 2013; Papadopoulos and Yeung, 2001). The bootstrap approach, introduced by Efron (1979), is a resampling method to estimate the statistic properties of a given sample dataset. The combination of bootstrap approach with the ANN is the bootstrapped neural network, which was employed to estimate safety margins with appropriate confidence intervals for the nuclear power plant (Secchi et al., 2008). Examples for applications of bootstrap method for uncertain analysis include flood forecasting (Han et al., 2007; Tiwari and Chatterjee, 2010) and electricity price prediction (Khosravi et al., 2013).

In this thesis, data uncertainty because of the small size of the dataset and imprecise analysis criteria, together with model parameter uncertainty due to training algorithm and initialization, are investigated. The aggregated consequence of these uncertainties is exhibited in the output (prediction) uncertainty. A comprehensive analysis involving all the aforementioned uncertainties with an actual SAGD dataset is novel. First, model parameter uncertainty is quantified with a Monte Carlo framework, in which training of the optimum network is repeated with many randomized initializations of model parameters. Next, parametric bootstrapping is performed to assess the data uncertainty introduced during the data analysis process (e.g., imprecise analysis criteria). Finally, bootstrapping with replacement is applied to evaluate the uncertainty stemming from limited dataset size.

2.4.1 Model Parameter Uncertainty

A Monte Carlo framework, in which training of the optimum network is repeated with many randomized initializations of model parameters, is used to quantify model parameter uncertainty. Aggregating the trained weights and biases derived from all initializations, the conditional probability P**(w|d)**, where **w** and **d** refer to the model parameter and data vectors, respectively, can be established. For a given testing sample, the corresponding output uncertainty is estimated by sampling multiple sets of **w** vectors from P**(w|d)**.

2.4.2 Data Uncertainty

The first source of uncertainty in the data is the results of outliers and imprecise cut-off values used in the logging interpretation, a common consideration in geologic data analysis. Uncertainty in input data is accounted for by estimating a likelihood function for each input attribute and performing parametric bootstrapping of this likelihood to assess uncertainty related to this input

attribute. In this thesis, a detailed sensitivity analysis reveals that each input attribute follows approximately a uniform distribution with a +/- 10% variation in attribute value after applying different cut-off criteria. Next, N_r data records are sampled from these uniform likelihood functions; each sample can be regarded as a realization from the probability P**(d)** and is subjected to ANN training. This probability can be combined with P**(w|d)** to obtain P**(w,d)** = P**(w|d)** × P**(d)**. If we ignore the model parameter uncertainty here and consider only data uncertainty, P**(w,d)** = P**(d)**, a total of N_r trained networks are obtained; therefore, for a given testing sample, the corresponding output uncertainty can be computed from predictions generated from all N_r trained models.

The second source of data uncertainty is a result of limited dataset size. The collected dataset is only one of an infinite number of possible datasets that may be drawn in a certain input domain (Srivastav et al., 2007). The variability of sampling the input and target values could lead to the uncertainty in training dataset. Srivastav et al. (2007) demonstrate how the bootstrap approach can be applied to quantify the uncertainty due to data size. If ANN approach is applied as the regression function (f) for prediction, for a particular input vector x_n, the output vector can be presented as $\boldsymbol{y_n} = f(\boldsymbol{x_n};\boldsymbol{w})$, where n = 1, …, N (total number of records). Considering that P**(w|d)** would vary when modeling with a different realization of all possible sample datasets, the uncertainty or variance of the distribution of y_n can be estimated by the bootstrap technique, where B_n additional datasets are sampled randomly based on the original dataset with replacement. The network is trained using B_n datasets to obtain B_n trained network and the output variance for a given new input from testing dataset can be directly calculated.

2.5 Synthetic SAGD Model Construction

In this thesis, a series of 2D synthetic SAGD models based on typical Athabasca oil reservoir properties and operating conditions are constructed using the BUILDER (CMG, 2015). The homogeneous (base) model consists of only oil sand and is shown in **Fig. 2-5**. The top of the reservoir is located at 200 m beneath the surface. Considering the symmetric growth of a steam chamber, only one-half of the distance between two neighboring well pairs is modeled. A set of horizontal wells of 900 m are placed along the lateral expansion direction (Y). The injector wellbore is located at 225 m in the Z axis, which is 5 m above the producer. Local grid

refinement is implemented near the injector/producer for better accuracy. Detailed synthetic model properties will be illustrated in chapter 5 and chapter 6.

Properties of individual shale barrier and lean zone including its proportions, shape, location, and size (i.e., lateral extent and thickness) are assigned according to probabilistic distributions inferred from typical SAGD reservoirs in the Athabasca deposits (Ma et al. 2017). Relevant reservoir properties and operating conditions will be presented in the corresponding chapter. Each model is subjected to numerical flow simulation in STARS (CMG, 2015) and the corresponding production profiles are recorded.

2.6 Parameterization from Production Time-Series Data

Time series data is usually a sequence of data points which are measured and recorded in a time interval. It is a common type of data in petroleum industry, for example, oil production rate, water injection rate, cumulative oil productions, SOR, reservoir pressure, and water cut profile. Identification of patterns and extraction of features from time-series data play crucial roles in any data-driven modeling workflows because the large dimensionality of time-series data would pose big changelings for further analysis. Normally, utilizing a smaller number of extracted patterns and features instead of the entire series for ensuing analysis would increase computation speed and improve quality of analysis dramatically.

Many types of parameterization of production time-series data methods are proposed and employed in this thesis to extract important input parameters for construction of data-driven models. These methods can be grouped into two categories including manual extraction and automatic feature extraction techniques, which consist of piecewise linear approximation, cubic spline interpolation, and discrete wavelet transform (DWT).

2.6.1 Manual Decline Pattern Extraction

Manual extraction is the simplest and intuitive method to extract relevant features from production time-series data. The basic mechanism of this approach is defining and retrieving some observable features from time-series data manually. The observable feature can be defined as a short period of time-series data with certain patterns, such as decline, increase, and “S”

shape. Next, by selecting some important points in this observable pattern (e.g., the staring, ending or local minimum/maximum), we are able to parametrize the time-series data properly.

For instance, the oil production rate profile corresponding to the heterogeneous model with a single shale barrier is presented in **Fig. 2-6**. It is easy to observe a decline pattern in the production profile due to the impact of shale barrier. Detailed sensitivity analysis demonstrates the shape, size, and location are related to the geological feature of the shale barrier. In order to extract the decline pattern, three feature points are determined from the production profile manually including the left starting point (l), local minimum point (b), and right ending point (r). Each feature point can be defined by three values (corresponding production time (t), rate (q) and cumulative oil production (Q)), as shown in **Fig. 2-6**. Therefore, a total number of 9 parameters can be extracted manually as input features to describe a certain decline pattern.

The obvious limitations of the manual extraction method are that it is tedious and subjective. Therefore, some automatic feature extraction techniques need to be included to deal with these issues.

2.6.2 Piecewise Linear Approximation

Piecewise linear approximation is one of the most commonly used representations of time-series data (Keogh et al., 2004). It can be employed to pre-process of raw time-series data for ensuing modeling frameworks. Basically, piecewise linear approximation segments a large raw time-series data into a small number of straight lines. Once the small segments are obtained, they can be used to represent the original time-series data. The advantage of using a piecewise linear approximation for time-series data analysis is that the dimensionality of the original time-series data is reduced significantly.

In this thesis, piecewise linear approximation algorithm would facilitate dividing an original production profile data into multiple linear segments, from which the features points of the decline patterns are easily identified. It is a robust feature extraction approach as no human factors are involved during the whole process. It also should be noted that certain criterion or threshold value need to specify before applying piecewise linear approximation algorithm. The procedure of how to employ piecewise linear approximation algorithm can be found in Keogh et al. (2004).

2.6.3 Cubic Spline Interpolation

Piecewise linear approximation method is useful for capturing the feature points associated with a given decline pattern; however, in order to represent the underlying curvature, some higher-order approximations are needed. That is because many possible curves can be fitted using just three feature points that describe a decline pattern. To that end, cubic spline interpolation (McKinley and Levine, 1998), which is piecewise third-order polynomials, is applied next to approximate the decline feature in the production profile. Cubic spline interpolation is a powerful data analysis tool to fit a series of unique continuous and smooth cubic polynomials between each set of the data points (McKinley and Levine, 1998). The coefficients obtained from cubic spline interpolation are used to interpolate the data without discontinuity and erratic behavior. The detailed algorithm of cubic spline interpolation can be found anywhere in McKinley and Levine (1998).

2.6.4 Discrete Wavelet Transform

The techniques of parametrization of production profiles explained above are based on the observable decline patterns in this thesis. However, in many cases, identifying such patterns is challenging due to the small variation in production time-series data. Therefore, discrete wavelet transform (DWT) can be applied to approximate the production profiles instead of extracting decline patterns in this thesis.

Wavelet transform is an important input feature extraction technique and decomposes a function (or signal) into the shifted and scaled versions of the basic (mother) wavelet, $\Psi(t)$, which is a wave-shaped function with a zero mean and a limited length (Radunovic, 2009). It is capable of capturing both high- and low-frequency phenomena (Chen et al., 1999; Rioul and Vetterli, 1991). The wavelet function can be expressed as:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \quad \text{(2-11)}$$

where s and τ represent the scaling and translation parameter, respectively. The factor $\frac{1}{\sqrt{s}}$ is used to maintain constant energy at different values of scale. Continuous wavelet transform (CWT)

and discrete wavelet transform (DWT) are two common forms of its application. Although CWT could produce more comprehensive decomposition of the original signal, it is more computationally intensive since the calculation of CWT coefficients requires continuous modification of s and τ. DWT provides a feasible, yet efficient, alternative for the decomposition of time-series data.

In DWT, the original time-series is subjected to a low-pass filter and a high-pass filter. The components deriving from the low-pass filter are the approximation coefficients (cA), while the components deriving from the high-pass filter correspond to the detail coefficients (cD), which are usually regarded as noises and can be discarded. Through a down-sampling procedure, the cA coefficients can be halved, allowing only half of the samples for decomposition in the next level. This procedure can be repeated for a number of levels, and the iterative DWT process is illustrated in **Fig. 2-7**. The level of decomposition may affect the subsequent data-driven modeling in the next step. A lower level of decomposition (i.e., retaining more coefficients) would result in more input variables, higher degrees of freedom and stronger nonlinearity (possibly overfitting) in the data-driven models.

2.7 Reference

Al-Fattah, S. M., & Startzman, R. A. (2003). Neural network approach predicts U.S. natural gas production. *SPE Production & Facilities, 18*(02), 84-91.

Amirian, E., Leung, J. Y., Zanon, S. (2015). Integrated cluster analysis and artificial neural network modeling for steam-assisted gravity drainage performance prediction in heterogeneous reservoirs, *Expert Systems with Applications, 42*(2), 723-740.

Ampazis, N., & Perantonis, S. J. (2000). Levenberg-marquardt algorithm with adaptive momentum for the efficient training of feedforward networks. Paper presented at the *IEEE-INNS-ENNS International Joint Conference on Neural Networks,* Como, Italy, 126-131.

An, P., & Moon, W. (1993). Reservoir characterization using feedforward neural networks. *In SEG Technical Program Expanded Abstracts 1993*, 258-262.

Bahrololoum, A., Nezamabadi-pour, H., & Saryazdi, S. (2015). A data clustering approach based on universal gravity rule. *Engineering Applications of Artificial Intelligence, 45*, 415-428.

Barrett, J. P. (1974). The coefficient of determination—some limitations. *The American Statistician, 28*(1), 19-20.

Bhattacharya, S., & Nikolaou, M. (2013). Analysis of production history for unconventional gas reservoirs with statistical methods. *SPE Journal, 18*(05), 878-896.

Chen, B., Wang, X., Yang, S., & McGreavy, C. (1999). Application of wavelets and neural networks to diagnostic system development, 1, feature extraction. *Computers & Chemical Engineering, 23*(7), 899-906.

CMG, 2015. STARS: Users' Guide, advanced processes & thermal reservoir simulator (Version 2015), Calgary, Alberta, Canada: Computer Modeling Group Ltd.

Demuth, H., Beale, M., & Hagan, M. (2008). *Neural network toolbox™ 6*, *User's Guide.*

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Ge, J., Xia, Y., & Nadungodage, C. (2010). UNN: A neural network for uncertain data classification. *Advances in Knowledge Discovery and Data Mining* (449-460), Springer.

Gharbi, R. B., & Elsharkawy, A. M. (1999). Neural network model for estimating the PVT properties of middle east crude oils. *SPE Reservoir Evaluation & Engineering, 2*(03), 255-265.

Guan, B. T., Gertner, G. Z., & Parysow, P. (1997). A framework for uncertainty assessment of mechanistic forest growth models: A neural network example. *Ecological Modelling, 98*(1), 47-58.

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the marquardt algorithm. *IEEE transactions on Neural Networks, 5*(6), 989-993.

Hammouda, K., & Karray, F. (2000). A comparative study of data clustering techniques. University of Waterloo, Ontario, Canada.

Han, D., Kwong, T., & Li, S. (2007). Uncertainties in real-time flood forecasting with neural networks. *Hydrological Processes, 21*(2), 223-228.

Hanna, S. R., Chang, J. C., & Fernau, M. E. (1998). Monte carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmospheric Environment, 32*(21), 3619-3628.

Hasani, M., & Emami, F. (2008). Evaluation of feed-forward back propagation and radial basis function neural networks in simultaneous kinetic spectrophotometric determination of nitroaniline isomers. *Talanta, 75*(1), 116-126.

Haykin, S.S. (2008). *Neural networks and learning machines* (3rd ed.), Upper Saddle River, NJ, USA: Pearson.

Heaton, J. (2008). *Introduction to neural networks with java*, Heaton Research, Inc.

Jolliffe, I. (2005). *Principal component analysis*, Wiley Online Library.

Joo, S., Oh, S. E., Sim, T., Kim, H., Choi, C. H., Koo, H., & Mun, J. H. (2014). Prediction of gait speed from plantar pressure using artificial neural networks. *Expert Systems with Applications, 41*(16), 7398-7405.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). Segmenting time series: A survey and novel approach. *Data Mining in Time Series Databases, 57*, 1-22.

Khosravi, A., Nahavandi, S., & Creighton, D. (2013). Quantifying uncertainties of neural network-based electricity price forecasts. *Applied Energy, 112*, 120-129.

Kjræulff, U. B., & Madsen, A. L. (2008). *Bayesian networks and influence diagrams: A guide to construction and analysis*, Springer Science+Business Media.

Kröse, B., Krose, B., van der Smagt, P., & Smagt, P. (1993). *An introduction to neural networks*. The Netherlands: University of Amsterdam.

Lechner, J. P., & Zangl, G. (2005). Treating uncertainties in reservoir performance prediction with neural networks. Paper presented at the *SPE Europec/EAGE Annual Conference,* Madrid, Spain.

Lee, S. H., Kharghoria, A., & Datta-Gupta, A. (2002). Electrofacies characterization and permeability predictions in complex reservoirs. *SPE Reservoir Evaluation & Engineering, 5*(03), 237-248.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Anonymous California, USA, 1, 281-297.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

McKinley, S., & Levine, M. (1998). Cubic spline interpolation. *College of the Redwoods, 45*(1), 1049-1060.

Mezić, I., & Runolfsson, T. (2008). Uncertainty propagation in dynamical systems. *Automatica, 44*(12), 3003-3013.

Mohammadpoor, M., Firouz, Q., Reza, A., & Torabi, F. (2012). Implementing simulation and artificial intelligence tools to optimize the performance of the CO2 sequestration in coalbed methane reservoirs. Paper presented at the *Carbon Management Technology Conference*, Orlando, Florida, USA.

Nigrin, A. (1993). *Neural networks for pattern recognition*, MIT press.

Norman, C. D. (2013). Correlation of porosity uncertainty to productive reservoir volume. Paper presented at the *SPE Middle East Oil and Gas Show and Conference,* Manama, Bahrain.

Papadopoulos, C. E., & Yeung, H. (2001). Uncertainty estimation and monte carlo simulation method. *Flow Measurement and Instrumentation, 12*(4), 291-298.

Parada, C. H., & Ertekin, T. (2012). A new screening tool for improved oil recovery methods using artificial neural networks. Paper presented at the *SPE Western Regional Meeting*. Bakersfield, California, USA.

Petrović, D. V., Tanasijević, M., Milić, V., Lilić, N., Stojadinović, S., & Svrkota, I. (2014). Risk assessment model of mining equipment failure based on fuzzy logic. *Expert Systems with Applications, 41*(18), 8157-8164.

Radunovic, D. P. (2009). *Wavelets: From math to practice (1st Ed.)*, Springer Publishing Company, Incorporated.

Ramgulam, A. (2006). Utilization of artificial neural networks in the optimization of history matching (Doctoral dissertation, the Pennsylvania State University).

Refsgaard, J. C., van der Sluijs, Jeroen P, Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process–a framework and guidance. *Environmental Modelling & Software, 22*(11), 1543-1556.

Rioul, O., & Vetterli, M. (1991). Wavelets and signal processing. *IEEE Signal Processing Magazine, 8*(LCAV-ARTICLE-1991-005), 14-38.

Sarma, P., Durlofsky, L. J., Aziz, K., & Chen, W. H. (2007). A new approach to automatic history matching using kernel PCA. Paper presented at the *SPE Reservoir Simulation Symposium,* Houston, Texas, USA.

Scheevel, J., & Payrazyan, K. (2001). Principal component analysis applied to 3D seismic data for reservoir property estimation. *SPE Reservoir Evaluation & Engineering, 4*(01), 64-72.

Secchi, P., Zio, E., & Di Maio, F. (2008). Quantifying uncertainties in the estimation of safety parameters by using bootstrapped artificial neural networks. *Annals of Nuclear Energy, 35*(12), 2338-2350.

Smith, L. I. (2002). A tutorial on principal components analysis. Cornell University, USA, 51(52), 65.

Solomatine, D., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics, 10*(1), 3-22.

Srivastav, R., Sudheer, K., & Chaubey, I. (2007). A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. *Water Resources Research, 43*(10).

Stundner, M., & Al-Thuwaini, J. S. (2001). How data-driven modeling methods like neural networks can help to integrate different types of data into reservoir management. Paper presented at the *SPE Middle East Oil Show*, Manama, Bahrain.

Tan, S. S., & Smeins, F. E. (1996). Predicting grassland community changes with an artificial neural network model. *Ecological Modelling, 84*(1-3), 91-97.

Tang, H., Meddaugh, W. S., & Toomey, N. (2011). Using an artificial-neural-network method to predict carbonate well log facies successfully. *SPE Reservoir Evaluation & Engineering, 14*(01), 35-44.

Tiwari, M. K., & Chatterjee, C. (2010). Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs). *Journal of Hydrology, 382*(1), 20-33.

Verga, F., Viberti, D., & Gonfalini, M. (2002). Uncertainty evaluation in well logging: Analytical or numerical approach? Paper presented at the *SPWLA 43rd Annual Logging Symposium,* Oiso, Japan.

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, Jeroen P, van Asselt, M. B., Janssen, P., & Krayer von Krauss, Martin P. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment, 4*(1), 5-17.

Waszczyszyn, Z. (1999*). Neural networks in the analysis and design of structures*, Springer.

Wright, W. (1999). Bayesian approach to neural-network modeling with input uncertainty. *Neural Networks, IEEE Transactions On, 10*(6), 1261-1270.

Xu, S., & Chen, L. (2008). A novel approach for determining the optimal number of hidden layer neurons for FNN,Äôs and its application in data mining. Paper presented at the *International Conference on Information Technology and Applications: iCITA,* 683-686.

Zerafat, M. M., Ayatollahi, S., Mehranbod, N., & Barzegari, D. (2011). *Bayesian network analysis as a tool for efficient EOR screening*. Paper presented as the *SPE Enhanced Oil Recovery Conference*, Kuala Lumpur, Malaysia.

Zhang, X., Liang, F., Yu, B., & Zong, Z. (2011). Explicitly integrating parameter, input, and structure uncertainties into bayesian neural networks for probabilistic hydrologic forecasting. *Journal of Hydrology, 409*(3), 696-709.

Figures

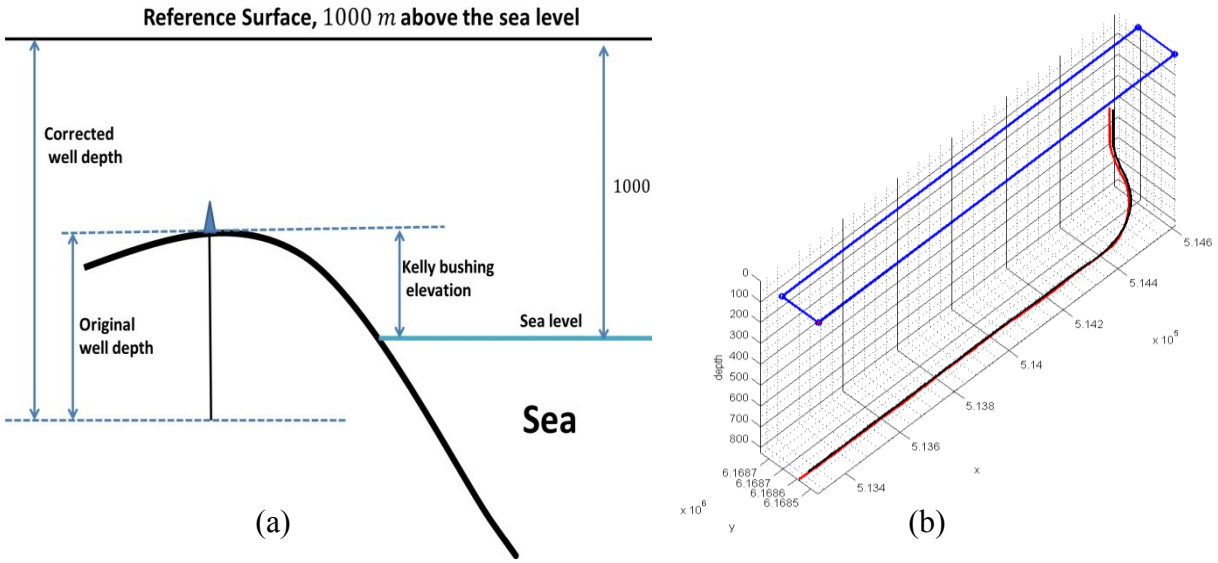


Fig. 2-1 Pre-process of logging data for data analysis: (a) – standardization of depth data according to an arbitrary reference surface; (b) – rectangular search domain around an injector-producer well pair in 3D view.

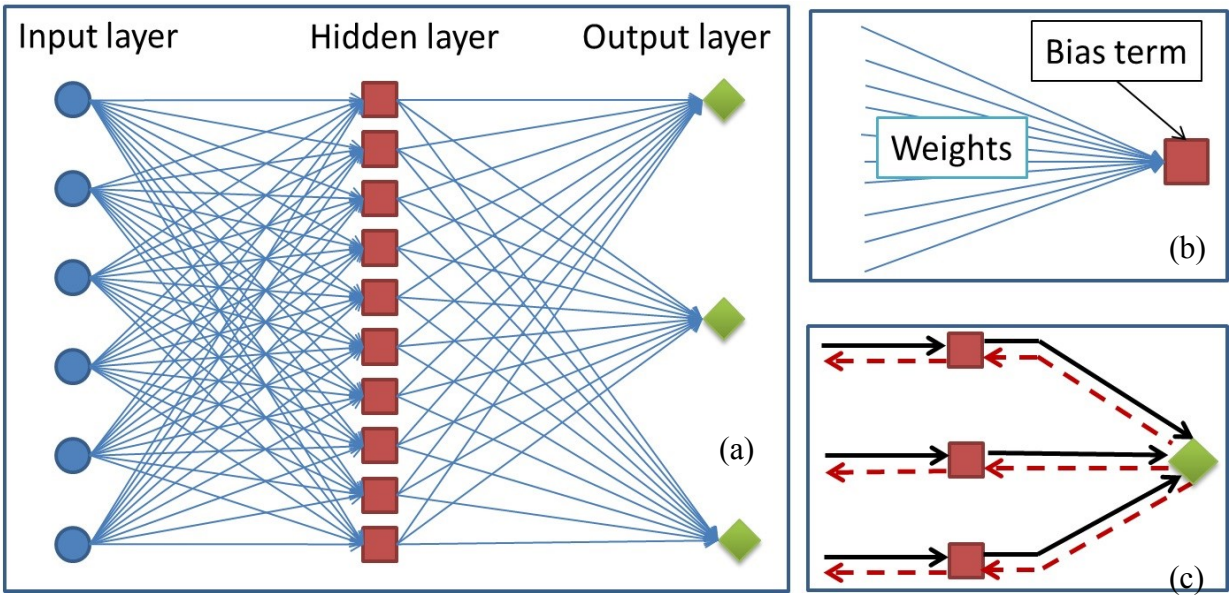


Fig. 2-2 Neural network structure: (a) – schematic of neural network architecture; (b) – a structure of neuron; (c) – transmission of signals; black solid arrows = function signals; red dashed arrows = error signals.

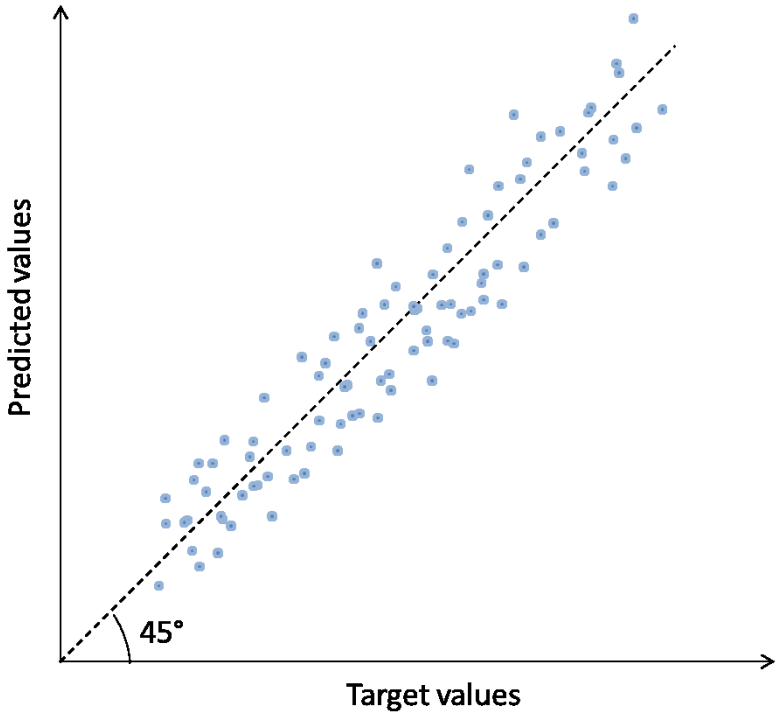


Fig. 2-3 A schematic of cross-plot.

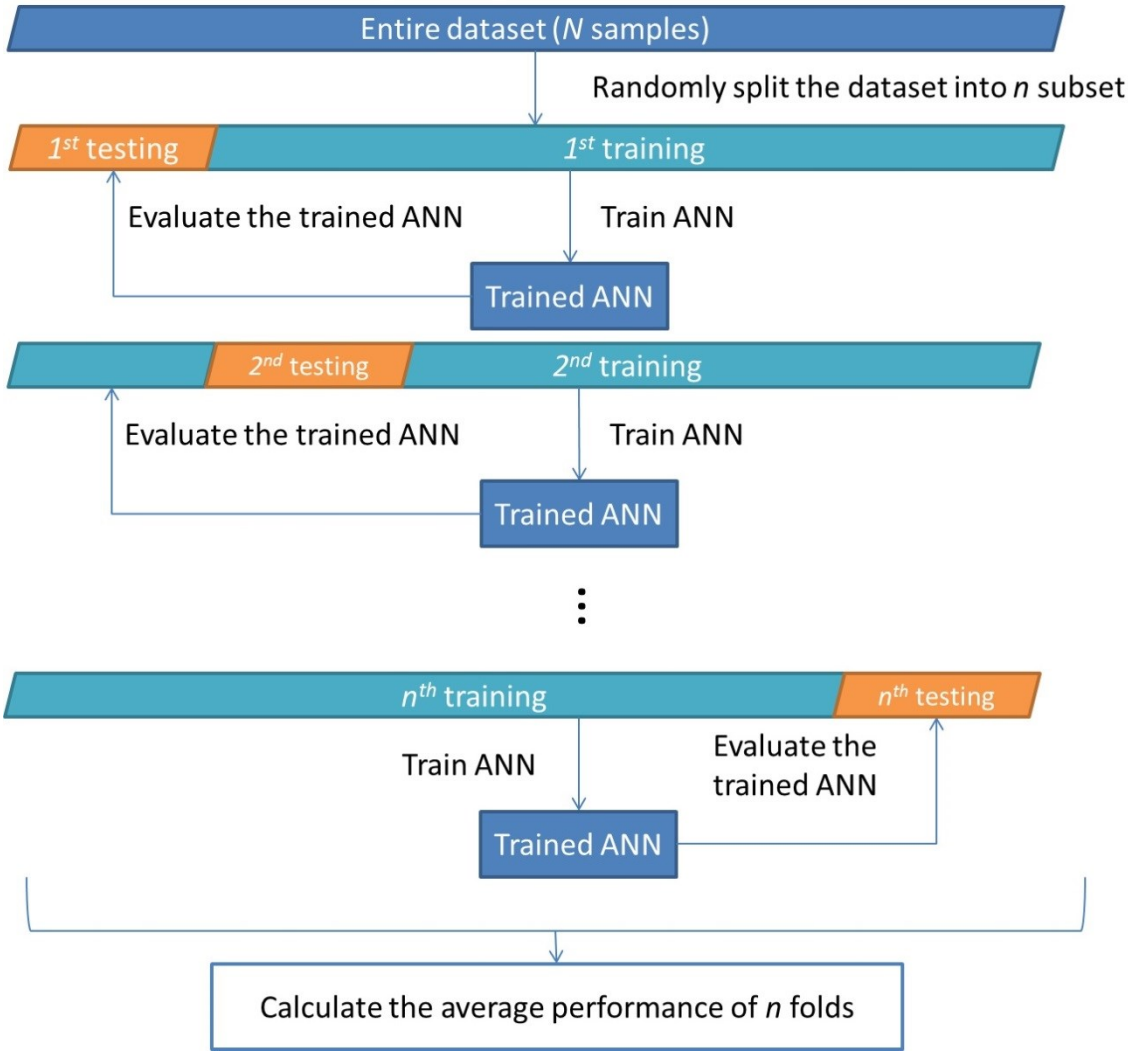


Fig. 2-4 Flowchart for n-fold cross-validation.

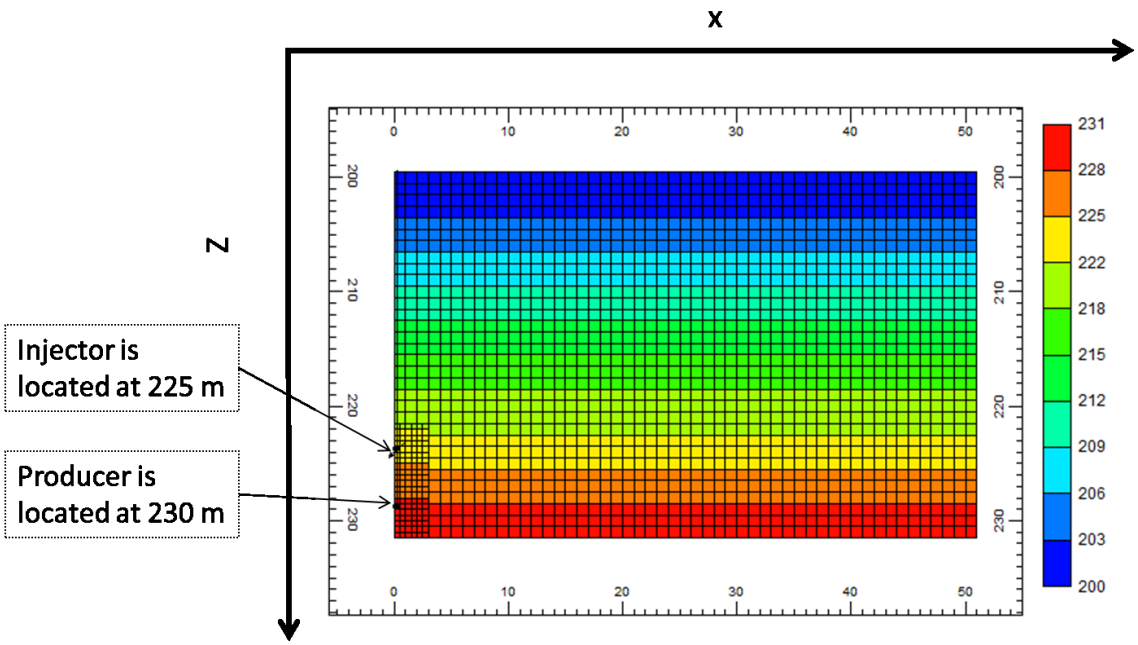


Fig. 2-5 Illustration of the 2D SAGD homogeneous models (only half of the distance between neighboring well pairs is incorporated).

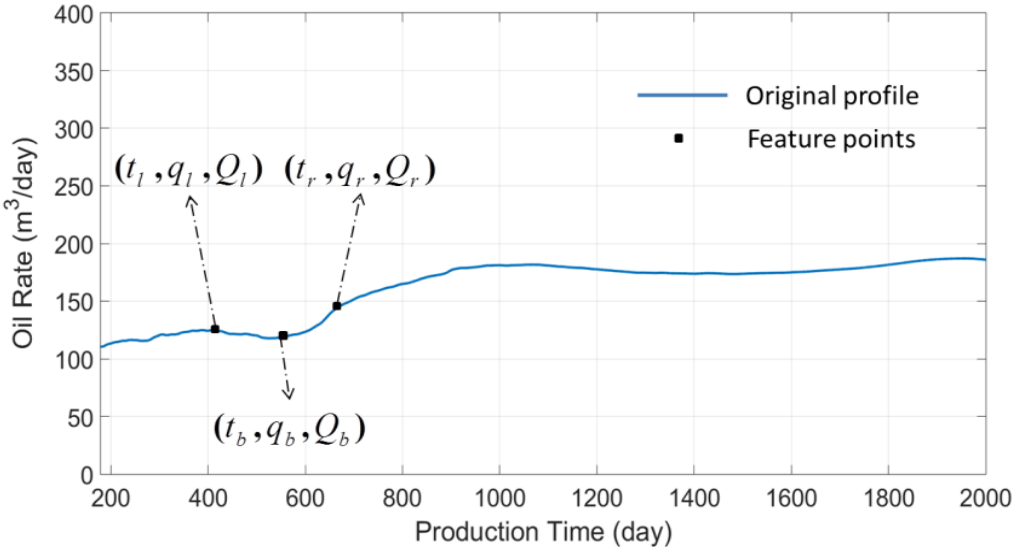


Fig. 2-6 Schematic of manual feature extraction from oil production rate time-series data.

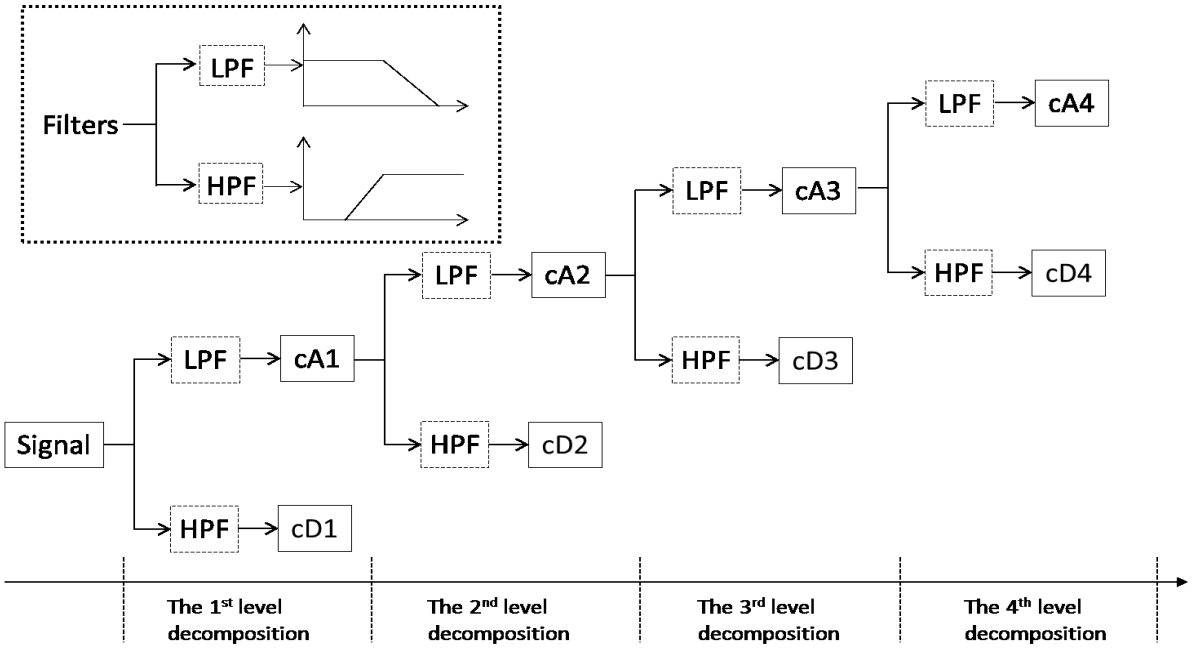


Fig. 2-7 Example of a time-series decomposition (i.e., feature extraction) using DWT involving a 4-level decomposition.

Chapter 3 Development of a Workflow for SAGD Production Analysis Using Knowledge-Based Approaches[1]

Chapter Overview

Quantitative appraisal of different operating areas and assessment of uncertainty due to reservoir heterogeneities are crucial elements in the optimization of production and development strategies in SAGD operations. Due to the apparent limitations of detailed compositional simulators, in this chapter, a novel knowledge-based approach for SAGD production analysis is developed to correlate reservoir properties and production performance.

A comprehensive training set encompassing SAGD field data compiled from a small number of publicly available sources is analyzed to develop a workflow and demonstrate its capabilities in the chapter. Exploratory data analysis (EDA) is carried out to interpret and extract relevant attributes describing characteristics associated with reservoir heterogeneities and operating constraints. An extensive dataset consisting of over 70 records is assembled. The extracted dataset is used to construct data-driven models using artificial neural network (ANN) to predict SAGD production. Predictions from the proposed approaches are both successful and reliable. Principal component analysis (PCA) is implemented to reduce the dimensionality of the input vector; statistical analysis is performed to analyze the uncertainties related to ANN model parameters and data.

This chapter illustrates that the proposed workflow is capable of predicting SAGD recovery performance from log-derived and operational variables. The analysis presents an

[1] Aversion of this chapter has been published as:
Ma, Z., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Practical implementation of knowledge-based approaches for steam-assisted gravity drainage production analysis. *Expert Systems with Applications*, *42*(21), 7326-7343.

important potential to be integrated directly into existing reservoir management and decision-making routines.

3.1 Introduction

SAGD is one of the most important thermal enhanced oil recovery technologies for producing heavy oil. A pair of horizontal wells, including an injection well and a production well that is located a few meters apart, are drilled into the reservoir. High-pressure steam is injected to reduce the bitumen viscosity. The heated crude oil then drains along the steam chamber edge into the production well by gravitational force. Evaluation of SAGD performance has been widely studied involving experiments (Akin and Bagci, 2001; Bagci, 2006; Shin and Polikar, 2006) and numerical simulation (Chang et al., 2012; Chow and Butler, 1996; Egermann et al., 2001; Fatemi, 2009; Siu et al., 1991). However, it is often impossible to reproduce all conditions and heterogeneities at the field scale in lab-scale models, while numerical flow simulations usually provide only approximate solutions to recovery responses, as numerous simplifications and assumptions must be invoked. The modeling process itself is also quite time-consuming, limiting its application in field-scale analysis involving multiple wells. Despite the availability of a large amount of production and reservoir data from different producing fields, practical application of knowledge-based models for reliable SAGD analysis and prediction is lacking.

Knowledge-based, or data-driven, modeling techniques, which entail comprehensive data analysis and implementation of machine learning methods for system forecast, provide an attractive alternative for the purposes of recovery performance prediction and uncertainty assessment, particularly when dealing with high-dimensional data space consisting of a large number of operational and geological parameters. In this work, neural network is employed to analyze SAGD production performance. Despite its recent implementation as a viable proxy for recovery prediction in design optimization (Popa et al., 2011; Queipo et al., 2002), its employment in data-mining frameworks for assessing heavy oil production performance in heterogeneous reservoirs is lacking (Ahmadloo et al., 2010; Popa et al., 2011). In particular, applied data-driven models involving actual field data from the McMurray bitumen deposits are rare. The dataset is compiled from extensive field data analysis including logging interpretation and production analysis. PCA is applied to reduce the dimensionality of the input vector,

alleviate the effects of over-fitting, and improve forecast quality. Uncertainty assessment of data-driven models is another area that is less explored in the applied expert systems literature, particularly when reservoir engineering data is involved. In this work, various sources of uncertainty including input data uncertainty, model parameter uncertainty, and model outcome uncertainty are quantified using Monte Carlo and bootstrapping approaches. In this application, data uncertainty primarily derives from inaccurate and incorrect data, limited number of records in dataset and imprecise (indefinite) analysis criteria. Model parameter uncertainty is common with most data-driven modeling techniques like ANN, whose training can be posed as an under-determined inverse problem with non-unique solutions.

The first objective of this research is to identify a description of pertinent predicting parameters (geologic, fluid, and operating) in relation to SAGD performance prediction to improve the predictability and accuracy of these models; EDA method is applied to extract a dataset from field data assembled from various public sources. The second objective is to demonstrate the potential of customizing applied knowledge-based modeling approaches for actual field data in providing practical tools suitable for SAGD performance prediction. In this work, ANN modeling approach is applied to predict cumulative production from a number of log-derived input attributes descriptive of both reservoir heterogeneities and operational conditions. The final objective is to propose new workflows to properly quantify and assess the uncertainties in data, model, and output. An important contribution of this work is that it demonstrates the feasibility of employing data-driven models for SAGD analysis using a realistic field dataset, a subject matter that insufficiently explored in the literature. Considering that many important data such as bottom-hole pressures, fluid properties, permeability, multi-phase flow functions, and thermal conductivities are commonly unavailable and, hence, are missing in the dataset, this work demonstrates how practical knowledge-based techniques can be used to construct data-driven models that are capable of predicting SAGD recovery performance from log-derived and operational variables. In addition, a novel uncertainty analysis workflow is implemented to quantify the impacts of individual uncertainty on the final model predictions.

The chapter is organized as follows: related works are summarized in section 3.2; materials and methods including the data analysis and dataset assembly, ANN technique, PCA approach and uncertainty analysis are presented in section 3.3; details of case studies and ANN modeling results are discussed in section 3.4; the paper is concluded in section 3.5.

3.2 Related Work

Data-driven modeling involves analysis of data characterizing the system of interest and focuses on using the machine learning methods to understand and build models that describe the behavior of the corresponding physical processes using experience, knowledge, and observed data generated from the processes of interest (Kjræulff and Madsen, 2008). Examples of some popular methods used in data-driven modeling are statistical methods, artificial neural network (Joo et al., 2014), and fuzzy logic (Petrović et al., 2014). The methods used nowadays have advanced significantly beyond the ones used in the conventional empirical regression. They are used for solving prediction problems, reconstructing highly non-linear relationships, performing data classification, and building rule-based expert systems. The general subject of data-driven modeling is developed with contributions from many overlapping disciplines including virtual intelligence, data mining, computational intelligence, machine learning, statistical data analysis, soft computing, and pattern recognition (Solomatine and Ostfeld, 2008).

ANN is a widely adopted data-driven modeling technique useful for identifying or approximating a complex non-linear relationship between input and target variables with only a limited number of assumptions about the "physical" behavior of the system. Learning the dependencies between inputs and corresponding outputs is the primary focus in data-driven modeling, and it is often accomplished using various supervised learning techniques (Solomatine and Ostfeld, 2008). Once a model is trained, it can be used to describe the behaviors and properties of this physical process. Compared to other function approximation techniques (e.g., response surface and Taylor expansion), ANN offers certain advantages including its capacity of inferring highly complex, nonlinear, and possibly uncertain relationships between system variables, requiring essentially zero prior knowledge regarding the unknown function (Hasani and Emami (2008). Many different learning tasks such as classification and non-linear function approximation can be well suited for ANN modeling. The first neural network model was introduced by McCulloch and Pitts (1943). After some major improvements and developments of ANN in recent decades, many formulations of neural network utilizing different transfer functions, learning algorithms, and network architectures (including hybridized fuzzy neural network) have been proposed and applied in various fields.

Applications of neural network can also be found in petroleum engineering. ANN has achieved significant popularity in areas such as production prediction (Al-Fattah and Startzman, 2003), reservoir characterization or properties prediction (An and Moon, 1993; Gharbi and Elsharkawy, 1999; Tang et al., 2011), history-matching (Ramgulam, 2006), classification (Stundner and Al-Thuwaini, 2001), proxy for prediction of recovery performance (Lechner and Zangl, 2005), production operation optimization and well design (Yeten et al., 2002). In recent years, the neural network has also been utilized to evaluate enhanced oil recovery projects (Parada and Ertekin, 2012; Zerafat et al., 2011) and assess CO_2 sequestration process (Mohammadpoor et al., 2012). ANN has been employed in the area of heavy oil recovery. For instance, it was used as a proxy model to forecast SAGD performance from operational parameters (Queipo et al., 2002) and to analyze production characteristics of cyclic steam injection process in homogeneous reservoirs (Popa et al., 2011; Popa and Patel, 2012); In the area of SAGD, Fedutenko et al. (2014) applied time-dependent radial basis function neural network to estimate oil production for the entire field, instead of for individual well pairs with varying reservoir and production characteristics. It is obvious that application of ANN in the analysis of SAGD process in heterogeneous reservoirs is still lacking. Amirian et al. (2014) applied ANN to estimate oil production from individual well pairs in layered heterogeneous reservoirs using a synthetic training dataset constructed from experimental design and numerical simulations. Although their results demonstrated significant potential in applying these data-driven approaches for recovery prediction through synthetic dataset, the feasibility of their approaches to actual field data was not demonstrated. In this work, a realistic field dataset is used for ANN model development and uncertainty analysis.

Previous studies have investigated the effects of heterogeneity on SAGD performance (Chen et al., 2008; Pooladi-Darvish and Mattar, 2002; Yang and Butler, 1992). Their results confirmed that SAGD performance is adversely affected by the presence of many long continuous shale layers, which hindered steam chamber expansion and fluid drainage. ANN and other data-driven modeling techniques have been applied recently to predict SAGD recovery performance in heterogeneous reservoirs (Amirian et al., 2013; 2014). The input attributes included a number of variables descriptive of reservoir heterogeneity and operational (well) parameters, while the corresponding output attributes were recovery factor and production profiles. In the study of Wang and Leung (2014), a number of input attributes were formulated to

parameterize characteristics of shale barriers and lean zones (e.g., locations, continuity, dimensions, proportions, and saturation) that are relevant to SAGD recovery performance. A ranking scheme accounting for both cumulative oil production and steam injection efficiency was proposed as the output attribute. Their results demonstrated that SAGD recovery efficiency decreases (i.e., the impedance of steam chamber advancement and obstruction of oil drainage) if the distance between the shale barrier and the well pair decreases, or if the volume (length and thickness), proportions, or continuity of the shale barrier increases.

The lack of data-driven models related to heterogeneous SAGD reservoir analysis in the literature motivates the development of ANN as an alternative tool to predict recovery performance in SAGD process in heterogeneous reservoirs. Most importantly, according to authors' knowledge, practical examples of data-driven models with an actual SAGD field dataset have not been published. Therefore, an actual SAGD field dataset is assembled to construct a set of ANN models. Practical challenges associated with the data assembly and analysis processes due to noises, errors, and missing data are explained in detail. In particular, strategies for (1) identifying/parameterizing pertinent predicting parameters; (2) handling of the high-dimensional dataset; and (3) assessing uncertainties in dataset and model parameter are discussed.

Analyzing and assembling a comprehensive dataset consisting of sufficient reliable data samples is the fundamental step in the model training process. This data analysis process typically entails three key steps: (1) planning and organization, (2) collection and analysis, and (3) integration and storage (Aly and Mahmoud Abu El Ela, 2007). Data relevant to petroleum applications are derived from diverse sources including seismic data, geological interpretation, log measurement, core analysis, fluid analysis or other laboratory measurements, well test data, and production data. Collection, combination, and analysis of such large amount of data would be a formidable task. Therefore, it is important to identify a description of all pertinent predicting parameters (geologic, fluid, and operating) in relation to SAGD performance prediction. This task can be accomplished by EDA, which refers to a group of statistical techniques useful for summarizing and extracting important attributes, detection of outliers, and definitions of appropriate assumptions and models relevant to the data (Tukey, 1977). In petroleum engineering field, EDA is often applied when handling a large dataset. For example, EDA was used to analyze reservoir and production data to identify and rank areas of possible production

improvement for a mature waterflood (Jansen and Kelkar, 1996). Holdaway (2009) discussed common EDA steps for characterizing reservoirs with large datasets from various sources.

The data should consist of a set of records composed of both input (predicting) attributes and the corresponding output (target) attributes. The entire dataset is usually divided into three portions to be used during the three stages in ANN model construction. These stages include (1) training (applying certain learning algorithms to calibrate the network parameters including weights and biases; (2) testing (evaluating the performance of trained network and optimizing the network architecture); and (3) validating (verifying the network performance using data that has not been previously presented to the network during the first two stages).

The prediction quality of ANN is often compromised due to the high-dimensional input vector, probable inter-correlation between predicting variables, and limited records in training data. Principal component analysis (PCA) can help to increase model robustness by reducing the dimensionality of the original dataset while retaining much of its information (variation) (Jolliffe, 2005). This is achieved by orthogonally transforming the dataset into a new set of uncorrelated variables or principal components, which are computed from an eigenvalue decomposition of the covariance matrix (Smith, 2002). The PCA technique has been successfully applied in areas including history matching (Sarma et al., 2007; Yadav, 2006), reservoir property estimation (Dadashpour et al., 2011; Lee et al., 2002; Scheevel and Payrazyan, 2001), and production data analysis (Bhattacharya and Nikolaou, 2013).

The analysis of uncertainty has been received much attention in recent years (Mezić and Runolfsson, 2008). Walker et al. (2003) explained various sources of uncertainty including input data uncertainty, model parameter uncertainty, and model outcome uncertainty, which is the accumulated uncertainty in the predicted values. In this application, data uncertainty primarily derives from inaccurate and incorrect data, a limited number of records in dataset and imprecise (indefinite) analysis criteria. Model parameter uncertainty is common with most data-driven modeling techniques like ANN, whose training can be posed as an under-determined inverse problem with non-unique solutions.

Three main groups of techniques are commonly adopted for uncertainty assessment: the Gaussian approach, the Monte Carlo method, and bootstrap method. Gaussian methods assume that distributions of uncertainty, including those exhibited by the input data and model parameters, are Gaussian. Specific applications of the Gaussian approaches for uncertainty

quantification can be divided into several categories: analytical error propagation equation method (Refsgaard et al., 2007; Verga et al., 2002); Bayesian approach (Nigrin, 1993; Wright, 1999; Zhang et al., 2011); and uncertain neural network method (Ge et al., 2010). The Monte Carlo method refers to a general stochastic approach for approximating the probability of a certain outcome by random sampling of a large number of realizations. The application of the Monte Carlo simulation in uncertainty analysis and error propagation can be found in a number of works (Guan et al., 1997; Hanna et al., 1998; Norman, 2013; Papadopoulos and Yeung, 2001). Bootstrap approach, introduced by Efron (1979), is a resampling method to estimate the statistic properties of a given sample dataset. The combination of bootstrap approach with the ANN is the bootstrapped neural network, which was employed to estimate safety margins with appropriate confidence intervals (Secchi et al., 2008) for the nuclear power plant. Examples for applications of bootstrap method for uncertain analysis include flood forecasting (Han et al., 2007; Tiwari and Chatterjee, 2010) and electricity price prediction (Khosravi et al., 2013).

Despite the availability of aforementioned uncertainty analysis approaches, discussion of comprehensive aggregated assessment of uncertainties stemming from various sources including data and model parameters is lacking among the expert systems literature, especially when error-prone actual field data is involved.

3.3 Methodology

3.3.1 Field Data Analysis and Dataset Assembly for Data-Driven Modeling

In this work, a set of SAGD field data assembled from the public domain is studied. The available data can be sub-divided into the following categories: Logging, Production, Injection, Well Header, Well Pair, and Deviation Survey:

- Logging: well log data including original petrophysical logging measurements such as spontaneous potential (SP), gamma ray (GR), true resistivity (RT), and a number of interpreted logs such as reservoir porosity and water saturation;
- Production: production rate, production time, and cumulative production for oil, gas, and water;
- Injection: steam and water injection volumes;

- Well Header: surface and bottom locations in different geographic coordinate systems, well depths, Kelly bushing elevation, and other relevant drilling and well completion information;
- Well Pair: producing field, pad, and associated horizontal well pairs;
- Deviation Survey: well trajectories of horizontal wells.

Well depth is measured from the elevation of Kelly bushing, which varies from well to well. Therefore, it must be normalized against a reference surface. This pre-processing step is illustrated in **Fig. 3-1**(**a**). It should be mentioned that vertical wells are drilled as delineation wells, while horizontal wells are drilled as producers and injectors. Logging information is available only at the vertical wells, while rest of the data are associated with the horizontal wells. A data record is assembled by combining logging information from the nearby vertical wells and operating/production information from the associated horizontal wells.

A number of input/output attributes describing the reservoir properties and production characteristics, including porosity (ϕ), net-to-gross (N/G) ratio, pay zone thickness (h), fluid saturation, injection and production data, can be extracted after combining and analyzing these field datasets. Although other information such as core analysis and seismic data might be available for a portion of the dataset, we are interested in constructing records that have the same number of input and output attributes. In this chapter, the input attributes related to reservoir properties are extracted directly from logging interpretation, while for the inputs pertinent to operation conditions are extracted from analyzing the injection and production datasets. For each well pair, reservoir information is extracted from the closest vertical logging wells. This is facilitated by assigning a rectangular search domain around the well pair, as shown in **Fig. 3-1**(**b**). The blue rectangle outlines the boundaries of the search domain for this well pair example. The red and black solid lines denote the well trajectories of the producer and the injector in this well pair, respectively. The logging wells that are located in the rectangular domain are selected for interpretation. If no logging wells can be found within the domain, or if a significant number of logs are missing such that no reliable interpretation is possible, this well pair would be excluded from the dataset. If a particular logging well is located nearby multiple horizontal well pairs, its interpreted values would be assigned to only the closest well pair.

Logging interpretation is carried out using the GR, SP, and RT logs to extract the required input attributes. Various cutoff values are assigned to these logs to identify sand/shale and

water/oil zones associated with each well. The selection of cutoff value depends on the specific logging tool and formation characteristics. Net pay is subsequently defined as sand saturated with oil. **Fig. 3-2**(**a**) presents an example of the interpreted results. The blue layers denote the non-pay zones, while the red layers represent pay zones. If the thickness of a shale layer embedded within the gross pay is less than 2 m, it is regarded as part of the net pay. Five variables including porosity (ϕ), water saturation (S_w), pay zone thickness (h), net-to-gross ratio (N/G), and shale index (SI) are computed. The definitions of gross pay, net pay and non-pay are explained in **Fig. 3-2**(**a**), in which gross pay is the thickness of the entire pay interval, while net-to-gross ratio is total thickness of net pay intervals divided by the gross pay. The porosity and water saturation are calculated as arithmetic averages of interpreted log values over the entire gross pay. The variable SI is a normalized shale continuity indicator defined as the thickness-to-distance ratio of the shaly layer located at the shortest distance to the injector, i.e., h_{sh}/d_{sh_inj}, described in the **Fig. 3-2**(**a**). The black solid line denotes the injection well. A large value of SI would indicate a thick shale barrier that is located close to the injector, impeding the advancement and growth of the steam chamber.

These five log-derived variables are considered as part of geologic input attributes for the ANN modeling. If multiple logging wells are found within the search domain, average values calculated over all logging wells are assigned to the input attributes. Although the formulation of additional input attributes (e.g., distance from top of pay zone to the injector) have been investigated, sensitivity analysis and findings from Amirian et al. (2014) suggest that these five variables are sufficient to capture the influences of heterogeneity, as observed from log data, on SAGD performance. Four additional input attributes relevant to well and operational conditions are incorporated in this work: effective number of oil production wells (N_e^{prod}) and effective number of steam injection wells (N_e^{inj}), total production period (T_{total}) of the given well pair, and cumulative steam injection (CSI). With 9 input attributes, there is a single output attribute of cumulative oil production (COP). The effective well numbers are introduced to describe well configurations with more than two wells (injector-producer pair) sharing the same drainage area. For example, in addition to the primary wells, reentry and infill wells are occasionally drilled as secondary wells to enhance production. Given the production/injection period and contribution to COP and CSI varies between each well, a time-weighted average (effective) well number is formulated as:

$$N_e^{prod} = \sum_{i=1}^{m_{prod}} \frac{T_i^{prod}}{T_{total}} \quad \text{.....(3.1)}$$

$$N_e^{inj} = \sum_{i=1}^{m_{inj}} \frac{T_i^{inj}}{T_{total}} \quad \text{.....(3.2)}$$

in the above definitions, N_e^{prod} and N_e^{inj} are effective number of producers and injectors of the well pair, respectively; m_{prod} and m_{inj} refer to the actual number of producers and injectors in this well pair, including all primary, re-entry, and infilled wells; i represents the well index; T_i^{prod} and T_i^{inj} are the corresponding actual production and injection time of i_{th} producer or injector. Finally, the total production period T_{total} is defined as the time period between first and last production dates. **Fig. 3-2**(**b**) illustrates the calculation of N_e^{prod} for a number of production scenarios. In this work, well pairs with N_e^{prod} much less than one is removed from the dataset to avoid bias due to extensive shut-in. CSI and COP refer to the cumulative volumes of injected steam and produced bitumen, respectively. Following the aforementioned analysis procedure, a data sample containing 9 input variables and 1 output variable can be formulated for a well pair once all required information are available.

3.3.2 Artificial Neural Network

The neural network is a kind of machine learning algorithm; it is widely used for pattern recognition and prediction by mimicking the information transfer in the central nervous system of human (Haykin, 2008). The basic neural network architecture is composed of an input layer, an output layer, and any number of hidden layers. The output layer is made of the target variables, while the input layer consists of attributes that are related to the target variables. A neural network with only the input and output layer is called a single-layer perceptron (SLP), which can be applied in problems that are linearly separable. In other cases, the multi-layer perceptron (MLP) neuron network is implemented. The MLP may contain any number of hidden layers, which serve to transform the original input data space into new spaces, where it is easy to perform the classification or regression process. The MLP is the most widely adopted perceptron in solving problems with real data. **Fig. 3-3**(**a**) shows an example of the basic architecture of the

MLP neuron network, which has one hidden layer; the blue circles, red squares, and green triangles denote neurons in the input layer, hidden layer, and the output layer, respectively. It is a fully connected neural network since every neuron in the network is connected to the nodes in its adherent layers.

The performance of ANN modeling would be affected by the number of hidden layer nodes (Tan and Smeins, 1996). Too many neurons (or connections) may lead to overfitting problem, while the prediction performance of ANN is compromised with insufficient nodes. There are no concrete guidelines to determine the number of free parameters in the hidden layer. In order to enhance computational efficiency and network predictability, the network architecture including the number of hidden layers, the number of neurons in hidden layers, choice of activation functions should be optimized by balancing between prediction accuracy and overfitting. This number of hidden nodes is typically considered to vary as a function of input vector dimension and the amount of training data. Several relationships or rules of thumb exist in the literature relating the training-dataset size to some user-defined error parameters calculated for a given network configuration (Waszczyszyn, 1999; Xu and Chen, 2008). A recent review of a range of design issues related to ANN development in petroleum industry can be found in Al-Bulushi et al. (2012). It was demonstrated that a single hidden layer could approximate any function with a finite number of discontinuities (Kröse et al., 1993). Heaton (2008) showed the number of neurons should be less than two times of the number of input parameters. Although there are some rules of thumb to select the number of hidden nodes have been proposed in some previous studies, it is common to select the optimum number of hidden neurons by trial and error. In this work, the n-fold cross-validation approach (which will be discussed hereafter) is used to determine the optimal network structure.

We implemented a feedforward backpropagation neural network, where the error is back propagated to train the network parameters (weights and biases assigned to each connection) in a supervised learning algorithm. Two kinds of signals are transferred in the feedforward neural network. The first one is function signal (or input signal), which comes from the input neurons and propagates forward through the hidden layers to the output layer; another one is error signal, which is generated from the output neurons and propagates backward through the hidden layers to the input layer (Haykin, 2008). A schematic of signal transfer is shown in **Fig. 3-3**(**b**), where black solid arrows and red dashed arrows denote the function signals and error signals,

respectively. Values of weights and biases are updated using a training dataset such that the mismatch between network predictions and known values of the target variables is minimized (Francis, 2001).

The nodes between neighboring layers are connected by weights, as shown in **Fig. 3-3**(**b**). The input signal x of a certain node in the hidden or output layer is the weighted summation of output signals from previous layer, as **Eq. 3-3** shows:

$$x_i = b_i + \sum_{j=1}^{m} w_{ij} \cdot y_j \quad \text{.....(3-3)}$$

where x_i is the weighted sum of input signals at node i in the current layer; y_j denotes the output value of node j in preceding layer; m represents the total number of nodes in the preceding layer; b_i is the threshold (bias) value; w_{ij} is the weight associated with the connection between node i and node j. To calculate the output value of this neuron, the transfer function (or activation function) is applied to the weighted sum. Various transfer functions such as pure line function, threshold function, and sigmoid function (e.g., the hyperbolic tangent function and logistic function) can be used. In this work, the hyperbolic tangent function, as shown in its general form in **Eq. 3-4**, is used:

$$f(x) = \tanh(x) \quad \text{.....(3-4)}$$

where x is an independent variable. The outputs of the hyperbolic tangent function are in the range of (-1, 1). It scales large positive and negative input values to +1 and -1, respectively. The output of node i or y_i can be computed by **Eq. 3-5**:

$$y_i = f(b_i + \sum_{j=1}^{m} w_{ij} \cdot y_j) \quad \text{.....(3-5)}$$

The value calculated from **Eq. 3-5** is the output signal from node i, which can be considered as the input signal to the next layer. **Eqs. 3-3** to **3-5** are repeated until the final output layer is reached and predicted value for the output variable is calculated. These equations allow

the function signals to be propagated from the input layer to the output layer. The backpropagation algorithm is applied to update the weight and biases during the learning stage. The classical backpropagation algorithm is a gradient descent supervised learning algorithm. An error signal is computed as the mismatch between target and prediction at the output layer. The weights of the output layer are updated based on estimated derivatives of error with respective to weight. This step is repeated to transmit the error signal in the reverse direction until all weights are updated. This entire updating procedure must be repeated for many epochs until a certain stop criterion is reached. Since backpropagation is a gradient descent method, the transfer functions must be differentiable.

Even though the gradient descent backpropagation is feasible for most problems, if the step-size is properly selected, it bears some limitations such as slow convergence. In order to enhance computational efficiency, various modified schemes have been developed including the Levenberg–Marquardt backpropagation (LM-BP) algorithm. The Levenberg–Marquardt backpropagation algorithm was proved as an efficient method to update weights and biases of neural network (Ampazis and Perantonis, 2000; Hagan and Menhaj, 1994). The details of the LM-BP algorithm can be found in a number of references (Hagan and Menhaj, 1994; Haykin, 2008), and it is applied in this study.

Due to the large disparity in scales of different data sources, normalization or standardization procedure is often performed (Francis, 2001). Normalization is an important pre-processing step for ANN modeling, with all data values being transformed to vary between a certain range such as [0, 1] or [-1, 1]. This step can help to reduce bias in the minimized solution as a result of the overwhelmingly large data values (Al-Fattah and Startzman, 2003). A data point x can be normalized by **Eq. 3-6**:

$$x^N = 2 \times \frac{x - x_{\min}}{x_{\max} - x_{\min}} - 1 \quad \text{.....(3-6)}$$

where x^N is the normalized data value ranging between -1 and 1; x_{max} and x_{min} represent the maximum and minimum value of this data vector, respectively.

To design the optimum architecture of the neural network, the n-fold cross-validation method is implemented in this study. The training dataset is divided into the n equal size subsets

randomly. From the n subsets, one is selected as the validation part while the remaining $n-1$ subsets are assigned as the training part. First, for a particular network structure, the training part is used to train the corresponding network parameters (weights and biases); the network performance is subsequently evaluated using the validation part. The mean squared error (*MSE*) between target and prediction is computed as a measure of network performance. This step is repeated numerous times with different random initial solution of the weights and biases, and the solution with the lowest *MSE* is selected. Next, another subset is selected as the validation part, and the remaining subsets are assigned as the training part. The training-validation process is repeated for n times to calculate an average performance (i.e., average *MSE*) of this network structure. This entire procedure is carried out again for another parameter exploration (network structure). Finally, the optimum neural network architecture with the best average performance is determined.

The ANN implementation workflow is shown in **Fig. 3-4**. An original dataset consisting of N samples is assembled from the data analysis process. The total dataset is divided into two parts: the first part includes k samples and is designated as the final testing dataset while the remaining $N-k$ samples are used in the training (experiment) stage to determine and train the optimum architecture of the neural network. The n-fold cross-validation, as previously outlined, serves to identify the optimal number of hidden layer(s) and number of neurons in the hidden layer(s). Once the optimum architecture is determined, the $N-k$ samples are partitioned into two subsets: training (90%) and validation (10%), which are used to train the optimal network in one final learning step. Numerous runs are conducted with different random initial solution of the weights and biases, and the solution with best performance (i.e., lowest *MSE*) is chosen. The final testing dataset is used to test the prediction performance of the trained network. In this study, the neural network modeling is implemented using the Neural Network Toolbox (Demuth et al., 2008) in Matlab™.

ANN prediction performance is also assessed through the coefficient of determination (R^2), whose value ranges between 0 and 1. R^2 is commonly used as a goodness-of-fit indicator in the linear regression model. In this work, *MSE* and R^2 between ANN outputs and target values are calculated to evaluate ANN prediction performance. A good prediction is indicated by low *MSE* and large R^2.

3.3.3 Principal Component Analysis

PCA is performed to reduce the dimensionality of the original dataset. First, the mean of each dimension is subtracted from the original data:

$$Z_{ij} = X_{ij} - \bar{X}_j \quad \text{.....(3-7)}$$

where X_{ij} is the j_{th} variable of i_{th} sample, $\bar{X}_j$ is the average of X_j over all N samples, while Z_{ij} is the new variable that represents deviation from the mean. The purpose of this step is to simplify the calculation for the covariance matrix and remove bias due to large disparity in mean values. Next, the covariance between two variables X_j and X_k is defined as:

$$COV(X_j, X_k) = \frac{\sum_{i=1}^{N} Z_{ij} Z_{ik}}{N-1} \quad \text{.....(3-8)}$$

where N is the number of data samples; i denotes sample index, j and k are dimension indices of two variables. Finally, eigenvalue decomposition of the covariance matrix (**Eq. 3-8**) is carried out. Individual eigenvalue represents the significance or contribution of the variance from the corresponding eigenvector to the total variance of the original data. The eigenvectors with highest eigenvalues are principal components (*PC*), which can be obtained by sorting the eigenvalues in a decreasing order. Once the principal components have been identified, the original dataset is transformed into the principal component space as principal scores (*PS*) according to the following equation, which are regarded as the inputs attributes in subsequent ANN modeling.

$$PS = PC \times Z^T \quad \text{.....(3-9)}$$

3.3.4 Uncertainty Analysis

In this work, data uncertainty because of the small size of dataset and imprecise analysis criteria, together with model parameter uncertainty due to training algorithm and initialization, are

investigated. The aggregated consequence of these uncertainties is exhibited in the output (prediction) uncertainty. A comprehensive analysis involving all the aforementioned uncertainties with an actual SAGD dataset is novel. First, model parameter uncertainty is quantified with a Monte Carlo framework, in which training of the optimum network is repeated with many randomized initializations of model parameters. Next, parametric bootstrapping is performed to assess the data uncertainty introduced during the data analysis process (e.g., imprecise analysis criteria). Finally, bootstrapping with replacement is applied to evaluate the uncertainty stemming from limited dataset size.

Model Parameter Uncertainty: Model parameter uncertainty is common with most data-driven modeling techniques like ANN, whose learning is often posed as an under-determined inverse problem with non-unique solutions of weights and biases for a given deterministic training dataset. Different stop criteria or learning algorithms could also give rise to uncertainty. In this work, results with the Levenberg–Marquardt back-propagation algorithm, particle swarm optimization (PSO) and genetic algorithm (GA) are compared. In addition, uncertainty could also stem from random initializations; this uncertainty can be quantified with a Monte Carlo framework, in which training of the optimum network is repeated with many randomized initializations of model parameters. Aggregating the trained weights and biases derived from all initializations, the conditional probability **P(w|d)**, where w and d refer to the model parameter and data vectors, respectively, can be established. For a given testing sample, the corresponding output uncertainty is estimated by sampling multiple sets of w vectors from **P(w|d)**.

Data Uncertainty: The first source of uncertainty in the data is the results of outliers and imprecise cut-off values used in the logging interpretation, a common consideration in geologic data analysis. Uncertainty in input data is accounted for by estimating a likelihood function for each input attribute and performing parametric bootstrapping of this likelihood to assess uncertainty related to this input attribute. In this work, a detailed sensitivity analysis reveals that each input attribute follows approximately a uniform distribution with a +/- 10% variation in attribute value after applying different cut-off criteria. Next, N_r data records are sampled from these uniform likelihood functions; each sample can be regarded as a realization from the probability **P(d)** and is subjected to ANN training. This probability can be combined with **P(w|d)** to obtain **P(w,d) = P(w|d) × P(d)**. If we ignore the model parameter uncertainty here and consider only data uncertainty, **P(w,d) = P(d)**, a total of N_r trained networks are obtained;

therefore, for a given testing sample, the corresponding output uncertainty can be computed from predictions generated from all N_r trained models.

The second source of data uncertainty is a result of limited dataset size. The collected dataset is only one of an infinite number of possible datasets that may be drawn in a certain input domain (Srivastav et al., 2007). The variability of sampling the input and target values could lead to the uncertainty in training dataset. Srivastav et al. (2007) demonstrate how the bootstrap approach can be applied to quantify the uncertainty due to data size. If ANN approach is applied as the regression function (f) for prediction, for a particular input vector x_n, the output vector can be presented as $y_n = f(x_n;w)$, where $n = 1, \ldots, N$ (total number of records). Considering that **P(w|d)** would vary when modeling with a different realization of all possible sample datasets, the uncertainty or variance of the distribution of y_n can be estimated by the bootstrap technique, where B_n additional datasets are sampled randomly based on the original dataset with replacement. The network is trained using B_n datasets to obtain B_n trained network and the output variance for a given new input from testing dataset can be directly calculated.

3.4 Results and Discussion

3.4.1 Case 1 – Original Input Variable Space:

The dataset employed in this work is extracted from three nearby producing fields consisting of Cristian Lake, Foster Creek, and Jackfish with comparable reservoir conditions. Locations of the three SAGD fields are displayed in **Fig. 3-5**. These fields are part of the south Athabasca oil sand, which is located in northeastern Alberta, Canada. They are selected to be part of this study because of their geographical proximity, similarity in reservoir conditions, and large well count. There are over 1300 wells in total, with approximately 500 well pairs; most of them have been producing for over 3 years. This large number of wells poses various challenges. In many instances, there is insufficient data to formulate a complete record of input and output attributes for a given well pair. As previously mentioned in the data analysis, certain issues can be addressed by utilizing a search domain and averaging of numerous logging wells.

After applying the aforementioned criteria, a total of 71 well pairs or complete records (samples) are extracted. Every record contains 9 attributes as the input parameters, including porosity (ϕ), water saturation (S_w), pay zone thickness or gross pay (h), net-to-gross ratio (N/G),

shale index (SI), effective number of producers (N_e^{prod}), effective number of injectors (N_e^{inj}), total production time (T_{total}), and cumulative steam injection (CSI), computed as described in the data analysis. The cumulative oil production (COP) is designated as the single output attribute. These 10 continuous variables computed from the aforementioned data analysis are regarded as input parameters directly without further transformation. **Table 3-1** summarizes some important statistical properties including average, standard deviation, minimum and maximum value of each variable. Histograms of all 10 input and output attributes are displayed in **Fig. 3-6**(**a**). The corresponding cross-plots between the individual input attributes and COP are shown in **Fig. 3-6**(**b**). Except for total production time and cumulative steam injection, which are positively correlated against COP, cross-plots of most variables with COP do not reveal obvious correlation patterns; however, non-linear relationships between these attributes might still exist.

As illustrated in **Fig. 3-4**, the entire dataset is first divided into the experiment part ($N - k$ = 60 samples) and final testing part (k = 11 samples) randomly. The experiment part is used to determine the optimum network structure and to train the corresponding network parameters (weights and biases), while the final testing part is used to test the performance of the trained model. A 5-fold cross-validation is implemented considering the size of the dataset. The *MSE* between the target values and the actual outcomes from neural work is evaluated. A number of architectures with one or two hidden layer(s) are tested. For the single hidden layer cases, the number of hidden neurons vary between 3 and 25, while for the two hidden layer cases, the number of hidden neurons in each layer ranges between 3 and 15. The optimum architecture for one hidden layer is determined to have 8 hidden neurons, while the optimum architecture with two hidden layers is a 4 × 9 combination, respectively.

After the optimum numbers of neurons in hidden layers are determined, the networks are trained again with all the samples from the experiment dataset. To avoid overfitting and to increase generalization of network, the training process should be terminated prematurely based on certain overfitting criteria. In this work, the experiment part (60 samples) is divided into a training subset and a validation subset. The training subset is used to train the optimal network (i.e., estimate values of the weights and biases) while the validation subset is used to avoid overfitting. As training progresses, the error calculated from the validation is monitored, and it would typically reduce in a manner resembling that of the training error. However, as the number

of epochs increases, this validation error would start to increase, despite that the training error continues to decline. This trend indicates that the network is overfitting the training data. Therefore, once the validation error begins to increase after a certain number of epochs, the training process should stop where the validation error is still at its minimum (Demuth et al., 2008; Haykin, 2008)

Fig. 3-7(**a-b**) illustrates the performance of the single hidden layer neural network with 8 neurons. The comparison between ANN prediction of COP (y-axis) and the actual target value from field data (x-axis) is shown. The results for the training subset (60 samples) and the final testing subset (11 samples) are shown in **Fig. 3-7**(**a**) and **Fig. 3-7**(**b**), respectively. Good agreements can be observed with the training and testing data, as most points follow the 45^{o} separating line that indicates a perfect correlation. The overall performance of the single layer ANN model is acceptable, as indicated by the high R^2 value and low *MSE*. Performance of the two hidden layer ANN model is shown in **Fig. 3-7**(**c-d**). With performance similar to that of the single layer case, ANN predictions are in good agreements with the target values from field data. Comparison with **Fig. 3-7**(**a-b**) reveals that the ANN models with these two architectures produce very similar results. Despite the increased number of hidden layers, the overall performance of the two hidden layer ANN model is not particularly superior to the single layer model, since the differences between ANN predictions and target values are reduced significantly for both training and testing datasets with either configuration. With the number of data samples kept constant, model performance would improve with a reduced number of unknown parameters (hidden neurons) due to fewer degrees of freedom. On the other hand, the two hidden layer model with 9 input attributes employ many more hidden neurons than the single-layer case; therefore, improvement in the overall performance is hindered by the increased unknown parameters and degrees of freedom.

Fig. 3-8 shows an example of overfitting for single hidden layer ANN case when the number of epochs is too large. As shown in **Fig. 3-8**(**a**), performance of ANN during the training stage is essentially perfect with excellent matches between targets and outcomes. However, when the trained network is applied to the testing dataset, comparison of ANN prediction with target values is extremely poor as shown in **Fig. 3-8**(**b**). **Fig. 3-8**(**c**) shows the change in training and validation errors with epochs during the training stage. As expected, both training and validation errors decrease with increasing number of epochs; however, the validation error reaches a

minimum at the 11th epoch and begins to rise thereafter (while the training error continues to decline). Results shown in **Fig. 3-8**(**a-b**) are derived from weights and biases obtained at the end (i.e., the maximum number of epochs allowed). However, to avoid overfitting, training should be terminated at the 11th epoch, and the results are shown in **Fig. 3-7**(**a-b**).

Uncertainty analysis is performed next. Three aspects of the uncertainty assessment, including uncertainty in model parameter, uncertainty in input variables due to imprecise analysis criteria and uncertainty in dataset due to a limited number of records, are estimated based the approaches presented in the methodology section.

First, uncertainty in model parameter is assessed with the Monte Carlo method, in which training is repeated 1000 times with randomized initialization of weights and biases in network. The LM –BP algorithm is chosen as the learning algorithm for all 1000 trainings with the identical train-stopping condition. Results with the two-hidden layer case (4 × 9) are shown in **Fig. 3-9**. As a result, 1000 outputs are generated corresponding to the 1000 trained ANN models for each testing data. **Fig. 3-9**(**a**) presents the box plots of the testing dataset, and many outliers (represented as red plus signs that are located beyond the lower and upper limits defined by the interquartile range) can be detected. Another way of representing this uncertainty is to define an error bar to be the standard deviation of all output values, and the results are shown in in **Fig. 3-9**(**b**). For particular testing data (e.g., the 10th and 11th sample), the variance in ensuing prediction is significant. Histograms of the testing dataset are also presented in **Fig. 3-9**(**c**). For all 11 distributions, the most probable (one with the highest probability value) output is closely approximating the actual target value.

Next, a parametric bootstrapping workflow is implemented to assess the input data uncertainty. A total of N_r = 1000 realizations of the original data records are sampled from the uniform likelihood functions (as described in the methodology section) and subjected to ANN training. Uncertainties of the 9 input attributes with uniform likelihood functions (10% variation) of two randomly-selected records are shown in **Table 3-2**. Once again, for a given testing sample, the corresponding output uncertainty can be computed from predictions generated from all N_r trained models, and the results are shown in **Fig. 3-10**. It should be noted that if a skewed (asymmetric) likelihood function is inferred from the data (instead of the uniform distribution with symmetric +/- 10% standard deviation), the variance of the ensuing output distribution might increase. Similar to the results in **Fig. 3-9**, the most probable output is in good agreement

with the target value for each testing data. However, the boxplots and errorbar plots in **Fig. 3-10**(**a**-**b**) reveal that impacts due to uncertainty in input variable are less significant as compared to model parameter uncertainty. This is evidenced by the fewer outliers and reduced variances observed for each testing sample.

The bootstrap approach is applied again to quantify the uncertainty due to a limited number of data records. One thousand new datasets with replacement are sampled based on the original dataset. Each new dataset will serve as an independent dataset to train a separate ANN model. The corresponding output uncertainty for the 11 testing samples can be computed from predictions generated from all 1000 trained models, and the results are shown in **Fig. 3-11**. It is interesting to note that the uncertainty due to dataset size is slightly higher than the uncertainty to input attribute variability, but it remains smaller than the uncertainty in model parameters.

Table 3-3 summarizes the uncertainty (in the form of standard deviation) for the ANN predictions of all 11 testing samples. Although the compounding effects of all three uncertainties have not been investigated in this thesis, analysis of individual aspects facilitates the assessment of their specific impacts on the ensuing prediction uncertainty. The values shown in the table confirm the previous observation that the largest uncertainty is attributed to the model parameter uncertainty. It is suspected that particular global optimization techniques, which are less prone to being stuck at local minima, might be more suitable for ANN training if the inverse problem is highly under-determined and the model parameter uncertainty is too high. Therefore, two other global optimization techniques, including PSO and GA, are tested for ANN learning, while all other conditions are kept identical to those of the previous case with LM-BP algorithm. The comparisons of model parameter uncertainties from these three training algorithms are summarized in **Table 3-4**. It is clear that substantial reduction in output uncertainties is resulted with the PSO algorithm (a relative improvement of approximately 40-50% as compared to the base case of LM-BP). A similar conclusion can be made regarding the GA algorithm, where uncertainties for 7 out of 11 predictions exhibit reduced variability. Model extrapolations can be another possible explanation for the model parameter uncertainty. Due to the limited number of records in the dataset and random allocation of records into training and testing sets, it is challenging to avoid or control extrapolations in input and output parameter spaces. In the present study, out of those 1000 realizations/iterations, a small number of negative COP values are predicted for some testing samples; those values are unphysical and must be discarded. Given

that uncertainty because of limited dataset size is also another important consideration, efforts to expand the dataset and obtain additional data records would be highly beneficial for future work.

3.4.2 Case 2 – Parameterization Using Principal Scores

In this case, the PCA technique is employed to reduce the dimensionality of input variable space in the original dataset from case study 1. The eigenvalues of the data covariance matrix are plotted in **Fig. 3-12**. It is deduced that the first six components possess the largest eigenvalues and contribute to most of the total variance; hence the number of principal components is chosen to be 6, while discarding the remaining ones. The final dimension of the principal scores is 6, which is less than the dimension of the original dataset. The principal scores and COP are considered as input and output attributes in subsequent ANN modeling. Applying the parameter exploration process by 5-fold cross-validation reveals the optimum network architectures (number of hidden neurons) are 7 and 4 × 11 for single and two hidden layers, respectively. The previous training, validation, and testing procedures are repeated in this case study, and the modeling results are presented in **Fig. 3-13**(**a-b**) and **Fig. 3-13**(**c-d**), respectively. Good agreements can be observed between the predicted COP and the target COP values during both training and testing stages for both network configurations. Similar to that of case 1, the values of R^2 for both ANN configurations are close to unity, while the values of *MSE* are small. Comparing the prediction results (**Figs. 3-7** and **3-13**) from both case studies, it is interesting to note that the ANN performance using the dimensionally-reduced principal scores to be at least as good as that using all 9 original variables. The reduced input attribute space helps to enhance the robustness of the data-driven ANN models for SAGD production analysis. Techniques such as cluster analysis can be employed in future studies to identify the internal structures among input variables and propose a reduced set of independent input attributes for ANN modeling (Amirian et al., 2014).

Despite of the various challenges associated with the dataset including limited availability, data redundancy, and the presence of noise, results in two case studies demonstrate a successful implementation of practical knowledge-based techniques to construct data-driven models capable of predicting SAGD recovery performance from log-derived and other operational variables.

3.5 Conclusion

A practical implementation of knowledge-based approaches has been proposed to facilitate SAGD production analysis. A comprehensive field dataset encompassing over 70 SAGD well pairs is compiled from numerous public domains. This field dataset is used to train a series of artificial neural networks to identify the non-linear relationships between various input (e.g., reservoir and operating parameters) and output variables (e.g., cumulative oil production). Principal component analysis is implemented to reduce the dimensionality of the original input variable space through orthogonal transformation. Uncertainty analysis is carried out to determine influences of the uncertainties originating from model parameter and data on the final ANN predictions.

Results from two case studies demonstrate that artificial neural network can be employed successfully to facilitate SAGD production performance prediction. Performances of ANN models are shown to be both reliable and satisfactory, as evidenced by the high values of R^2 and low values of *MSE* between predictions and targets. In fact, the performance of the ANN model with single hidden layer is comparable to the two-layer model. Comparison with modeling in original space confirms that the derived principal scores can reliably capture the essential information encompassed in the original data space. Results of the uncertainty assessment for this particular SAGD dataset reveal that model parameter uncertainty is dominating, while data uncertainty due to input attribute variability is the least significant.

This work has presented a number of important contributions. First, a novel workflow that incorporates exploratory data analysis, artificial neural networks, and comprehensive uncertainty assessment is presented. Its feasibility has been tested with an actual field dataset for analyzing SAGD production in heterogeneous reservoirs. The application of data-driven models involving actual field data from the McMurray bitumen deposits is novel. Second, uncertainty assessment is an area that is less explored in the intelligent and expert systems literature, particularly when reservoir-engineering data is involved. In this work, influences of the uncertainties originating from model parameter, data (uncertainty in input attributes due to imprecise analysis criteria and a limited number of records in the dataset) on the final ANN predictions are systematically quantified based on Monte Carlo and bootstrapping methods. The developed workflows can be extended to analyze other engineering datasets derived from experimental measurements. Finally,

this work demonstrates the feasibility of employing data-driven models for SAGD analysis using a realistic field dataset, a subject matter that insufficiently explored in the literature. Considering that many important data such as bottom-hole pressures, fluid properties, permeability, multi-phase flow functions, and thermal conductivities are commonly unavailable and, hence, are missing in the dataset, this work demonstrates how practical knowledge-based techniques can be used to construct data-driven models capable of predicting SAGD recovery performance from only log-derived and operational variables. The approach can be integrated directly into most existing reservoir management routines. Given that robust reservoir management and real-time decision-making are major challenges faced by the energy industry, the data-driven models presented here have great potential to be applied in other recovery projects such as solvent-aided steam injection.

Compared to detailed numerical simulations, data-driven models, as implemented in this study, offer a number of advantages. First, data-driven models can be readily updated, as new information (e.g., new wells or new data collection projects) becomes available. Second, they represent computationally-efficient alternatives for analyzing a large amount of competitor data, which is often prone to uncertainties and errors. They are particularly useful when the underlying relationships between input and output variables are highly complex and possibly uncertain.

There are still remaining limitations with the proposed approach. First, extracting variables from large datasets can be challenging. For example, as cumulative production time or a number of well pairs increase, manpower and computational costs associated with the data extraction procedure could also increase drastically. Second, a limited number of data records and model extrapolations may affect the model predictability and robustness. Finally, the set of input attributes employed in this study might not sufficiently capture the influences of all reservoir, fluid, and operating conditions, particularly in reservoirs with significant lateral heterogeneities or substantial variation in bottom-hole pressures and fluid properties.

Therefore, future studies should include data from other producing fields to increase the number of samples and incorporate other data mining techniques to identify internal patterns among data. Avenues for expanding the dataset and integrating additional records should be explored and would help to avoid instances of extrapolation and reduce prediction uncertainty stemming from various sources. The possibilities of incorporating other variables extracted from core measurements and fluid analysis will be explored. Formulation of other input and output

attributes should be considered. Steam-oil-ration profile, which is an important economic indicator, can be treated as additional output. Additional research should also focus on the parameterization of lithological facies as relevant input attributes. Future research will also analyze the compounding effects of all aspects of uncertainty.

3.6 Reference

Akin, S., & Bagci, S. (2001). A laboratory study of single-well steam-assisted gravity drainage process. *Journal of petroleum science and engineering, 32*(2), 23-33.

Al-Bulushi, N., King, P., Blunt, M., & Kraaijveld, M. (2012). Artificial neural networks workflow and its application in the petroleum industry. *Neural Computing and Applications, 21*(3), 409-421.

Al-Fattah, S. M., & Startzman, R. A. (2003). Neural network approach predicts U.S. natural gas production. *SPE Production & Facilities, 18*(02), 84-91.

Aly, M. A. E. E. (2007). Data analysis methodology for reservoir management. Paper presented at the *EUROPEC/EAGE Conference and Exhibition,* London, U.K.

Amirian, E., Leung, J. Y., Zanon, S., & Dzurman, P. (2013). Data-driven modeling approach for recovery performance prediction in SAGD operations. Paper presented at the *SPE Heavy Oil Conference-Canada,* Calgary, Alberta, Canada.

Amirian, E., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Integrated cluster analysis and artificial neural network modeling for steam-assisted gravity drainage performance prediction in heterogeneous reservoirs. *Expert Systems with Applications, 42*(2), 723-740.

Ampazis, N., & Perantonis, S. J. (2000). Levenberg-marquardt algorithm with adaptive momentum for the efficient training of feedforward networks. Paper presented at the *IEEE-INNS-ENNS International Joint Conference on Neural Networks,* Como, Italy, 126-131,

An, P., & Moon, W. (1993). Reservoir characterization using feedforward neural networks. *In SEG Technical Program Expanded Abstracts 1993*, 258-262.

Bagci, A. S. (2006). Experimental and simulation studies of SAGD process in fractured reservoirs. Paper presented at the *SPE/DOE Symposium on Improved Oil Recovery,* Tulsa, Oklahoma, U.S.A.

Bhattacharya, S., & Nikolaou, M. (2013). Analysis of production history for unconventional gas reservoirs with statistical methods. *SPE Journal, 18*(05), 878-896.

Chang, J., Ivory, J., & Tunney, C. (2012). Numerical simulation of steam-assisted gravity drainage with vertical slimholes. *SPE Reservoir Evaluation & Engineering, 15*(06), 662-675.

Chen, Q., Gerritsen, M. G., & Kovscek, A. R. (2008). Effects of reservoir heterogeneities on the steam-assisted gravity-drainage process. *SPE Reservoir Evaluation & Engineering, 11*(05), 921-932.

Chow, L., & Butler, R. (1996). Numerical simulation of the steam-assisted gravity drainage process (SAGD). *Journal of Canadian Petroleum Technology, 35*(06), 55-62.

Dadashpour, M., Rwechungura, R. W., & Kleppe, J. (2011). Fast reservoir parameter estimation by using effect of principal components sensitivities and discrete cosine transform. Paper presented at the *SPE Reservoir Simulation Symposium,* The Woodlands, Texas, USA.

Demuth, H., Beale, M., & Hagan, M. (2008). *Neural network toolbox™ 6*, *User's Guide.*

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Egermann, P., Renard, G., & Delamaide, E. (2001). SAGD performance optimization through numerical simulations: Methodology and field case example. Paper presented at the *SPE International Thermal Operations and Heavy Oil Symposium,* Margarita Island, Venezuela.

Fatemi, S. M. (2009). Simulation study of steam assisted gravity drainage (SAGD) in fractured systems. *Oil & Gas Science and Technology-Revue De l'IFP, 64*(4), 477-487.

Fedutenko, E., Yang, C., Card, C., & Nghiem, L. X. (2014). Time-dependent neural network based proxy modeling of SAGD process. Paper presented at the *SPE Heavy Oil Conference-Canada,* Calgary, Alberta, Canada.

Francis, L. (2001). The basics of neural networks demystified. *Contingencies (11/12 2001), 1*(1), 56-61.

Ge, J., Xia, Y., & Nadungodage, C. (2010). UNN: A neural network for uncertain data classification. *Advances in Knowledge Discovery and Data Mining* (449-460), Springer.

Gharbi, R. B., & Elsharkawy, A. M. (1999). Neural network model for estimating the PVT properties of middle east crude oils. *SPE Reservoir Evaluation & Engineering, 2*(03), 255-265.

Guan, B. T., Gertner, G. Z., & Parysow, P. (1997). A framework for uncertainty assessment of mechanistic forest growth models: A neural network example. *Ecological Modelling, 98*(1), 47-58.

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the marquardt algorithm. *IEEE transactions on Neural Networks, 5*(6), 989-993.

Han, D., Kwong, T., & Li, S. (2007). Uncertainties in real-time flood forecasting with neural networks. *Hydrological Processes, 21*(2), 223-228.

Hanna, S. R., Chang, J. C., & Fernau, M. E. (1998). Monte carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmospheric Environment, 32*(21), 3619-3628.

Hasani, M., & Emami, F. (2008). Evaluation of feed-forward back propagation and radial basis function neural networks in simultaneous kinetic spectrophotometric determination of nitroaniline isomers. *Talanta, 75*(1), 116-126.

Haykin, S.S. (2008). *Neural networks and learning machines* (3rd ed.), Upper Saddle River, NJ, USA: Pearson.

Heaton, J. (2008). *Introduction to neural networks with java*, Heaton Research, Inc.

Holdaway, K. R. (2009). Exploratory data analysis in reservoir characterization projects. Paper presented at the *SPE/EAGE Reservoir Characterization & Simulation Conference,* Abu Dhabi, UAE.

Jansen, F., & Kelkar, M. (1996). Exploratory data analysis of production data. Paper presented at the *Permian Basin Oil & Gas Recovery Conference,* 331-342, Midland, Texas, USA.

Jolliffe, I. (2005). *Principal component analysis*, Wiley Online Library.

Joo, S., Oh, S. E., Sim, T., Kim, H., Choi, C. H., Koo, H., & Mun, J. H. (2014). Prediction of gait speed from plantar pressure using artificial neural networks. *Expert Systems with Applications, 41*(16), 7398-7405.

Khosravi, A., Nahavandi, S., & Creighton, D. (2013). Quantifying uncertainties of neural network-based electricity price forecasts. *Applied Energy, 112*, 120-129.

Kjræulff, U. B., & Madsen, A. L. (2008). *Bayesian networks and influence diagrams: A guide to construction and analysis*, Springer Science+Business Media.

Kocadağlı, O., & Aşıkgil, B. (2014). Nonlinear time series forecasting with bayesian neural networks. *Expert Systems with Applications, 41*(15), 6596-6610.

Kröse, B., Krose, B., van der Smagt, P., & Smagt, P. (1993). *An introduction to neural networks*. The Netherlands: University of Amsterdam.

Lechner, J. P., & Zangl, G. (2005). Treating uncertainties in reservoir performance prediction with neural networks. Paper presented at the *SPE Europec/EAGE Annual Conference,* Madrid, Spain.

Lee, S. H., Kharghoria, A., & Datta-Gupta, A. (2002). Electrofacies characterization and permeability predictions in complex reservoirs. *SPE Reservoir Evaluation & Engineering, 5*(03), 237-248.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

Mezić, I., & Runolfsson, T. (2008). Uncertainty propagation in dynamical systems. *Automatica, 44*(12), 3003-3013.

Mohammadpoor, M., Firouz, Q., Reza, A., & Torabi, F. (2012). Implementing simulation and artificial intelligence tools to optimize the performance of the CO2 sequestration in coalbed methane reservoirs. Paper presented at the *Carbon Management Technology Conference,* Orlando, Florida, USA.

Nigrin, A. (1993). *Neural networks for pattern recognition*. MIT press.

Norman, C. D. (2013). Correlation of porosity uncertainty to productive reservoir volume. Paper presented at the *SPE Middle East Oil and Gas Show and Conference,* Manama, Bahrain.

Papadopoulos, C. E., & Yeung, H. (2001). Uncertainty estimation and Monte Carlo simulation method. *Flow Measurement and Instrumentation, 12*(4), 291-298.

Parada, C. H., & Ertekin, T. (2012). A new screening tool for improved oil recovery methods using artificial neural networks. Paper presented as the *SPE Western Regional Meeting*, Bakersfield, California, USA.

Petrović, D. V., Tanasijević, M., Milić, V., Lilić, N., Stojadinović, S., & Svrkota, I. (2014). Risk assessment model of mining equipment failure based on fuzzy logic. *Expert Systems with Applications, 41*(18), 8157-8164.

Pooladi-Darvish, M., & Mattar, L. (2002). SAGD operations in the presence of overlying gas cap and water layer-effect of shale layers. *Journal of Canadian Petroleum Technology, 41*(06)

Popa, A. S., & Patel, A. N. (2012). Neural networks for production curve pattern recognition applied to cyclic steam optimization in diatomite reservoirs. Paper presented at the *SPE Western Regional Meeting,* Bakersfield, California, USA.

Popa, A. S., Cassidy, S. D., & Mercer, M. (2011). A data mining approach to unlock potential from an old heavy oil field. Paper presented at the *SPE Western North American Region Meeting,* Anchorage, Alaska, USA.

Queipo, N. V., Goicochea, J. V., & Pintos, S. (2002). Surrogate modeling-based optimization of SAGD processes. *Journal of Petroleum Science and Engineering, 35*(1), 83-93.

Ramgulam, A. (2006). Utilization of artificial neural networks in the optimization of history matching (Doctoral dissertation, the Pennsylvania State University).

Refsgaard, J. C., van der Sluijs, Jeroen P, Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process–a framework and guidance. *Environmental Modelling & Software, 22*(11), 1543-1556.

Sarma, P., Durlofsky, L. J., Aziz, K., & Chen, W. H. (2007). A new approach to automatic history matching using kernel PCA. Paper presented at the *SPE Reservoir Simulation Symposium,* Houston, Texas, USA.

Scheevel, J., & Payrazyan, K. (2001). Principal component analysis applied to 3D seismic data for reservoir property estimation. *SPE Reservoir Evaluation & Engineering, 4*(01), 64-72.

Secchi, P., Zio, E., & Di Maio, F. (2008). Quantifying uncertainties in the estimation of safety parameters by using bootstrapped artificial neural networks. *Annals of Nuclear Energy, 35*(12), 2338-2350.

Shin, H., & Polikar, M. (2006). Experimental investigation of the fast-SAGD process. Paper presented at the *Canadian International Petroleum Conference,* Calgary, Alberta, Canada.

Siu, A., Nghiem, L., Gittins, S., Nzekwu, B., & Redford, D. (1991). *Modelling steam-assisted gravity drainage process in the UTF pilot project*. Paper presented at the *SPE Annual Technical Conference and Exhibition,* Dallas, Texas, USA.

Smith, L. I. (2002). A tutorial on principal components analysis. Cornell University, USA, 51(52), 65.

Solomatine, D., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics, 10*(1), 3-22.

Srivastav, R., Sudheer, K., & Chaubey, I. (2007). A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. *Water Resources Research, 43*(10).

Stundner, M., & Al-Thuwaini, J. S. (2001). How data-driven modeling methods like neural networks can help to integrate different types of data into reservoir management. Paper presented as the *SPE Middle East Oil Show*, Manama, Bahrain.

Tan, S. S., & Smeins, F. E. (1996). Predicting grassland community changes with an artificial neural network model. *Ecological Modelling, 84*(1-3), 91-97.

Tang, H., Meddaugh, W. S., & Toomey, N. (2011). Using an artificial-neural-network method to predict carbonate well log facies successfully. *SPE Reservoir Evaluation & Engineering, 14*(01), 35-44.

Tiwari, M. K., & Chatterjee, C. (2010). Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs). *Journal of Hydrology, 382*(1), 20-33.

Tukey, J. W. (1977). Exploratory data analysis, Addison-Wesley.

Verga, F., Viberti, D., & Gonfalini, M. (2002). Uncertainty evaluation in well logging: Analytical or numerical approach? Paper presented at the *SPWLA 43rd Annual Logging Symposium,* Oiso, Japan.

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, Jeroen P, van Asselt, M. B., Janssen, P., & Krayer von Krauss, Martin P. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment, 4*(1), 5-17.

Wang, C., & Leung, J. (2015). Characterizing the effects of lean zones and shale distribution in steam-assisted-gravity-drainage recovery performance. *SPE Reservoir Evaluation & Engineering, 18*(03), 329-345.

Waszczyszyn, Z. (1999*). Neural networks in the analysis and design of structures*, Springer.

Wright, W. (1999). Bayesian approach to neural-network modeling with input uncertainty. *Neural Networks, IEEE Transactions On, 10*(6), 1261-1270.

Xu, S., & Chen, L. (2008). A novel approach for determining the optimal number of hidden layer neurons for FNN,Äôs and its application in data mining. Paper presented at the *International Conference on Information Technology and Applications: iCITA,* 683-686.

Yadav, S. (2006). History matching using face-recognition technique based on principal component analysis. Paper presented at the *SPE Annual Technical Conference and Exhibition,* San Antonio, Texas, USA.

Yang, G., & Butler, R. (1992). Effects of reservoir heterogeneities on heavy oil recovery by steam-assisted gravity drainage. *Journal of Canadian Petroleum Technology, 31*(08), 37-43.

Yeten, B., Durlofsky, L. J., & Aziz, K. (2002). Optimization of nonconventional well type location and trajectory. Paper presented at the *SPE Annual Technical Conference and Exhibition,* San Antonio, Texas, USA.

Zerafat, M. M., Ayatollahi, S., Mehranbod, N., & Barzegari, D. (2011). *Bayesian network analysis as a tool for efficient EOR screening*. Paper presented as the *SPE Enhanced Oil Recovery Conference*, Kuala Lumpur, Malaysia.

Zhang, X., Liang, F., Yu, B., & Zong, Z. (2011). Explicitly integrating parameter, input, and structure uncertainties into bayesian neural networks for probabilistic hydrologic forecasting. *Journal of Hydrology, 409*(3), 696-709.

Tables

Table 3-1 Statistical properties of input and output variables of the original dataset.

Statistical Properties	Input Variables									Output Variable
	ϕ	S_w	N/G	h	SI	N_i^{prod}	N_i^{inj}	T_{total}	CSI	COP
Average	0.28	0.34	0.62	55.09	1.97	1.04	1.05	4.19	516474.71	230131.21
Standard deviation	0.04	0.18	0.19	17.14	6.19	0.11	0.06	3.02	320640.57	153847.99
Minimum	0.21	0.13	0.12	12.60	0.00	0.99	0.96	0.42	58845.00	26523.00
Maximum	0.35	0.73	1.00	93.70	50.00	1.51	1.20	11.66	1270096.00	643304.00

Table 3-2 Uncertainties in input data of two data record example.

Sample	Data type	ϕ	S_w	N/G	h	SI	N_i^{prod}	N_i^{inj}	T_{total}	CSI
Data record example 1	Actual value from data analysis	0.31	0.26	0.46	44.33	0.52	1.00	1.02	9.13	818782.00
	Lower bound (minus 10%)	0.28	0.24	0.41	39.90	0.47	0.90	0.92	8.22	736903.80
	Upper bound (plus 10%)	0.34	0.29	0.50	48.77	0.58	1.10	1.12	10.05	900660.20
Data record example 2	Actual value	0.29	0.35	0.69	55.58	0.27	1.43	1.02	10.74	838972.00
	Lower bound (minus 10%)	0.26	0.31	0.62	50.02	0.24	1.29	0.92	9.66	755074.80
	Upper bound (plus 10%)	0.32	0.38	0.76	61.13	0.29	1.58	1.13	11.81	922869.20

Table 3-3 Target values for 11 test samples and the corresponding uncertainties represented as standard deviations.

Sample Index	1	2	3	4	5	6	7	8	8	10	11
Target value	40404	86482	108436	116675	146340	183603	237045	315963	382430	426597	640777
Uncertainty in model parameter	76617	33573	54591	67661	68826	69474	55984	67986	66470	125812	215025
Uncertainty in input attributes due to imprecise analysis criteria	29384	18106	22003	26000	33145	31806	32610	31411	33304	44080	51655
Uncertainty due to limited records in original dataset	43939	23640	37568	45679	42604	41888	39936	43065	40490	60421	71421

Table 3-4 Model parameter uncertainties represented as standard deviations for 11 test samples from three algorithms (LM-BP, PSO, and GA) and the relative improvement of PSO and GA as compared against LM-BP.

Sample Index	1	2	3	4	5	6	7	8	8	10	11
LM-BP	76617	33573	54591	67661	68826	69474	55984	67986	66470	125812	215025
PSO	41986	16078	25068	29904	42439	34497	32717	35140	30010	55870	93393
GA	56336	25916	39151	79752	71124	63109	44775	65554	57662	142551	229791
Relative improvement between LM-BP and PSO	45.2%	52.1%	54.1%	55.8%	38.3%	50.3%	41.6%	48.3%	54.9%	55.6%	56.6%
Relative improvement between LM-BP and GA	26.5%	22.8%	28.3%	-17.9%	-3.3%	9.2%	20.0%	3.6%	13.3%	-13.3%	-6.9%

Figures

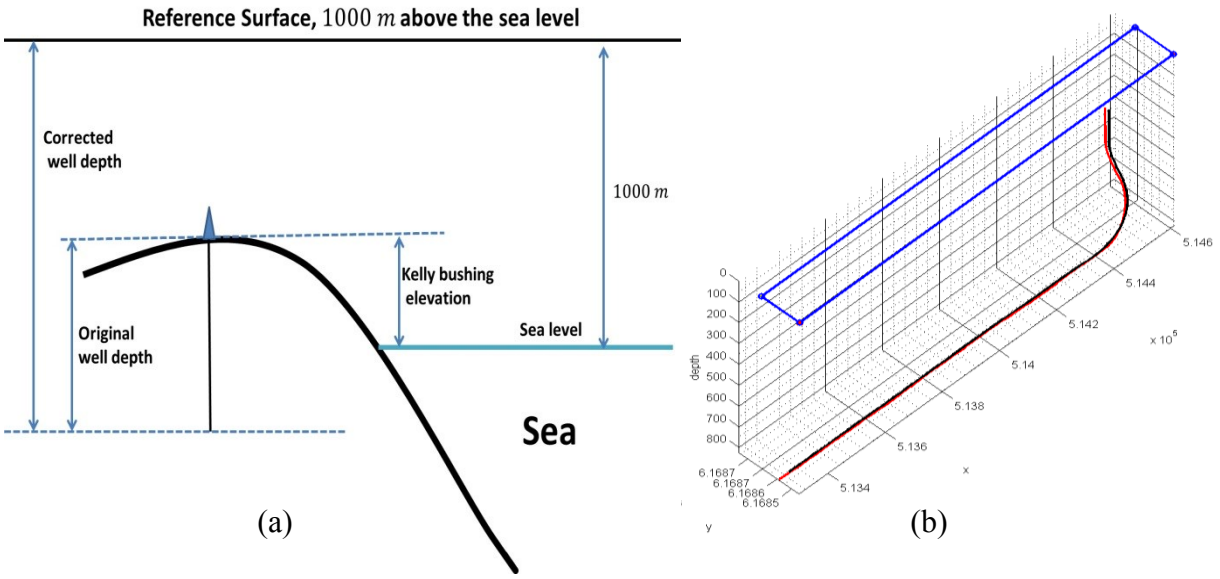


Fig. 3-1 Pre-process of logging data for data analysis: (a) – standardization of depth data according to an arbitrary reference surface; (b) – rectangular search domain around an injector-producer well pair in 3D view.

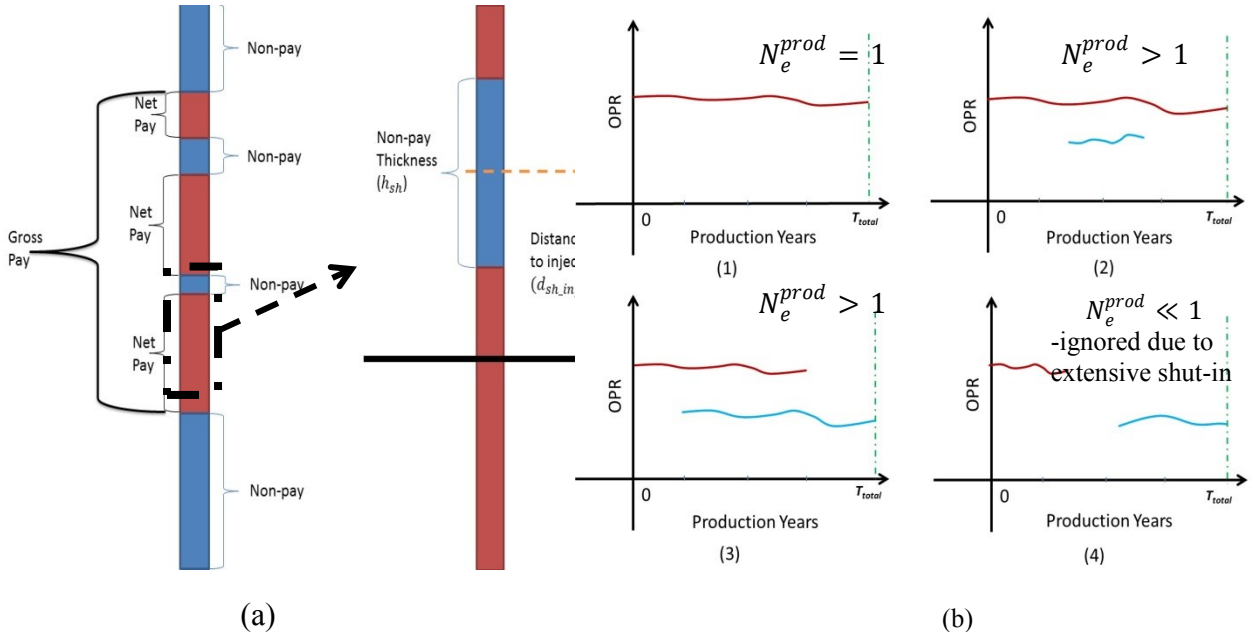
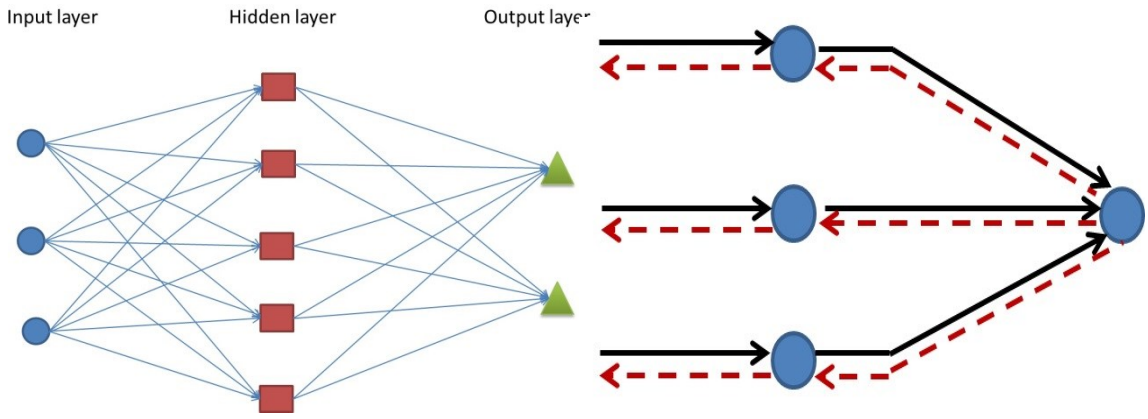


Fig. 3-2 Data analysis example: (a) – pay definition from log analysis (net pay zone represents oil-filled sand, while non-pay zones include shale or water-filled sand) and formulation of shale

index; (b) – production data analysis for calculating N_e^{prod}: the X axis denotes the production time while the Y axis is the oil production rate (OPR); the red curve represents OPR oil production rate of the primary producer, while the blue curve represents the oil production rate of the secondary producer.



(a) (b)

Fig. 3-3 Neural network structure: (a) – schematic of neural network architecture; (b) – transmission of signals; black solid arrows = function signals; red dashed arrows = error signals.

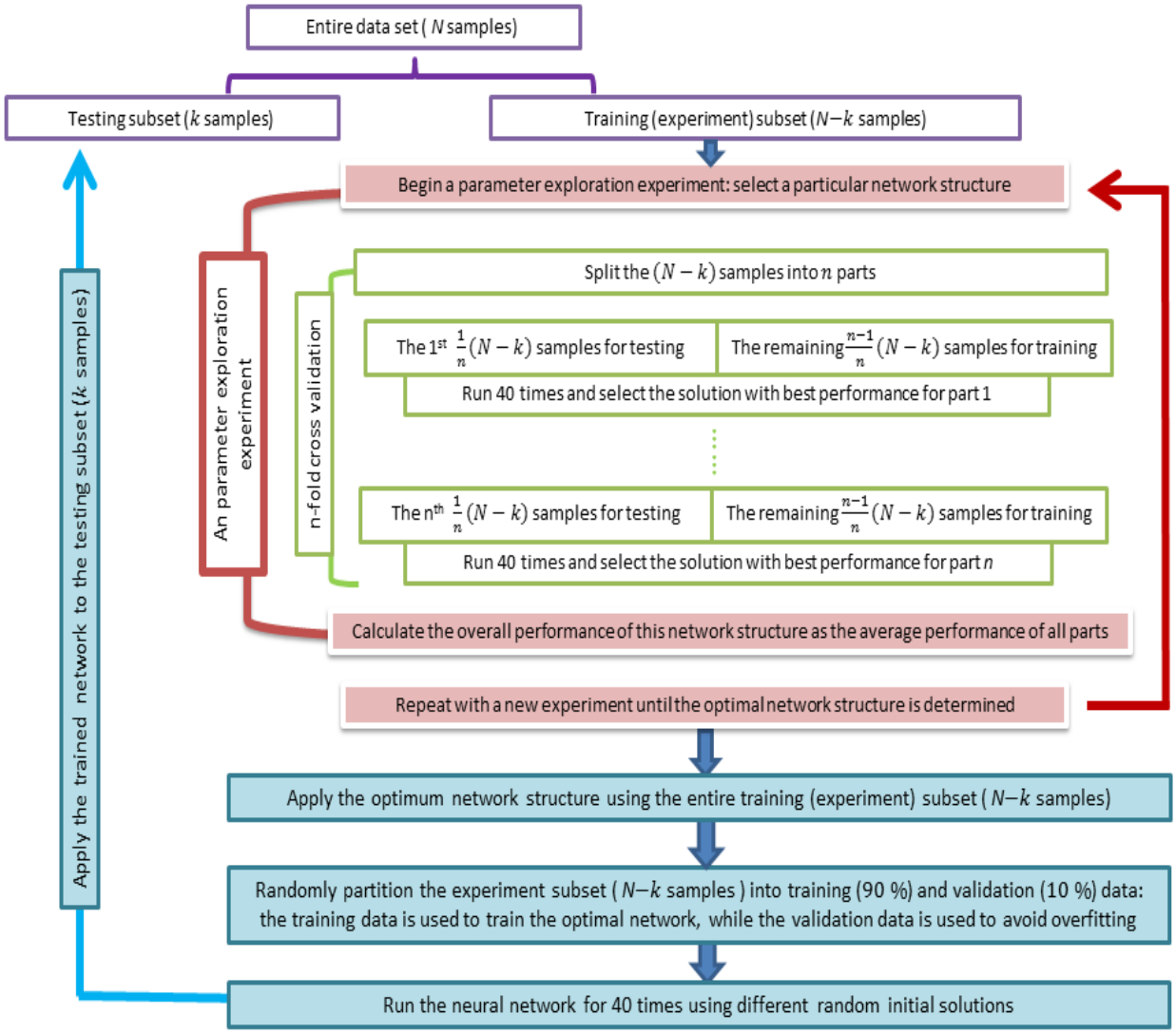


Fig. 3-4 Flowchart of ANN modeling.

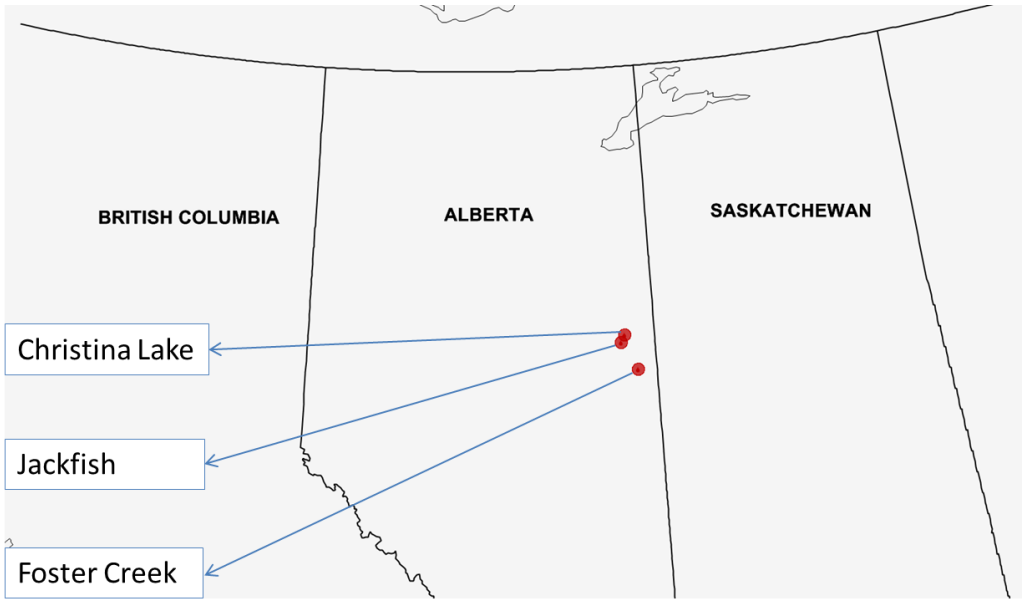
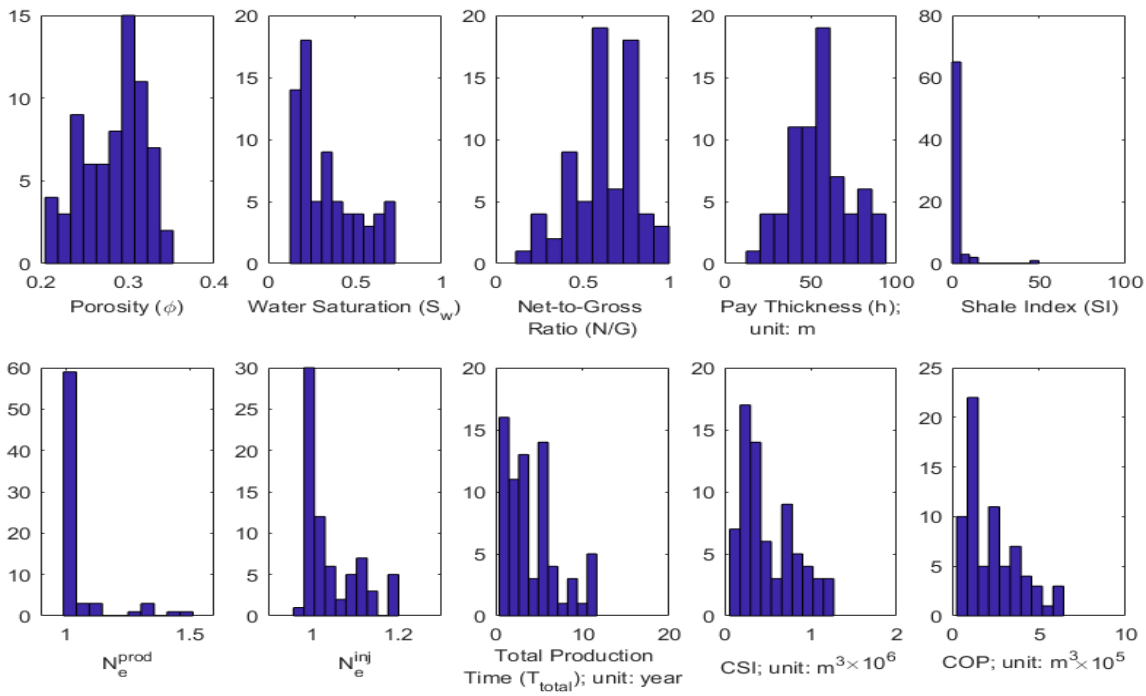
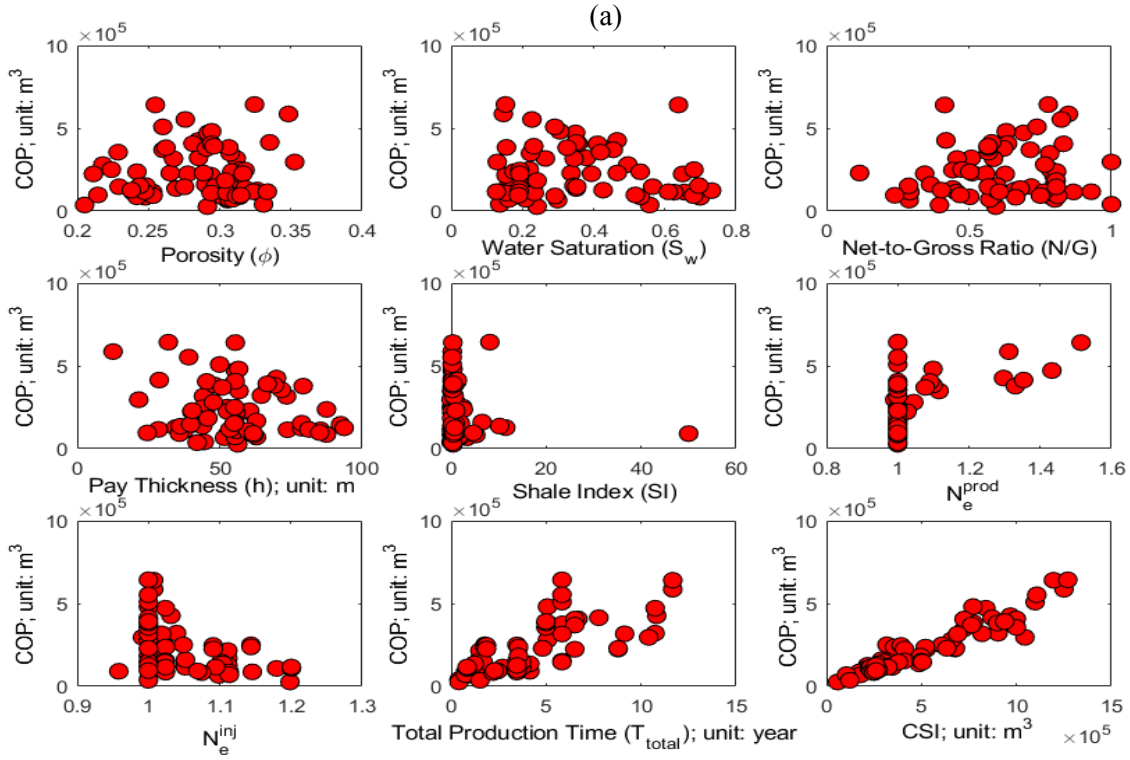


Fig. 3-5 Locations of the 3 SAGD producing fields.



(a)



(b)

Fig. 3-6 Exploratory data analysis of the dataset in case study 1: (a) – histograms of all input and output attributes; (b) – cross-plots between input attributes and output attribute (COP).

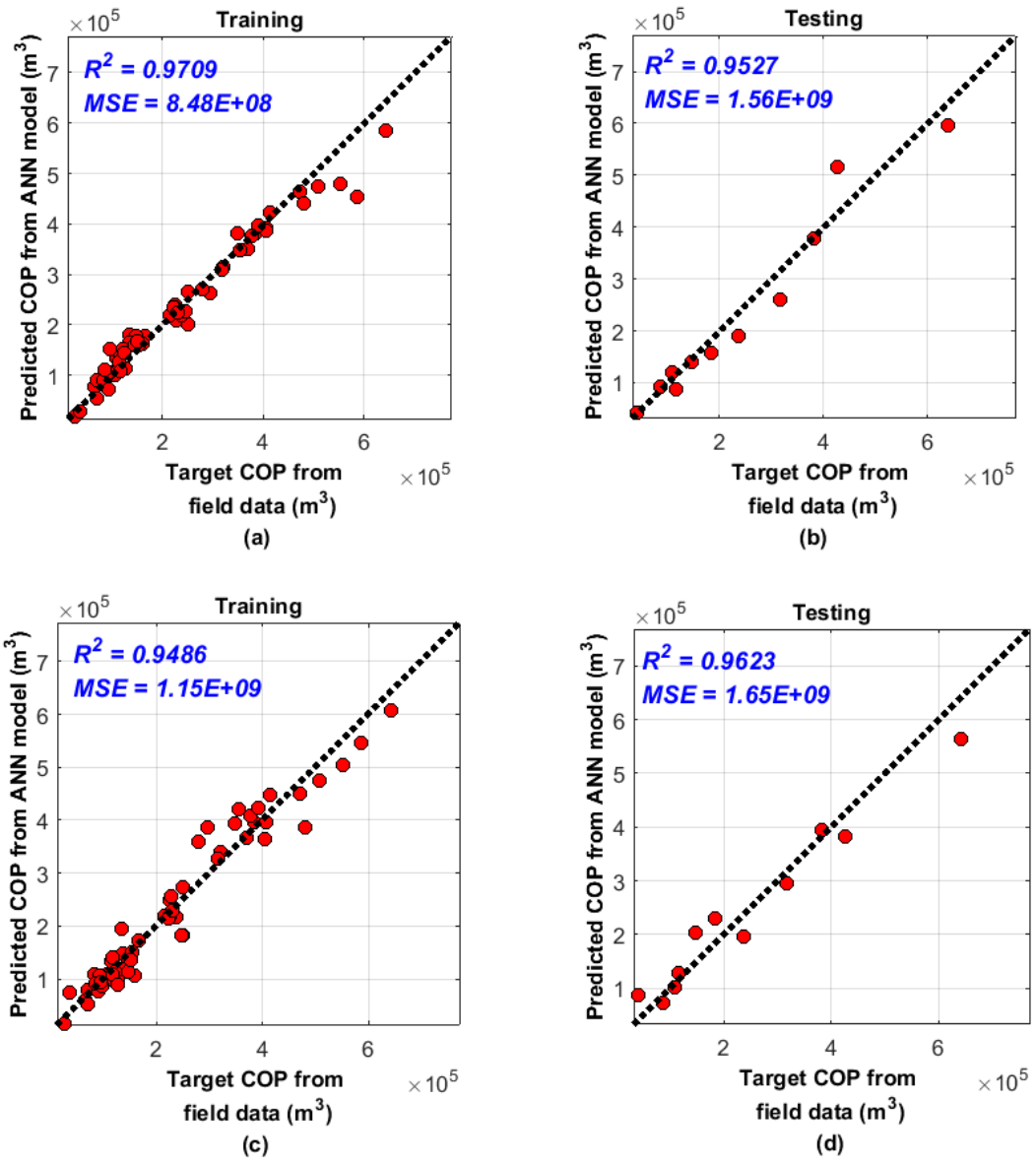


Fig. 3-7 Cross-plots of COP from ANN prediction (with 9 input attributes) and target COP from field data of case study 1. Single hidden layer ANN: (a) – training dataset, (b) – testing dataset; Two hidden layers ANN: (c) – training dataset, (d) – testing dataset.

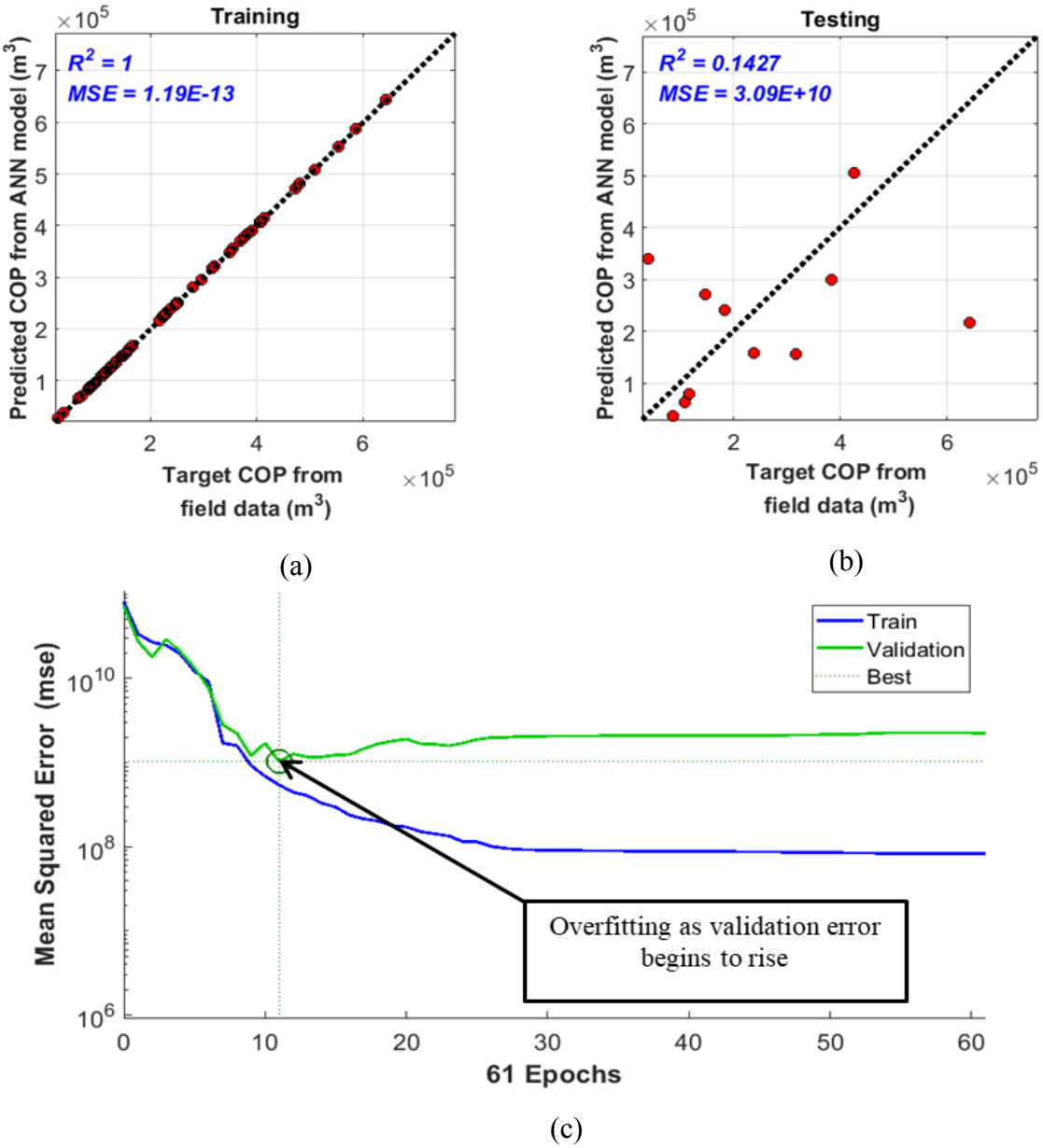


Fig. 3-8 Illustration of overfitting for the single hidden layer in Case Study: (a) – training dataset; (b) – testing dataset; (c) – evolution of training and validation error as a function of epochs.

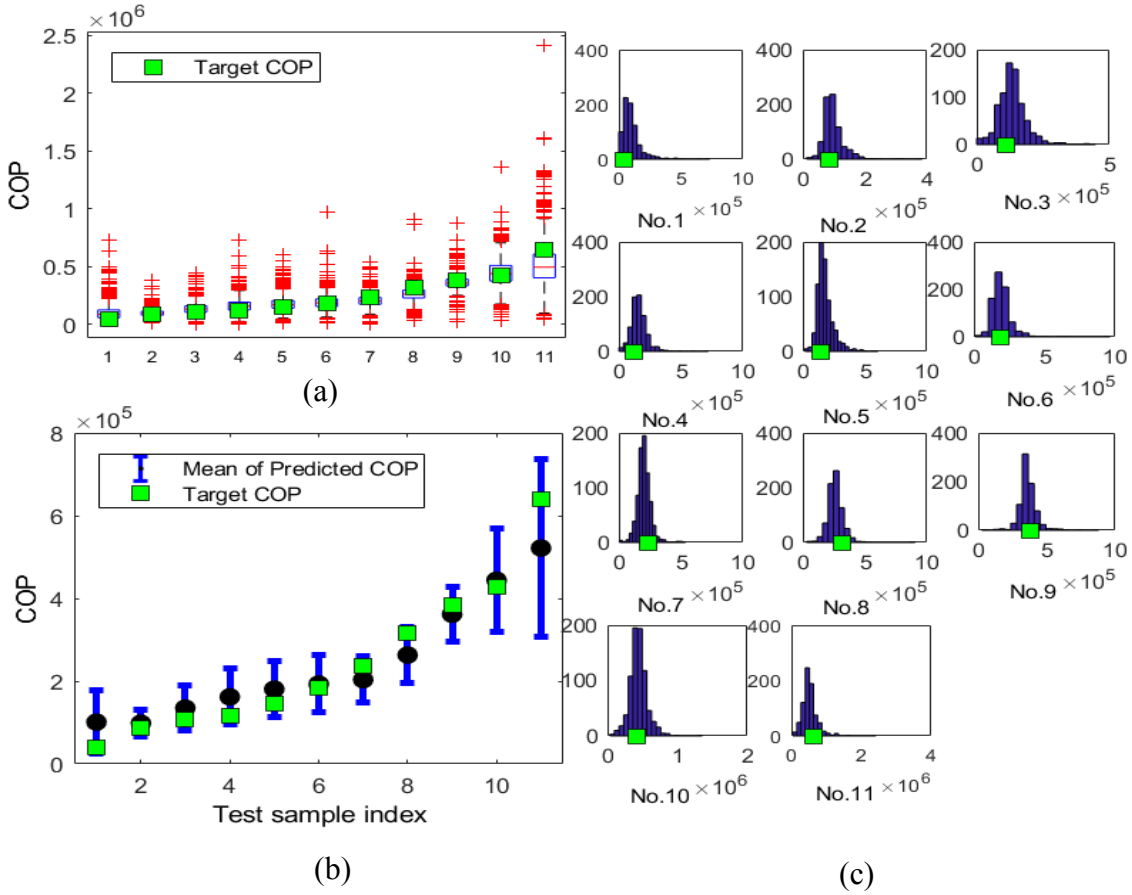


Fig. 3-9 Prediction outcome uncertainty due to model parameter uncertainty: (a) – box plot, (b) – errorbar plot, and (c) – histograms of predicted COP for the 11 testing samples; the corresponding target COP is represented by the green square.

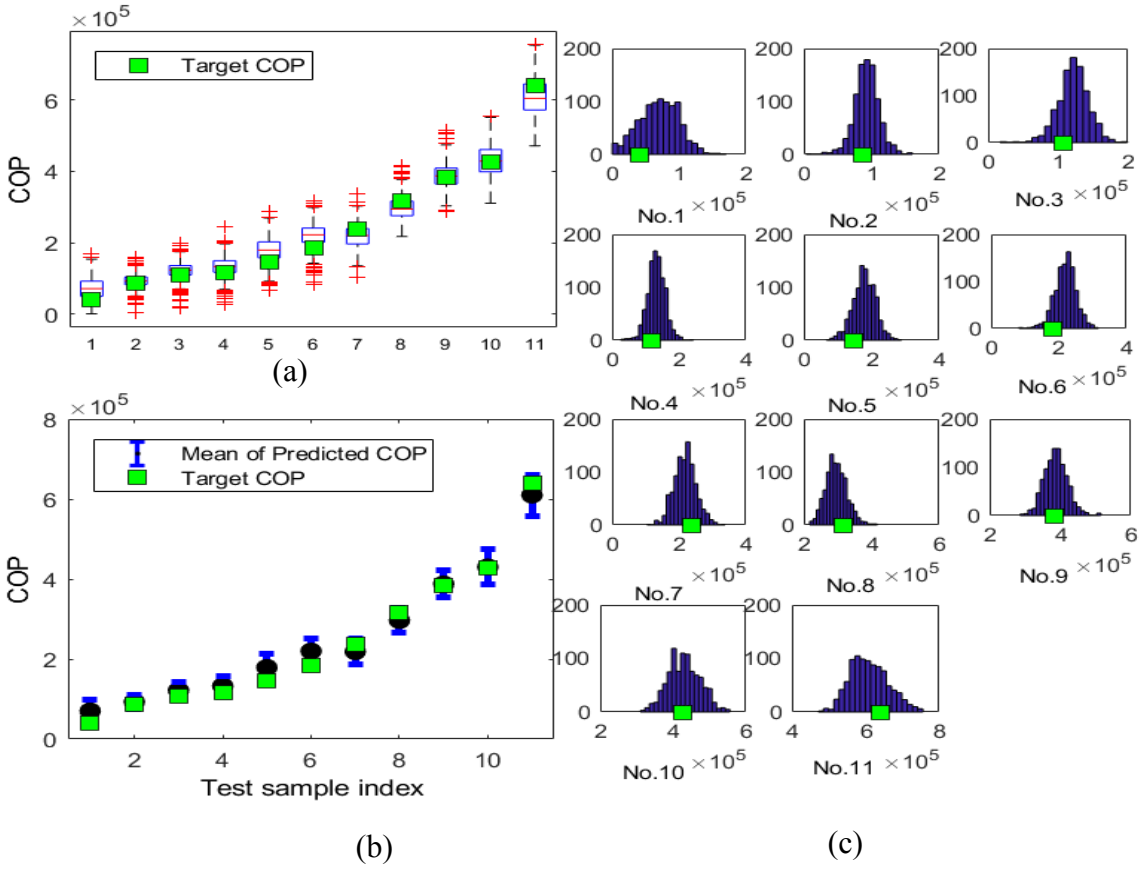


Fig. 3-10 Prediction outcome uncertainty due to data uncertainty (uncertainty in input attributes as a result of imprecise analysis criteria): (a) – box plot, (b) – errorbar plot, and (c) – histograms of predicted COP for the 11 testing samples; the corresponding target represented by the green square.

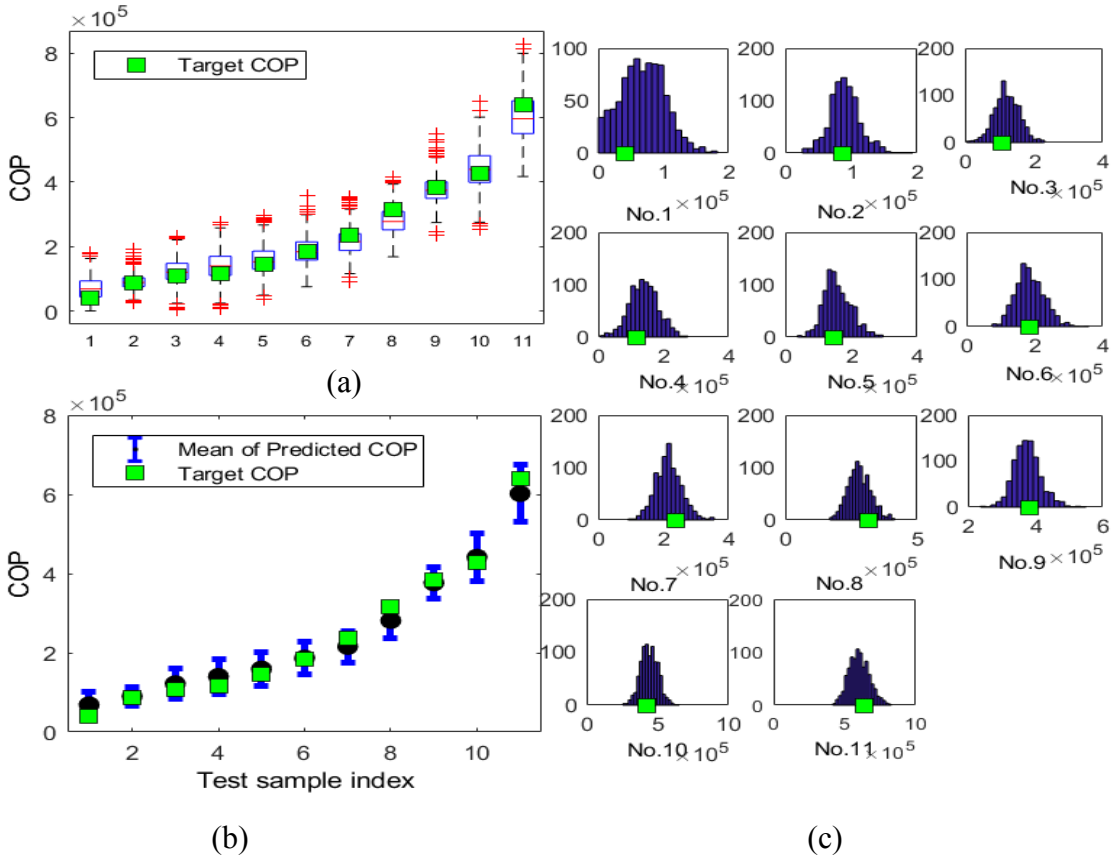


Fig. 3-11 Prediction outcome uncertainty due to data uncertainty (uncertainty as a result of limited number of records in the dataset): (a) – box plot, (b) – errorbar plot, and (c) – histograms of predicted COP for the 11 testing samples; the corresponding target COP is represented by the green square.

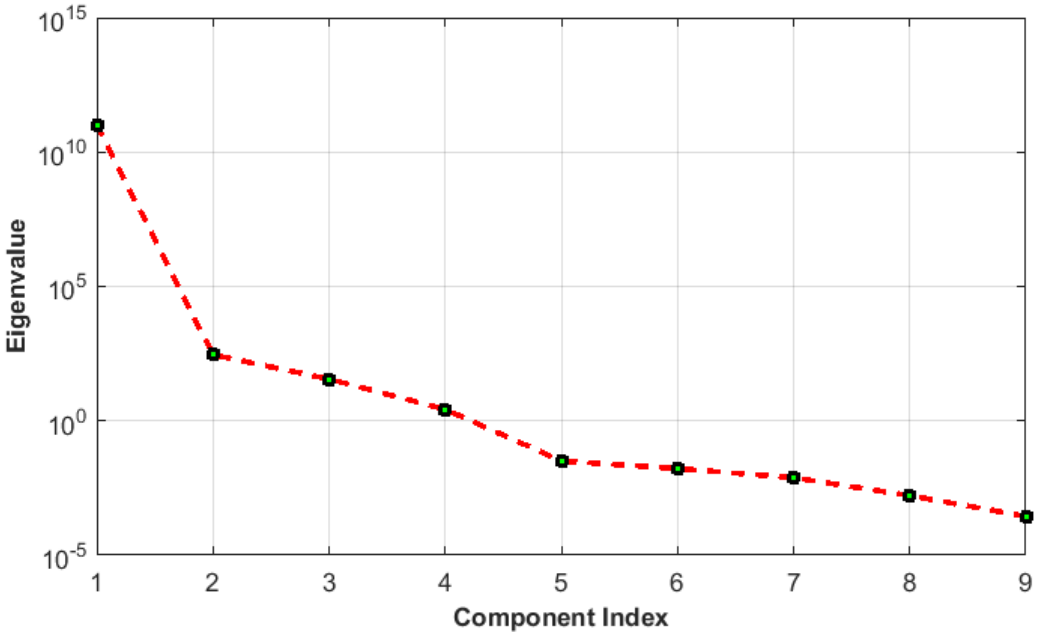


Fig. 3-12 Eigenvalue and its corresponding index in case study 2.

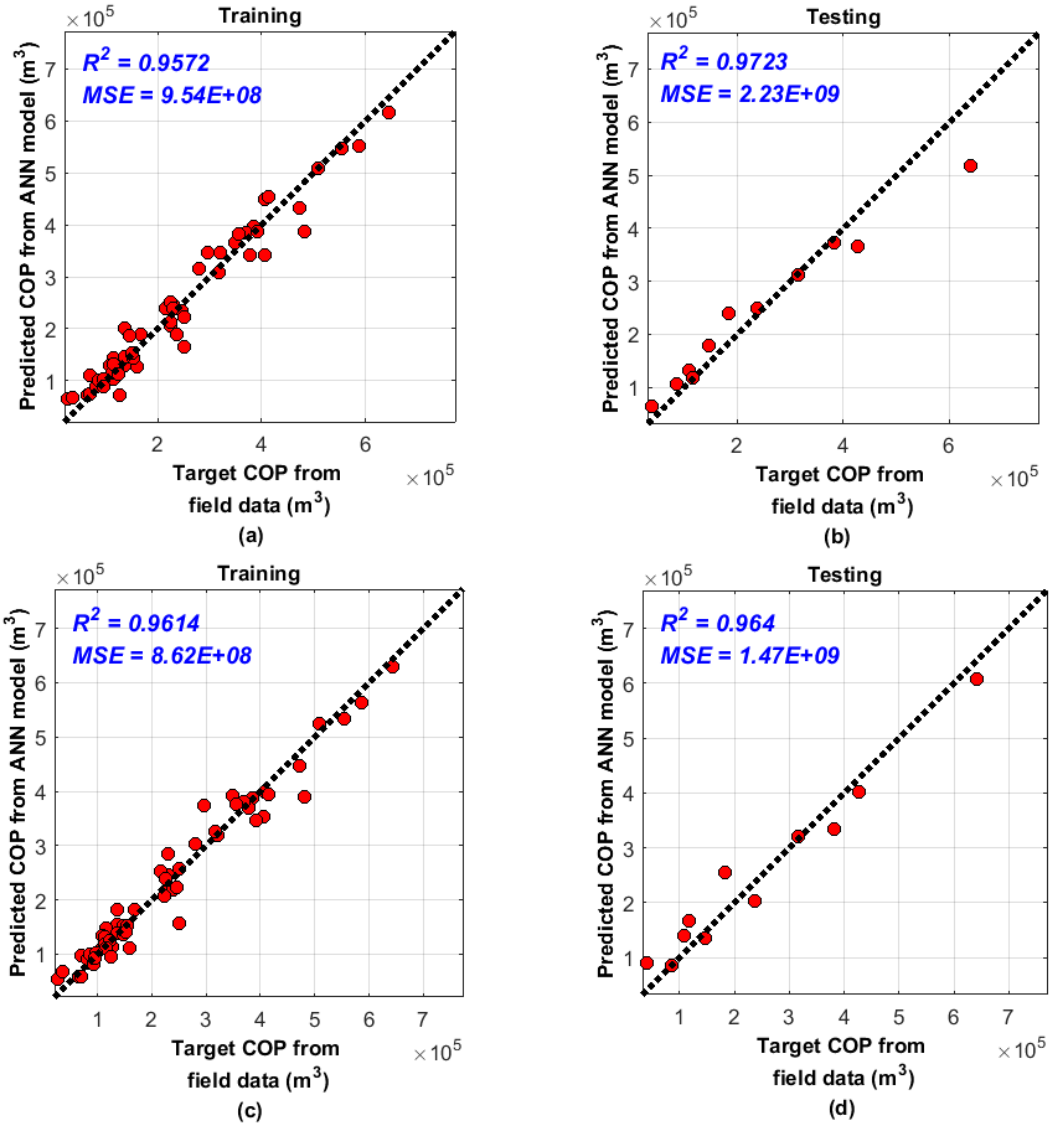


Fig. 3-13 Cross-plots of COP from ANN prediction (with 6 input attributes) and target COP from field data of case study 2. Single hidden layer ANN: (a) – training dataset, (b) – testing dataset; Two hidden layers ANN: (c) – training dataset, (d) – testing dataset.

Chapter 4 Application of the SAGD Production Analysis Workflow for a Large Dataset[2]

Chapter Overview

Data-driven modeling approaches are employed as complementary tools for SAGD production forecast and pattern recognition of highly non-linear relationships between system variables in chapter 3. The developed SAGD production analysis workflow in preceding chapter is extended and applied to a large dataset in the current chapter.

Several improvements are implemented and summarized as: first, different from the previous chapter where only a limited number of wells are analyzed, field data from more than two thousand wells are extracted from various publicly available sources to formulate 153 complete data samples in the current chapter. Analysis of a raw dataset of this magnitude for SAGD reservoirs has not been published in the literature. This chapter attempts to discuss and address a number of the challenges encountered. Second, the impact of extrapolation of the petrophysical parameters from the nearby vertical well is assessed. As a result, an additional input attribute is introduced to capture the uncertainty in extrapolation, while a new output attribute is incorporated as a quantitative measure of SAGD process efficiency. Third, k-means clustering analysis algorithm is applied to improve prediction quality and model robustness by removing data correlation and identifying internal structures among the dataset, which is a novel extension to the previous SAGD analysis study.

[2] A version of this chapter has been published as:
Ma, Z., Leung, J. Y., & Zanon, S. (2017). Practical data mining and artificial neural network modeling for SAGD production analysis, *Journal of Energy Resources Technology*, *139*(3), 032909.

ANN is employed to facilitate the production performance analysis by calibrating the reservoir heterogeneities and operating constraints with production performance. The modeling results are demonstrated to be both reliable and acceptable. This chapter demonstrates the combination of artificial intelligence-based approaches and data-mining analysis can facilitate practical field data analysis, which is often prone to uncertainties, errors, biases, and noises, with high reliability and feasibility.

4.1 Introduction

Steam Assisted Gravity Drainage (SAGD) is one of the most important and proven thermal enhanced oil recovery techniques for heavy oil (or bitumen) production (Butler et al., 1981). As shown in **Fig. 4-1**(**a**), a pair of horizontal wells is drilled into the target formation: a steam injection well that is located at a few meters above the oil production well. High-temperature steam is continuously injected into the reservoir to form a steam chamber, which facilitates heating the crude oil and reducing its viscosity. The heated oil would drain along the edge of the steam chamber into the production well by gravity.

Although experimental studies (Bagci, 2006; Sasaki et al., 1996) and numerical simulations (Fatemi, 2009; Shin et al., 20012; Panwar et al., 2015) have been adopted extensively to understand the SAGD process, there are still some limitations associated with these methods for the purposes of field-scale recovery performance evaluation and prediction. It is difficult to reproduce reservoir conditions and heterogeneities (variation in rock properties) in laboratory-scale models. On the other hand, the numerical simulation may invoke various assumptions regarding the physical system to be modeled, due to the imprecise or incomplete knowledge of the underlying physical processes and reservoir description; for example, precise quantification of fluid behavior, recovery mechanisms, and multi-scale heterogeneities can be challenging. The steps for model construction may also be time-consuming (both computationally- and labor-intensive) (Queipo et al., 2002; Lacroix et al., 2003; Lee et al., 2015).

It is of great interest to propose complementary tools for reliable SAGD performance modeling. Artificial intelligence (AI) or data-driven modeling techniques entail analysis of data characterizing the system of interest and application of machine learning algorithms to build models suitable for system forecast or pattern recognition. AI-based methods have been widely

adopted in many engineering applications, such as software design (Rodríguez et al., 2016), river flow modeling (Kisi, 2004), and failure diagnosis in complex energy systems (Toffolo, 2009). These techniques are particularly useful with large datasets, as in the case of SAGD, where a large amount of data is available in the public domain. They may possibly provide alternative avenues for recovery performance prediction and uncertainty assessment, particularly when dealing with high-dimensional data consisting of a large number of operational and geological parameters.

In this work, artificial neural network (ANN) is employed to analyze SAGD production performance. ANN is a machine learning algorithm that is widely used for pattern recognition, classification, and prediction by mimicking information transfer in the central nervous system of the human beings (Haykin, 2008). ANN has been used to identify and approximate non-linear, complex and uncertain relationships between the input (predicting) and output (target) variables in a given dataset, which is a collection of records (samples); each record consists of a pair of input and output vectors.

ANN and other artificial intelligence approaches have been used in the petroleum exploration and production industry for several decades (Bravo et al., 214). Common ANN applications include reservoir parameter estimation (Adibifard et al., 2014), analysis of differential pipe sticking (Jahanbakhshi and Keshavarzi, 2016), sand production prediction (Khamehchi, et al., 2017), production forecast (Li et al., 2013), history-matching (Foroud, et al., 2014), study of drilling hydraulics (Wang and Salehi, 2015); recovery performance evaluation (Awoleke and Lane, 2011), production operation optimization and well design (Ayala and Ertekin, 2017) and fluid property calculations (Manshad et al., 2016). However, its application in heterogeneous SAGD reservoir analysis is limited. In particular, analysis involving field data from the McMurray deposits is rare.

Assembling a comprehensive dataset is an important step in any AI-based modeling workflow. In this work, a field dataset consisting of both geological and production information of Canadian SAGD wells is compiled from the public domain. In previous studies (Ma et al., 2014; 2015), a subset of the dataset extracted from only 3 nearby fields with comparable reservoir conditions was analyzed. In this work, field data gathered from a total of ten SAGD production fields are utilized to create an expanded dataset. Practical challenges associated with the data assembly and analysis processes due to noises, errors, and missing data are addressed.

Strategies for parameterizing pertinent predicting parameters, handling of high-dimensional datasets, and identifying internal structures among data are studied.

The forecast quality of ANN is often compromised due to probable inter-correlation between predicting variables, limited records in training data, and high dimensionality of input vectors. In order to improve model robustness and to achieve a better prediction performance, additional data-mining algorithms should be integrated. In this work, principal component analysis (PCA) is utilized to remove redundancy and correlation among the input variables in the original dataset, while retaining much of its information (Jolliffe, 2005). When faced with a large amount of data, it is shown that robustness and accuracy of the prediction capability are greatly enhanced by performing cluster analysis to identify internal data structures and groupings prior to ANN modeling (Amirian et al., 2015). Clustering analysis partitions a large dataset into a number of subsets with similar characteristics. This is a necessary step in this work since the original dataset encompasses data collected from ten different SAGD fields in Canada, with wide-ranging reservoir properties and production characteristics. K-means (MacQueen, 1967), which is a widely-adopted partitioning clustering algorithm, is applied in this work due to its simplicity and computational efficiency (Bahrololoum et al., 2015).

Due to the increased dataset size in this study, a number of new challenges arise. First, though the number of wells is large, complete data records can be extracted from only a portion of these wells, since measurements are not available everywhere. The general method for data analysis and assembly presented in (Ma et al., 2015) is adopted here, but a few modifications or adjustments are needed to construct a usable dataset. The impact of extrapolation of petrophysical parameters from the nearby vertical well is assessed in this study. Therefore, some additional input and output attributes are introduced. Second, dissimilar reservoir and operating conditions have led to significant variability among the data, identification of internal structures by means of clustering is needed to improve the prediction capabilities. The objectives of this chapter are: (1) extracting a comprehensive dataset of over 2000 SAGD wells and formulating a set of pertinent input attributes (descriptive of reservoir heterogeneities and operating conditions) and output attributes (representative of SAGD production performance); (2) demonstrating the potential to customize applied data-driven models constructed from actual field data for SAGD production forecast; (3) illustrating an improvement in robustness and reliability of the proposed modeling approach by integration of various data-mining techniques.

The chapter is organized as follows: first, the process of data assembly is discussed, and the theory or mathematical formulation of the proposed modeling workflow is explained in section 4.2; next, a case study illustrating the application of the described workflow is presented in section 4.3; finally, key findings are summarized in section 4.4.

4.2 Materials and Methods

4.2.1 Data Analysis and Assembly

A set of SAGD field data is assembled from the public domain. The available data consist of well trajectories/pairs, petrophysical logs, production/injection profiles. It should be noted that the dataset used in this study is limited to only those that are available in the public domain.

To construct a dataset suitable for ANN modeling, a number of input/output attributes describing the reservoir properties and production characteristics are extracted. The input attributes describing reservoir properties are obtained directly from logging interpretation, while the inputs pertinent to production characteristics are gathered from analyzing the injection and production data. As shown in **Fig. 4-1(b)**, for most SAGD projects, petrophysical logs are available only at the vertical wells (denoted as red circles), while production and injection data are available only at horizontal well pairs (denoted as tan lines). Vertical wells are often drilled near the horizontal pairs as delineation wells, from which petrophysical log and core data are sampled. Log data such as gamma ray and resistivity measurements are used to identify rock types and saturation distribution. Various cutoff values are assigned to identify sand/shale and water/oil zones. The selection of cutoff values is based on the specific formation characteristics and logging method. Details of the data extraction and analysis procedure can be found elsewhere (Ma et al., 2014; 2015). A number of improvements and modifications, however, have been implemented in this study. For instance, the search domain for the vertical wells has been refined. It is observed that generally speaking, petrophysical information derived from vertical wells that are located far away from the horizontal well pair do not provide a reliable description of heterogeneity at the horizontal well pair. This is because the separating distance may be larger than the physical length of correlation/continuity of the heterogeneous features (e.g., shale barriers). Therefore, a limited search distance of 60 m is applied.

Five input variables including porosity (ϕ), water saturation (S_w), pay zone thickness (h), net-to-gross ratio (N/G) and shale index (SI) can be extracted. The porosity (ratio of void space to the bulk volume) and water saturation are calculated as averages over the entire gross pay. N/G is the thickness ratio of net pay to gross pay, where net pay refers to the oil-saturated sand interval and gross pay refers to the entire production interval consisting of water-/oil-saturated sand/shale layers. The advancement of a steam chamber is often impeded by the presence of shale barriers; therefore, SI is formulated to capture the influence of a particular shale barrier. It is a normalized shale continuity indicator defined as the thickness-to-distance ratio of the shaly layer located at the shortest distance to the injector, i.e., h_{sh} /d_{sh_inj}. A large SI value would represent a thick shale barrier that is located very close to the injection well, with the potential to hamper steam chamber's growth and advancement. To further capture the uncertainty in the extrapolation of the aforementioned five parameters from the nearby vertical well, a new variable, which is the shortest distance between the horizontal well pair and the vertical well (d), is introduced as an input parameter. It serves to quantify the degree of correlation that may exist in properties extracted from the nearby vertical well locations and those at the corresponding horizontal well pair location. These six log-derived variables are considered as input attributes for subsequent ANN modeling.

Four additional input attributes related to well/operating conditions [i.e., effective number of oil production wells (N_e^{prod}), effective number of steam injection wells (N_e^{inj}), total production period of the given well pair (T_{total}), and cumulative steam injection (CSI)], plus two output attributes that summarize SAGD operational efficiency and recovery performance [i.e., duration over which the monthly average steam-to-oil ratio exceeds a particular threshold (TISOR) and cumulative oil production (COP)], are extracted from the available production data. CSI is defined as the total steam injection volume during the entire production/injection period. Due to the existence of some re-entry and infilled wells, which are drilled to enhance oil production, there could be multiple injector-producer pairs sharing the same original drainage area. In order to capture this irregularity, the effective well numbers are introduced as time-weighted averages. The total production period, T_{total}, is the time elapsed between the first and last production dates.

In this work, TISOR is an important variable in describing the efficiency of steam injection, and it is defined as the time period when the monthly average steam-to-oil ratio exceeds 2 (a

commonly-accepted upper limit for typical SAGD wells). For a given value of COP, a low value of TISOR would correspond to higher steam injection efficiency.

4.2.2 Artificial Neural Network

ANN is applied to build various data-driven models for SAGD production analysis in this work. The basic ANN architecture contains an input layer, an output layer and any number of hidden layers. An example of a fully-connected neural network with one hidden layer is shown in **Fig. 4-2**. The input layer consists of parameters that are related to the target variables in the output layer. Each node (or neuron) in the network configuration (except input nodes) contains a bias term and is connected by weights to other nodes in the preceding and posterior layers. The backpropagation approach, a gradient-based supervised learning algorithm, is adopted here to train the ANN. Details of the ANN formulations were discussed by Haykin (2008).

4.2.3 Principal Component Analysis and Clustering Analysis

Though the original input vector is small in this study, information in the data samples may still be redundant. PCA can be applied to remove this redundancy by converting a high-dimensional input vector into a set of linearly uncorrelated features, or principal components (*PC*), with relatively lower dimensionality through orthogonal transformation. Once the *PCs* are determined, the original dataset is transformed into a set of principal scores (*PS*), which are assembled into an input vector for subsequent ANN modeling.

Cluster analysis facilitates the identification of internal structures among the data by partitioning a large dataset into numerous sub-datasets with similar characteristics. For a specified number of subgroups, k-means algorithm assigns data points to individual groups by minimizing the within-cluster sum of squares according to **Eq. 4-1**:

$$S_{\mathrm{c}} = \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} I_{ij}(\mathbf{X}_i - \overline{\mathbf{X}}_j)(\mathbf{X}_i - \overline{\mathbf{X}}_j)^T \quad \text{..... (4-1)}$$

where S_c is the within-cluster sum of squares; g stands for the number of groups, I_{ij} equals to 1 if a data sample $\mathbf{X_i}$ belongs to cluster j and 0 otherwise; $\overline{\mathbf{X_j}}$ represents the center of cluster j, which

is the arithmetic average of all data objects in this cluster. The within-cluster sum of squares is the objective function to be minimized through iterations.

Details about the theory and formulation of k-means clustering were explained by Hammouda (2000). One shortcoming with this algorithm is that the final clustering results are dependent on the initialization of cluster centers; therefore, it is necessary to repeat the minimization procedure with numerous initial guesses in order to identify the optimal groupings.

4.2.4 Workflow for Production Analysis

The adopted workflow for SAGD production analysis is summarized in **Fig. 4-3**. After performing the exploratory data extraction and analysis, an original dataset consisting of 10 input variables and 2 output variables is assembled. Next, PCA is performed to reduce the dimensionality of this dataset. This new dataset is subdivided randomly into two parts for ANN model construction: (1) training set (80%) – optimizing the network architecture and estimating network parameters; (2) testing set (20%) – evaluating the performance of the trained network.

An 80-20 split for training and testing purposes is reasonable for typical ANN applications. A hold-out validation scheme is implemented, as it offers an acceptable approximation of the model's true prediction error with less computational requirement in comparison to the n-fold cross-validation (Noureldin et al., 2007). K-means clustering is applied to the training set, and two distinct clusters are identified; next, this clustering result is applied to the testing samples as well. The rationale for performing PCA prior to clustering is two folds: (1) it ensures identical principal scores for network training in both clusters; and (2) the orthogonality in the principal component space would facilitate the subsequent cluster analysis. For each cluster, the optimal network configuration is determined by a n-fold cross-validation, as illustrated in chapter 2. The ANN model with the optimal configuration in each cluster is then trained and evaluated using corresponding training and testing dataset, respectively.

4.3 Case Study

4.3.1 Dataset Description and Data Mining

The dataset employed in this work is extracted from ten producing SAGD fields encompassing a variety of reservoir conditions and operational constraints. There are over 2700 wells in total,

with approximately 1100 well pairs among these fields. This large number of wells poses various challenges including incomplete and noisy data. In many instances, there is insufficient data to formulate a complete record of input and output attributes for a given well pair. For example, logging wells are unavailable in certain areas, while production history is too short for some well pairs. Following the data analysis strategy discussed in section 4.2.1, a total of 153 data samples are assembled. The dataset is shown in **Table A-1**. It is a significantly larger dataset in comparison to the 71 records presented in a previous study (Ma et al., 2015). Each record is composed of 10 input attributes (representing reservoir properties and production constraints) plus 2 output attributes (describing SAGD recovery performance and operational efficiency).

Table 4-1 summarizes the key information of all data samples extracted from the ten fields. Most of these fields are accompanied by long production histories and large values of COP. It appears that the average cumulative steam-to-oil ratio (cSOR) values are quite high, which may be attributed to the presence of reservoir heterogeneities. It should be noted that the number of producers is often greater than that of injectors due to the existence of re-entry and infilled production wells.

The correlation plots between 12 variables, or the 'corrplot' (Wei, 2013), are generated and shown in **Fig. 4-4**. The actual data points are not shown, but the ellipse's size and color reflect the nature and magnitude of the correlation between an individual pair of variables. There exist a noticeable positive correlation between TISOR and COP. A strong negative correlation is observed between ϕ and S_w; this trend is reasonable, as S_w often increases in tight rocks or shales. As expected, three input attributes, including T_{total}, N_e^{prod}, and CSI, which are pertinent to production characteristics, are exhibiting strong positive correlations with the two outputs. In order to assess data redundancy and its impact on training performance, input variable importance analysis is considered next. Following the formulation presented in Sung (1988), the relative importance of individual input parameter is assessed by calculating the change in mean squared error (*MSE*) between training with the original dataset and training in the absence of the specific input attribute. Due to the huge discrepancies in the absolute values of the two output variables, normalized values are computed in this *MSE* analysis. In order to avoid sensitivity due to random initialization of network parameters, training of each ANN with the optimal structure (single hidden layer with 18 hidden neurons) is repeated numerous times; the model with the least *MSE* is selected. The corresponding input variable importance plot is shown in **Fig. 4-5**. It

is obvious that 3 input variables including T_{total}, CSI, and N_e^{prod} are dominating. This observation confirms the conclusions that are derived from **Fig. 4-4**. However, as impacts of the remaining variables are still considerable, removing them from the dataset explicitly is not recommended. Instead, PCA technique is employed to reduce the redundancy in input variables and to generate new features for ANN modeling in this work.

Results of the PCA are presented in **Fig. 4-6**. It is observed from the variance plot in **Fig. 4-6(a)** that the first 6 components have captured the majority of the variability (≥ 90%) exhibited by the data; therefore, the number of principal components is selected to be 6 (discarding the remaining 3 components). **Fig. 4-6(b)** depicts the bi-plot between the first two principal components, illustrating the variance contribution of each input variable to the two respective principal components. For instance, S_w has the largest coefficient corresponding to *PC* 1, suggesting that *PC* 1 is highly capable of representing the information related to S_w. Additionally, *PC* 1 is capable of distinguishing samples with large positive values of S_w, h, and d and small negative values of ϕ and N/G. The locations of all 153 samples are also shown. The correlation coefficient among the resultant principal scores is essentially zero, which is much lower than that of the original dataset shown in the **Fig. 4-4**. Therefore, PCA is shown to be effective in eliminating the redundancy and correlation among the original data.

A set of 30 samples are randomly selected to be the testing dataset, while the remaining are used for training purposes (i.e., $K = 30$ and $M - K = 123$, following the notation in **Fig. 4-3**). K-means clustering is applied to partition the training subset into two clusters. Due to the limited number of training samples, only two clusters are identified, with 54 samples and 69 samples in the first and second cluster, respectively. The testing samples are assigned to these two clusters according to their Euclidian distances to the cluster centers. From the scatter plots between the first two principal scores in **Fig. 4-7**, two distinct clustered distributions can be detected. The result of the cluster analysis is summarized in **Fig. 4-8**. For most fields such as Jackfish, Firebag, and Leismer, their respective samples belong to a single cluster; while for fields such as Cristina Lake (1), their samples are grouped predominantly into a single cluster, with only a few exceptions. This observation reflects the capability of k-means clustering in identifying internal groupings pertinent to individual reservoir and production characteristics. However, for others such as Surmont, their samples are almost evenly split into two clusters. This outcome reflects the fact that clustering is also controlled by a nonlinear interplay between all input attributes

encompassing both production constraints and reservoir properties. For instance, two samples could belong to different clusters, despite sharing a similarity in reservoir characteristics (e.g., ϕ, S_w, h, N/G, and SI), if their production conditions (e.g., the total number of producers/injectors or CSI) vary dramatically.

The histograms of 10 original inputs variables from each cluster are compared visually in **Fig. 4-9**. Significant differences are detected between the distributions of the two clusters, confirming the presence of internal structures among the dataset. It is clear that reservoirs in cluster 2 are of better quality with higher ϕ, lower S_w, and higher N/G, even though they are much thinner reservoirs. Because of these differences, wells in cluster 2 generally require fewer producers and injectors (i.e., lower N_e^{prod} and N_e^{inj}) and are capable of maintaining a higher CSI. It is interesting to note that while operating conditions are normally considered to be artificial or controllable factors, which vary among different operators, the data would suggest that production strategies (e.g., well placement and flowing pressures) are still reflective of the underlying reservoir quality. Therefore, despite some overlapping among the two clusters in the last five input variables, obvious differences in operational variables can be detected among these two groups. Finally, given the distributions of d are similar for these two clusters, it is inferred that any bias that may occur due to extrapolation of the logging data is comparable for both clusters.

4.3.2 ANN Modeling Results

The data is first analyzed to assess the impact of extrapolation of petrophysical parameters from the nearby vertical well. This is achieved by dividing the dataset into two groups: (group 1 with $d \geq 22$ m and group 2 with $d < 22$ m, where the median of d is 22 m). A 5-fold cross-validation is employed to identify the optimal ANN architecture for each group. Configurations with both single and two hidden layers are considered in the network architecture exploration process. Finally, the optimal architectures for the first and second group are determined as two hidden layers with 15 × 12 hidden nodes and single hidden layer with 7 hidden nodes, respectively.

The corresponding ANN performances for both training and testing datasets are shown in **Fig. 4-10**. The overall performance is quite satisfactory for both groups, as evidenced by the coefficient of determination (R^2) being close to 1 and the low *MSE*. It appears that the adopted ANN approach is able to capture these non-linear relationships exhibited in the field data.

However, the more interesting observation from **Fig. 4-10** is that the ANN prediction for group 2 (R^2 being 0.90 for both TISOR and COP) is superior in comparison to group 1 (R^2 being 0.81 for TISOR and 0.73 for COP, respectively). The corresponding *MSE* values are also reduced in the predictions for group 2. It is expected that the production performance would be highly dependent on the petrophysical properties; a small *d* value may suggest that the information extracted from the nearby vertical well is more likely to represent what would actually be observed at the horizontal well pair. Therefore, a stronger relationship seems to exist between the petrophysical variables (inputs) and the production variables (outputs) for group 2. This stronger relationship contributes to the better prediction capability of the ANN model for group 2.

This difference highlights the significance of extrapolation and the input parameter *d* in capturing the corresponding uncertainty. In an ideal setting, one would include only data from vertical wells that are located close to the producing well pairs; unfortunately, high costs for data collection in subsurface engineering often limits the amount of data that is available, the incorporation of *d* as an input parameter offers a simple, yet effective, way to construct a larger usable dataset, while quantifying the associated uncertainty due to extrapolation.

Though it is possible to divide the entire dataset into two groups based on an arbitrary threshold of *d*, a more rigorous technique should be applied to identify the internal structures (if there is any) by consideration of all the variables. Next, k-means clustering is applied to the entire dataset, and separate ANN model is constructed for each cluster. Once again, a 5-fold cross-validation is employed to optimize the ANN architecture. Finally, the optimal architectures for the first and second cluster are determined to be a single hidden layer with 6 hidden nodes and two layers with 10×6 hidden nodes, respectively.

The ANN performances for the training and testing datasets after k-means clustering are compared in **Fig. 4-11**. Good agreement is observed with both the training and testing datasets, as most points follow the 45° separating line that indicates a perfect correlation. The value of R^2 is close to unity. The results are encouraging, as they serve to illustrate the capability of the implemented workflow in achieving a reliable SAGD production analysis. For both clusters, except for a few data points, the overall forecasting performance of the ANN model is acceptable. The presence of these exceptions may be attributed to the limited number of records in the dataset and random allocation of records into the training and testing subsets, which contribute to extrapolating the trained ANN models when the testing data is applied. Since extrapolation in

either the input or output parameter spaces is usually hard to detect and avoid, increasing the number of data samples could potentially improve the prediction accuracy. The residual errors (e_{res}) and relative errors (e_{rel}) between the targets and predicted values from this ANN model are presented in **Fig. 4-12**. It is important to note that the residual error is independent of the actual magnitude of the predicted value (i.e., homoscedastic) and approximately Gaussian (i.e., normally distributed). Except for a few outliers, the relative errors corresponding to most data samples are quite small.

An important novelty of this implementation is that the distance between the horizontal well pair and the vertical well (d) has been incorporated as an additional input variable, which is shown to improve the predictive capability of the model. It helps to capture the imprecision due to extrapolation from the nearby logging well to the horizontal well pair location. It should be noted that the model performance will be enhanced if more samples become available in the future (e.g., drilling of new wells and/or extraction of additional information from existing dataset). However, it should be cautioned that ANN predictive ability is often compromised by excessive noise in the data. Therefore, care should be taken to extract additional samples without forfeiting the data quality (e.g., correlating vertical wells and horizontal well pair that are located too far apart).

To illustrate the functionality of PCA, two other ANN models are trained using all 10 original input variables for the same clusters in **Fig. 4-11**; an optimized model with two hidden layers and 7 × 4 neurons is used for cluster 1, while a single layer with 3 neurons is used for cluster 2. The results are presented in **Fig. 4-13**. The performances of the models with only 6 principle scores (**Fig. 4-11**) is comparable to that using all 10 input variables in the original dataset, confirming that PCA is useful for removing data redundancy without compromising prediction quality, despite the limited number of data records in the dataset. However, PCA has not improved the prediction quality in this case. The small set of input attributes has been deliberately selected to span the system parameters; hence, a relatively weak correlation is exhibited by these variables (**Fig. 4-4**). It is likely that as we expand the dataset to encompass more variables, the improvement with PCA may be more dramatic. An important implication is that since most reservoir-engineering datasets suffer from data scarcity (limited dataset size), the application of supervised learning methods may benefit from selecting an appropriate set of

input/output attributes based on domain knowledge and utilizing techniques to reduce data dimensionality.

In order to illustrate the performance of clustering process, two additional experiments are conducted. First, the training dataset for each cluster remains unchanged, while the testing datasets are switched (i.e., the testing dataset of the second cluster is subjected to the ANN model of the first cluster and vice versa). The corresponding results are illustrated in **Fig. 4-14**. Comparing **Fig. 4-14** to their counterparts in **Fig. 4-11**, more scattering is observed with larger values of *MSE* and lower values of R^2. It is clear that significant variability exists among the training and testing samples; therefore, cluster analysis is essential in isolating the impacts of internal groupings or structures among the data. Next, all samples in the training dataset are used to construct a single ANN model (with an optimal architecture of a single hidden layer with 5 hidden neurons) without performing cluster analysis. The predictive capability of the ANN model, as presented in **Fig. 4-15**, is relatively inferior in comparison to that in **Fig. 4-11**, particularly for the case of COP. However, this improvement as a result of clustering is not overly significant. This is likely due to the weak partitioning or non-distinct boundary between the two clusters (**Fig. 4-7**). These two experiments have confirmed the benefits of cluster analysis for this particular dataset. In general, it is anticipated that the improvement would be more noteworthy as the dataset size increases.

It should be emphasized that the results presented here do not intend to dismiss the use of conventional flow simulations; it offers, instead, a complementary alternative for SAGD analysis. Construction of simulation models is often a labor- and time-intensive task, while many assumptions involving fluid and reservoir properties are required. Therefore, the use of data-driven models can assist decision-making in SAGD operations; for example, proposing future well locations, evaluating land sales or farm-in opportunities, establishing uncertainty (e.g., P10/P50/P90) in production forecasts. Higher confidence can be achieved by corroborating simulation results with models derived from actual field data.

4.3.3 Uncertainty Analysis Results

Uncertainty analysis on all two output variables is performed using the Monte Carlo simulation and bootstrapping method, as described in detail by Ma et al. (2015). Three sources of uncertainty, including uncertainty in model parameter, uncertainty in input variables due to

imprecise analysis criteria and uncertainty in dataset due to limited number of records, are estimated. First, the Monte Carlo method is employed to study the uncertainty originated from ANN model parameters, where ANN training for each cluster is repeated 1000 times with randomized initialization of weights and biases using the optimal network configurations determined previously. Therefore, 1000 output vectors are generated corresponding to 1000 trained ANN models for each testing data record. Next, the input data uncertainty due to imprecise analysis criteria (e.g., outliers and petrophysical log cut-offs) is assessed in a parametric bootstrapping scheme. A total of 1000 realizations of each input variable are sampled from a uniform likelihood distribution with a +/- 10% variation in the attribute value. This step essentially generates a total of 1000 datasets (all consist of the same number of records as the original dataset). Each dataset is subjected to PCA, k-means clustering, and ANN training to generate 1000 models. For a given testing sample, the corresponding output uncertainty can be computed from predictions generated from all 1000 trained models. Finally, another bootstrap approach is employed to evaluate the uncertainty derived from a limited number of data samples (e.g., insufficient well pairs). A total of 1000 new training datasets are sampled from the original training dataset with replacement while keeping the final testing dataset unchanged. Once again, all 1000 datasets that consist of the same number of records serve as the original datasets). Training of the optimal neural network for individual clusters is repeated with each new dataset, and the output uncertainty for the 30 final testing samples can be quantified from final predictions obtained from all 1000 trained models.

In order to provide a comprehensive comparison of output uncertainties resulting from all three sources, their corresponding values (in the form of standard deviation over all 1000 ANN predictions for the 30 testing samples) are summarized in **Table 4-2**. For most samples, the largest output uncertainty is attributed to the limited number of records in the dataset and input data uncertainty. This finding would suggest that expanding the dataset could be beneficial in reducing the uncertainty in ANN predictions. This can be achieved by collecting data from more wells or producing fields, as well as reprocessing the existing data and/or refining the log analysis criteria for specific fields to extract additional information. It is also observed in **Table 4-2** that uncertainties corresponding to a few particular samples (e.g., #6) appear to be too large. This may be the result of extrapolation beyond the subspace spanned by the data.

This observation is different from the one presented in Ma et al. (2015), which suggests that the primary sources of uncertainty are those in the ANN model parameters and limited data samples. Therefore, it is expected that a larger dataset could significantly reduce the overall uncertainty. However, this is not what has been observed in **Table 4-2**. Several factors may explain this inconsistency. First, though the number of data samples in the previous study (Ma et al., 2015), was smaller, those samples were extracted from only three fields within close proximity to each other; it is reasonable to expect that there are significant similarities, in terms of both reservoir characteristics and operating conditions, among those fields. In contrast, ten fields with widely varying characteristics are included in the current study. Although the number of data samples has increased to 153, it is still likely that a much larger dataset is required to sufficiently span the variable spaces. Although the cluster analysis has helped to identify the internal structures among this heterogeneous dataset, it also inadvertently reduces the number of samples available for ANN modeling corresponding to each cluster. Second, the aforementioned issue is exacerbated by the fact that more input and output attributes have been included in this study, while the degree of freedom increases. At the end, this analysis illustrates an important notion that model prediction capability and uncertainty are strongly dependent on the overall quality of the dataset, even if identical quantification approaches have been employed. Therefore, for a given optimized model training algorithm, in order to reduce the cumulative uncertainty in the final model prediction, attaining a larger dataset and identifying internal structures are equally important.

4.4 Conclusion

A comprehensive dataset encompassing ten SAGD fields is compiled from public sources. Important variables related to reservoir heterogeneity and production constraints are identified through logging interpretation and production analysis. The most important contribution is the inclusion of a new input variable that captures the uncertainty due to extrapolation of petrophysical parameters from the nearby vertical well. Models of artificial neural network are utilized to facilitate SAGD production forecast. Application of such models with a large SAGD field dataset is novel.

Principal component analysis and cluster analysis are applied to improve the forecast capacity, efficiency, and robustness of the proposed approach. Forecast performances of these models are shown to be both reliable and satisfactory. Influences of the uncertainties originating from model parameter, input attribute variability and limited data records on the final ANN predictions are investigated. The analysis reveals that uncertainty due to limited data records is dominating. This result motivates future efforts in expanding the available dataset for uncertainty management.

This work proposes a feasible complementary alternative to traditional numerical simulation for SAGD production analysis. In general, data such as bottom-hole pressures, fluid properties, permeability, multi-phase flow functions and thermal conductivities are unavailable in the public domain, rendering construction of detailed simulation models challenging without extensive assumptions. This work demonstrates how practical data-driven models can be useful for predicting SAGD recovery performance from log-derived and operational variables alone. Results presented in this study are applicable to data ranges represented in the dataset. Extrapolation beyond the subspace spanned by the data should be cautioned. Nevertheless, the modeling framework presented here can be readily extended to incorporate additional datasets or input variables.

Future studies should incorporate additional input/output attributes (e.g., steam-to-oil ratios) and analyze the compounding effects of all aspects of uncertainty. Increasing the number of data records may alleviate extrapolation in the input/output parameter space and reduce prediction uncertainty.

4.5 Reference

Butler, R., McNab, G., & Lo, H. (1981). Theoretical studies on the gravity drainage of heavy oil during in‐situ steam heating. *The Canadian Journal of Chemical Engineering, 59*(4), 455-460.

Bagci, A. S. (2006). Experimental and simulation studies of SAGD process in fractured reservoirs. Paper presented as the *SPE/DOE Symposium on Improved Oil Recovery*, Tulsa, Oklahoma.

Sasaki, K., Akibayashi, S., & Kosukegawa, H. (1996). Experimental study on initial stage of SAGD process using 2-dimensional scaled model for heavy oil recovery. Paper presented at the *International Conference on Horizontal Well Technology*, Calgary, Alberta, Canada.

Fatemi, S. M. (2009). Simulation study of steam assisted gravity drainage (SAGD) in fractured systems. *Oil & Gas Science and Technology-Revue De l'IFP, 64*(4), 477-487.

Shin, H., Hwang, T., & Chon, B. (2012). Optimal grid system design for SAGD simulation. Paper presented at *SPE Heavy Oil Conference Canada*, Calgary, Alberta, Canada.

Panwar, A., Trivedi, J. J., & Nejadi, S. (2015). Importance of distributed temperature sensor data for steam assisted gravity drainage reservoir characterization and history matching within ensemble Kalman filter framework. *Journal of Energy Resources Technology, 137*(4) pp. 042902.

Queipo, N. V., Goicochea, J. V., & Pintos, S. (2002). Surrogate modeling-based optimization of SAGD processes. *Journal of Petroleum Science and Engineering, 35(*1), 83-93.

Lacroix, S., Renard, G., Lemonnier, P. (2003). Enhanced numerical simulations of ior processes through dynamic sub-gridding. Paper presented at *Canadian International Petroleum Conference*, Calgary, Alberta, Canada.

Lee, H., Jin, J., & Shin, H. (2015). Efficient prediction of SAGD productions using static factor clustering. *Journal of Energy Resources Technology, 137*(3) pp. 032907.

Rodríguez, G., Soria, Á., & Campo, M. (2016). Artificial intelligence in service-oriented software design. *Engineering Applications of Artificial Intelligence*, *53*.

Kisi, Ö., (2004). River flow modeling using artificial neural networks. *Journal of Hydrologic Engineering, 9*(1), 60-63.

Toffolo, A. (2009). Fuzzy expert systems for the diagnosis of component and sensor faults in complex energy systems. *Journal of Energy Resources Technology, 131*(4).

Haykin, S.S. (2008). *Neural networks and learning machines* (3rd ed.), Upper Saddle River, NJ, USA: Pearson.

Bravo, C. E., Saputelli, L., & Rivas, F. (2014). State of the art of artificial intelligence and predictive analytics in the e&p industry: a technology survey. *SPE Journal*, *19*(04).

Adibifard, M., Tabatabaei-Nejad, S., & Khodapanah, E. (2014). Artificial neural network (ANN) to estimate reservoir parameters in naturally fractured reservoirs using well test data. *Journal of Petroleum Science and Engineering, 122*.

Jahanbakhshi, R., & Keshavarzi, R. (2016). Intelligent classifier approach for prediction and sensitivity analysis of differential pipe sticking: a comparative study. *Journal of Energy Resources Technology, 138*(5) pp. 052904.

Khamehchi, E., Kivi, I. R., & Akbari, M. (2014). A novel approach to sand production prediction using artificial intelligence. *Journal of Petroleum Science and Engineering, 123*, 147-154.

Li, X., Chan, C., & Nguyen, H. (2013). Application of the neural decision tree approach for prediction of petroleum production. *Journal of Petroleum Science and Engineering, 104*, 11-16.

Foroud, T., Seifi, A., & AminShahidi, B. (2014). Assisted history matching using artificial neural network based global optimization method–applications to Brugge field and a fractured Iranian reservoir. *Journal of Petroleum Science and Engineering, 123*, 46-61.

Wang, Y., & Salehi, S. (2015). Application of Real-Time Field Data to Optimize Drilling Hydraulics using Neural Network Approach. *Journal of Energy Resources Technology, 137*(6) pp. 062903.

Awoleke, O., & Lane, R. (2011). Analysis of Data from the Barnett Shale using Conventional Statistical and Virtual Intelligence Techniques. *SPE Reservoir Evaluation & Engineering, 14*(05) , 544-556.

Ayala H, L. F., & Ertekin, T. (2007). Neuro-Simulation Analysis of Pressure Maintenance Operations in Gas Condensate Reservoirs. *Journal of Petroleum Science and Engineering, 58*(1), 207-226.

Manshad, A. K., Rostami, H., & Hosseini, S. M. (2016). Application of artificial neural network-particle swarm optimization algorithm for prediction of gas condensate dew point pressure and comparison with Gaussian processes regression-particle swarm optimization algorithm. *Journal of Energy Resources Technology, 138*(3) pp. 032903.

Ma, Z., Leung, J. Y., Zanon, S., & Dzurman, P. (2014). Practical implementation of knowledge-based approaches for SAGD production analysis. Paper presented at the *SPE Heavy Oil Conference Canada*, Calgary, Alberta, Canada.

Ma, Z., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Practical implementation of knowledge-based approaches for steam-assisted gravity drainage production analysis. *Expert Systems with Applications, 42*(21), 7326-7343.

Jolliffe, I. (2005). *Principal component analysis*, Wiley Online Library.

Amirian, E., Leung, J. Y., & Zanon, S. (2015). Integrated cluster analysis and artificial neural network modeling for steam-assisted gravity drainage performance prediction in heterogeneous reservoirs, *Expert Systems with Applications, 42*(2), 723-740.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Anonymous California, USA, 1, 281-297.

Bahrololoum, A., Nezamabadi-pour, H., & Saryazdi, S. (2015). A data clustering approach based on universal gravity rule. *Engineering Applications of Artificial Intelligence*, *45*, 415-428.

Shlens, J. (2014). A tutorial on principal component analysis. *ArXiv Preprint arXiv:1404.1100*.

Hammouda, K., & Karray, F. (2000). A comparative study of data clustering techniques. University of Waterloo, Ontario, Canada.

Noureldin, A., El-Shafie, A., & Taha, M. R. (2007). Optimizing neuro-fuzzy modules for data fusion of vehicular navigation systems using temporal cross-validation. *Engineering Applications of Artificial Intelligence, 20*(1), 49-61.

Wei, T. (2013). Corrplot: visualization of a correlation matrix, *R Package Version 0.73*.

Sung, A. (1998). Ranking importance of input parameters of neural networks, *Expert Systems with Applications*, *15*(3), 405–411.

Tables

Table 4-1 Overview of ten producing fields incorporated in this study.

Field index	Christina Lake (1)	Foster Creek	Jackfish	Christina Lake (2)	Firebag
Average T_{total} (years)	3.4	6	3.5	4.8	6.5
Range of T_{total} (years)	0.4-10.8	1.2-11.7	0.4-6.2	2.3-5.8	1.4-9.6
Average COP (m^3)	203000	299500	123420	286440	607000
Range of COP (m^3)	26523-481350	22611-798880	28250-250130	95105-643300	75368-1025400
Average cSOR	1.8	3.5	2.7	3.2	5.3
Range of cSOR	0.2-6.8	1.0-6.9	0.2-3.7	1.8-4.8	1.3-8.1
Number of Injectors	114	282	133	127	177
Number of Producers	127	434	172	132	223
Number of logging wells	58	50	24	21	10
Range of Porosity	0.27-0.33	0.17-0.35	0.19-0.27	0.21-0.32	0.23-0.29
Range of Pay Thickness (m)	34.6-79.3	12.6-59.5	47.7-96.9	24.6-93.7	70.2-84.1
Range of d (m)	1.9-48.3	4.5-55.8	3.1-53.9	5.9-55.9	26.6-40.0
Field index	Great Divide	Hangingstone	Leismer	MacKay River	Surmont
Average T_{total} (years)	4.6	8	2.5	8.2	6.1
Range of T_{total} (years)	2.8-5.7	0.8-14.3	2.3-2.7	5.3-10.7	1.6-15.9

Average COP (m^3)	85138	207760	88995	199570	164080
Range of COP (m^3)	32861-140410	3434-548780	46893-119790	45420-398780	22383-376710
Average cSOR	4.2	6.6	2.1	5.2	3.9
Range of cSOR	2.5-5.3	0.4-11.9	1.4-2.5	2.3-9.2	0.4-13.1
Number of Injectors	40	69	30	128	109
Number of Producers	42	71	30	128	99
Number of logging wells	17	91	31	46	58
Range of Porosity	0.22-0.34	0.13-0.31	0.33-0.35	0.22-0.32	0.16-0.36
Range of Pay Thickness (m)	5.6-40.6	21.0-63.4	22.1-31.1	25.7-38.6	13.2-71.9
Range of *d* (m)	3.6-45.5	7.5-51.0	5.3-54.8	6.6-48.3	2.5-46.9

Table 4-2 Comparison of relative uncertainty (standard deviation divided by target value) from different sources (model parameter uncertainty, input data uncertainty and uncertainty due to limited dataset size).

Output Samples index	TISOR			COP		
	Model parameter	Input data	Dataset size	Model parameter	Input data	Dataset size
1	0.184	0.225	0.280	0.227	0.258	0.316
2	0.126	0.181	0.202	0.201	0.267	0.286
3	0.132	0.202	0.213	0.237	0.281	0.322
4	0.147	0.184	0.217	0.187	0.200	0.228
5	0.336	0.532	0.708	0.147	0.203	0.370

6	0.704	1.081	1.066	0.234	0.320	0.356
7	0.182	0.262	0.291	0.708	0.961	1.103
8	0.164	0.247	0.281	0.097	0.144	0.154
9	0.191	0.241	0.275	0.161	0.194	0.211
10	0.113	0.163	0.182	0.173	0.239	0.257
11	0.200	0.229	0.277	0.227	0.242	0.259
12	0.158	0.180	0.202	0.308	0.387	0.399
13	0.259	0.301	0.343	0.351	0.381	0.407
14	0.254	0.274	0.381	0.618	0.614	0.791
15	0.342	0.384	0.474	1.031	1.266	1.438
16	0.101	0.126	0.133	0.240	0.314	0.296
17	0.121	0.161	0.166	0.210	0.272	0.272
18	0.661	0.645	0.849	1.007	0.898	1.078
19	0.092	0.118	0.141	0.124	0.159	0.161
20	0.102	0.136	0.163	0.274	0.344	0.369
21	0.122	0.132	0.162	0.206	0.218	0.235
22	0.101	0.122	0.135	0.310	0.329	0.358
23	0.122	0.164	0.186	0.423	0.574	0.659
24	0.084	0.160	0.135	0.248	0.404	0.380
25	0.168	0.283	0.301	0.258	0.374	0.420

26	0.128	0.193	0.230	0.163	0.238	0.249
27	0.071	0.107	0.129	0.307	0.428	0.535
28	0.144	0.172	0.207	0.179	0.229	0.249
29	0.116	0.179	0.183	0.466	0.664	0.683
30	0.288	0.395	0.511	0.340	0.465	0.624

Figures

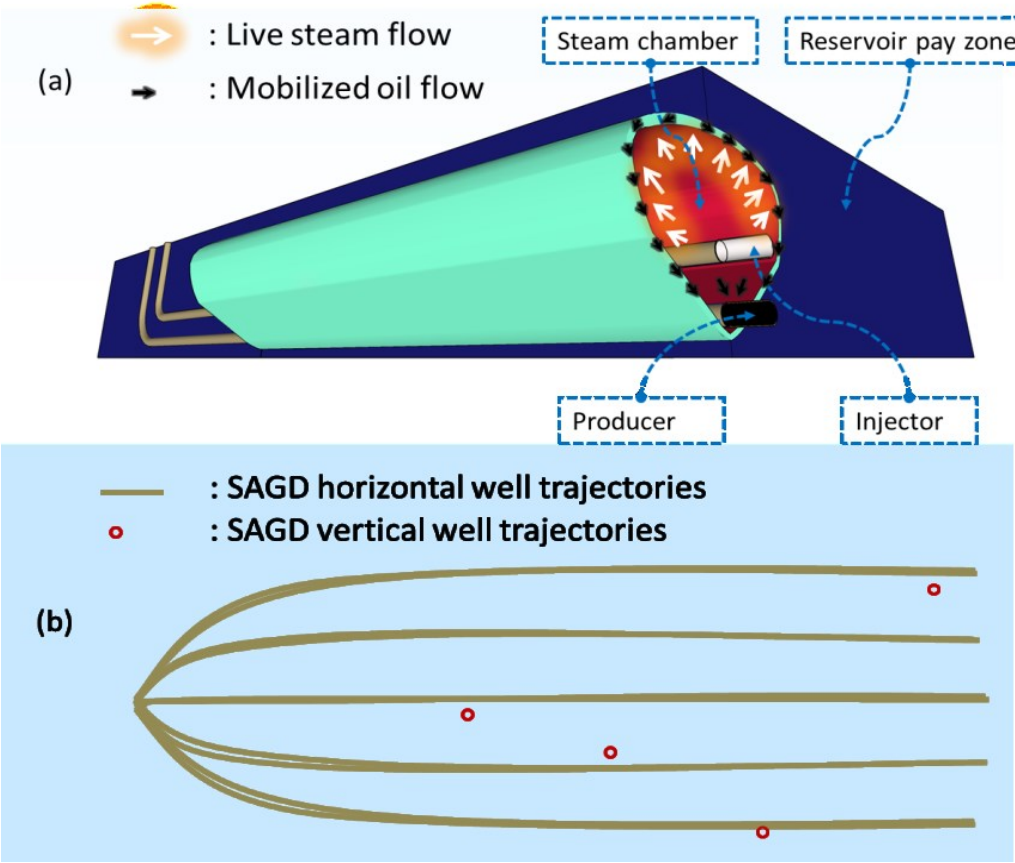


Fig. 4-1 Typical SAGD project: (a) – schematic of a single well pair in 3D; (b) – schematic of a well pad consisting of 5 well pairs in top view.

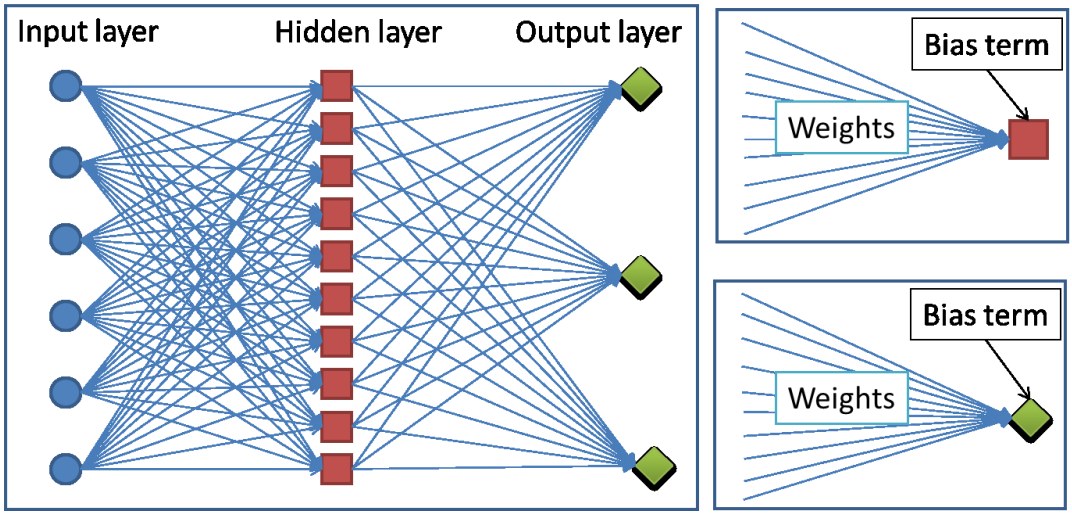


Fig. 4-2 Neural network architecture with only one hidden layer.

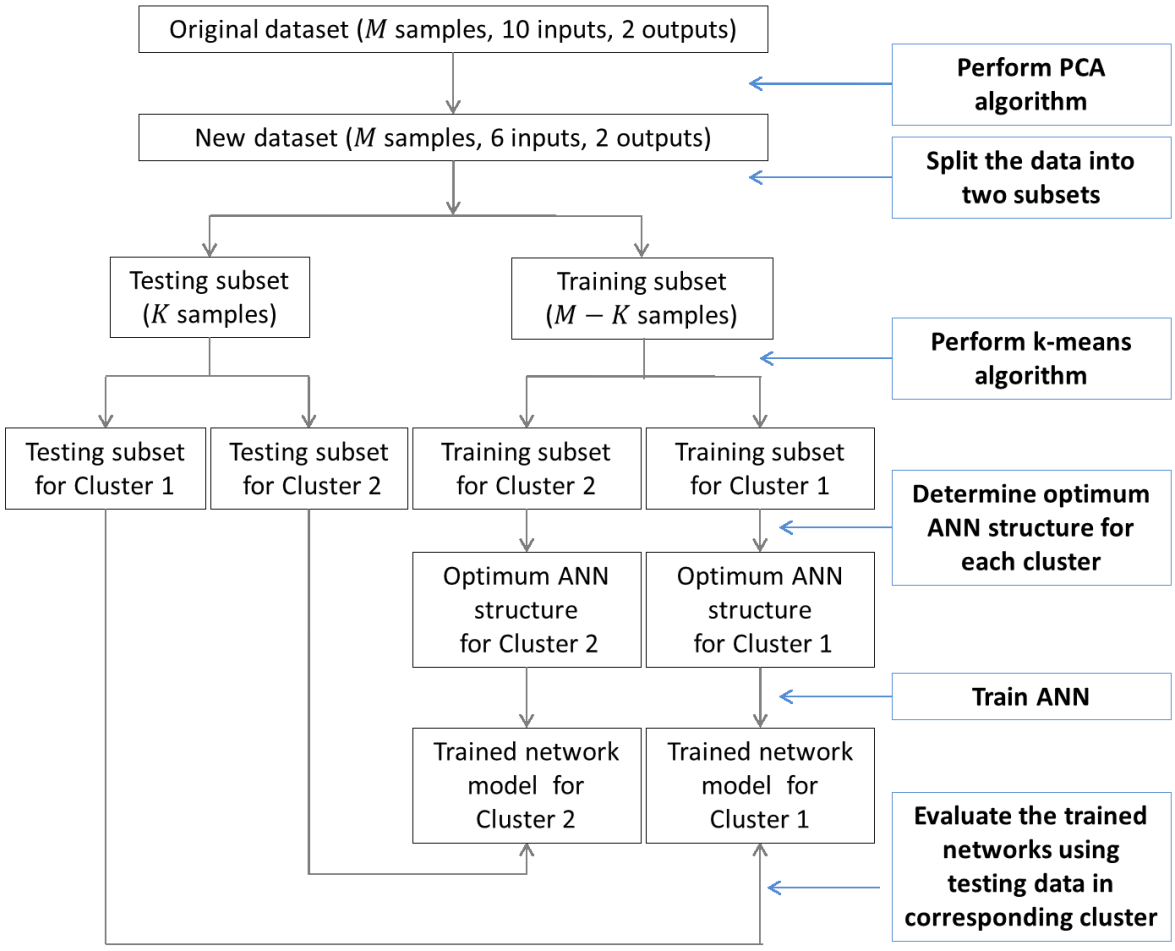


Fig. 4-3 Flowchart of the adopted analysis workflow.

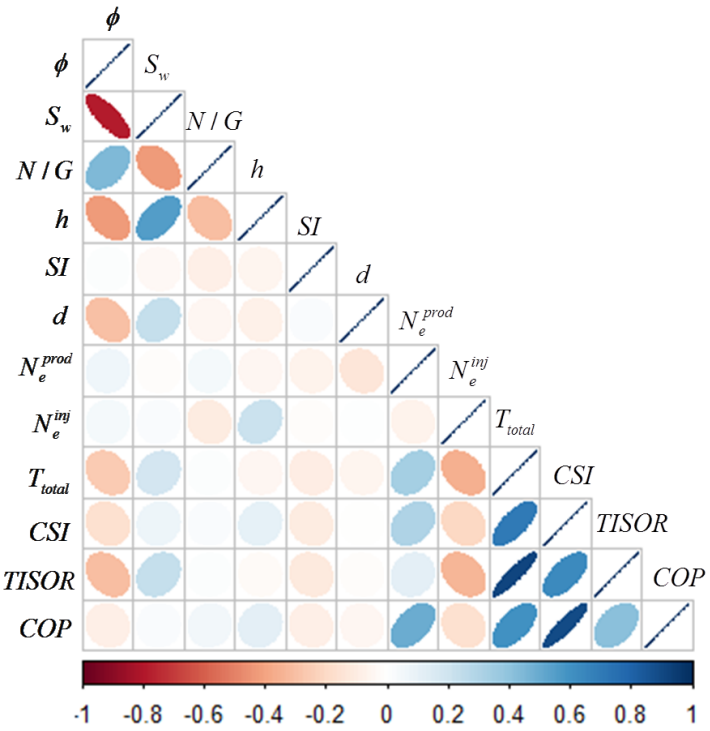


Fig. 4-4 Correlation plot of the original dataset: large positive value indicates a strong positive correlation between the two parameters; low negative value indicates a strong negative correlation between the two parameters.

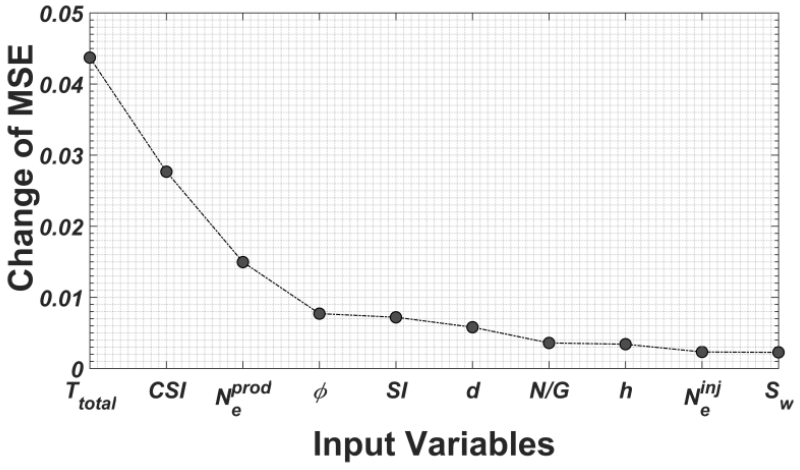


Fig. 4-5 Input variable importance plot.

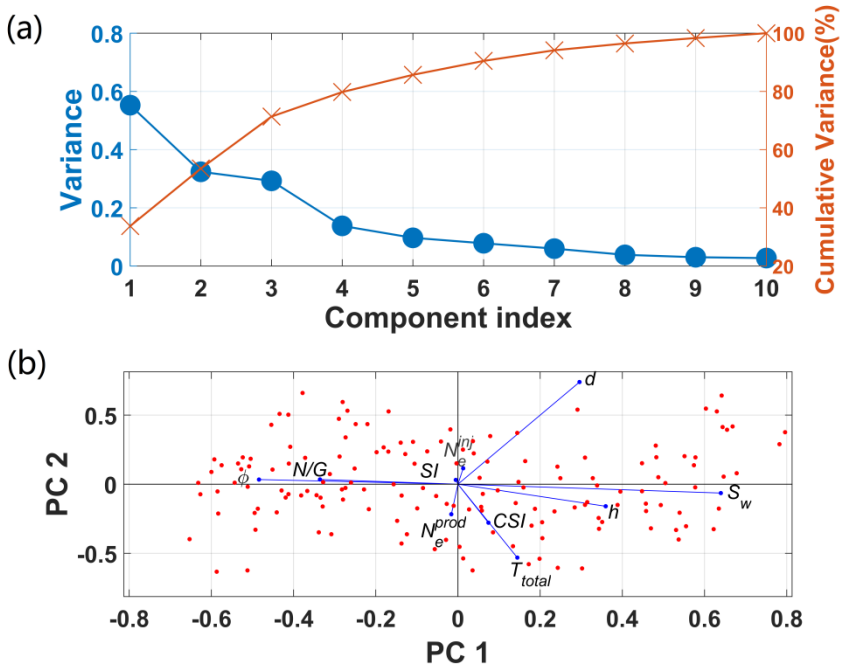


Fig. 4-6 Principal component analysis: (a) – variance plot; (b) – bi-plot: visualization of the orthonormal principal component coefficients for each variable with respect to the first two principal components. The 153 data samples are denoted by red dots.

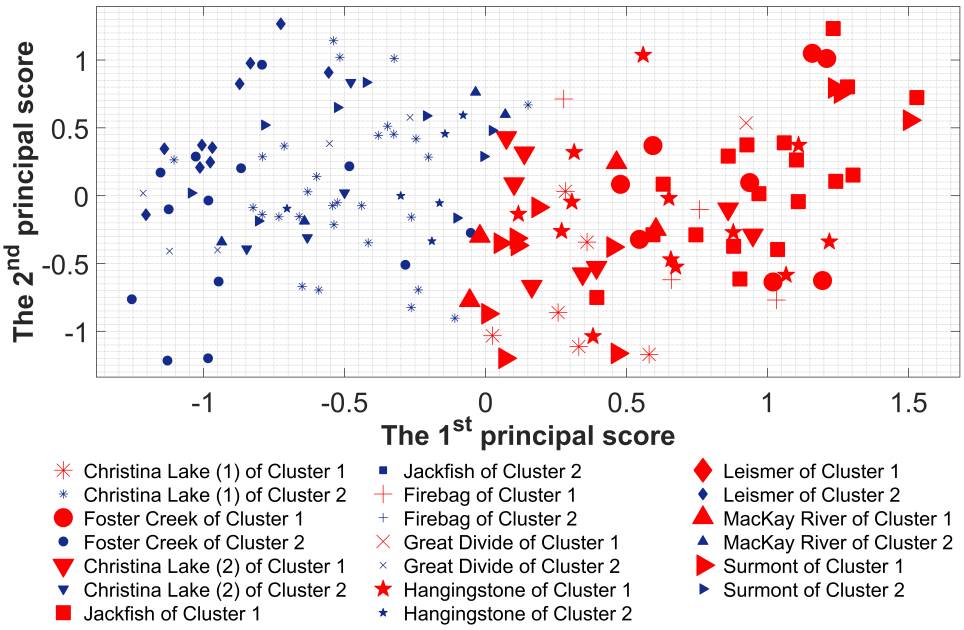


Fig. 4-7 Scatter plot between the first principal score (*PS* 1) and the second principal score (*PS* 2) for all 10 producing fields: small blue marker – cluster 1; large red marker – cluster 2.

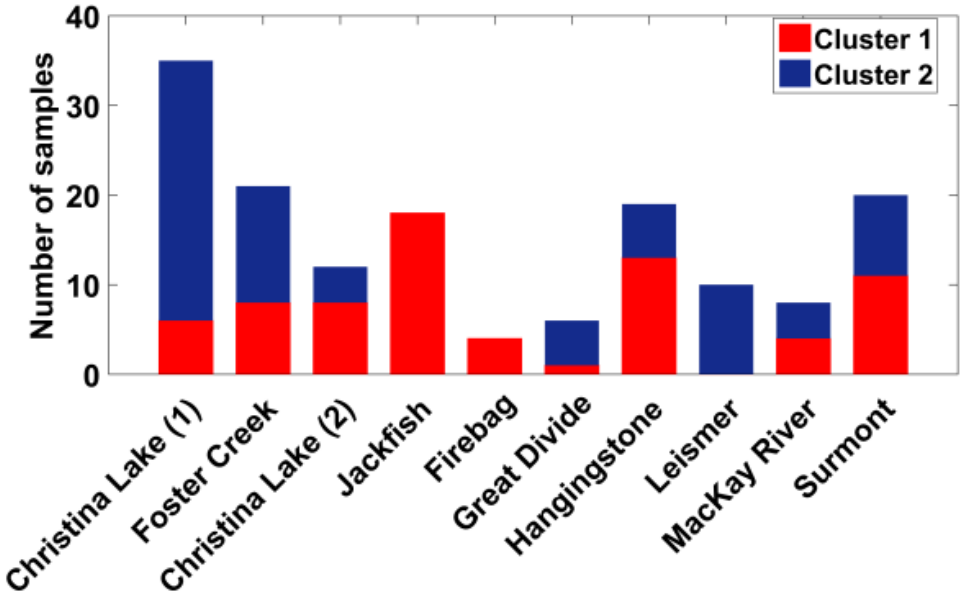


Fig. 4-8 Clustering analysis results of the ten producing fields.

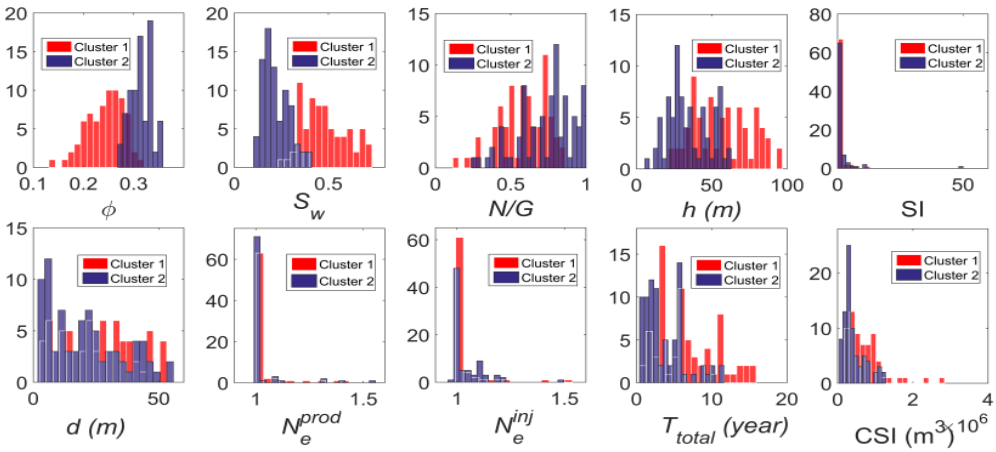


Fig. 4-9 Comparison of histograms of the 10 original input variables from cluster 1 and cluster 2: red – cluster 1; blue – cluster 2.

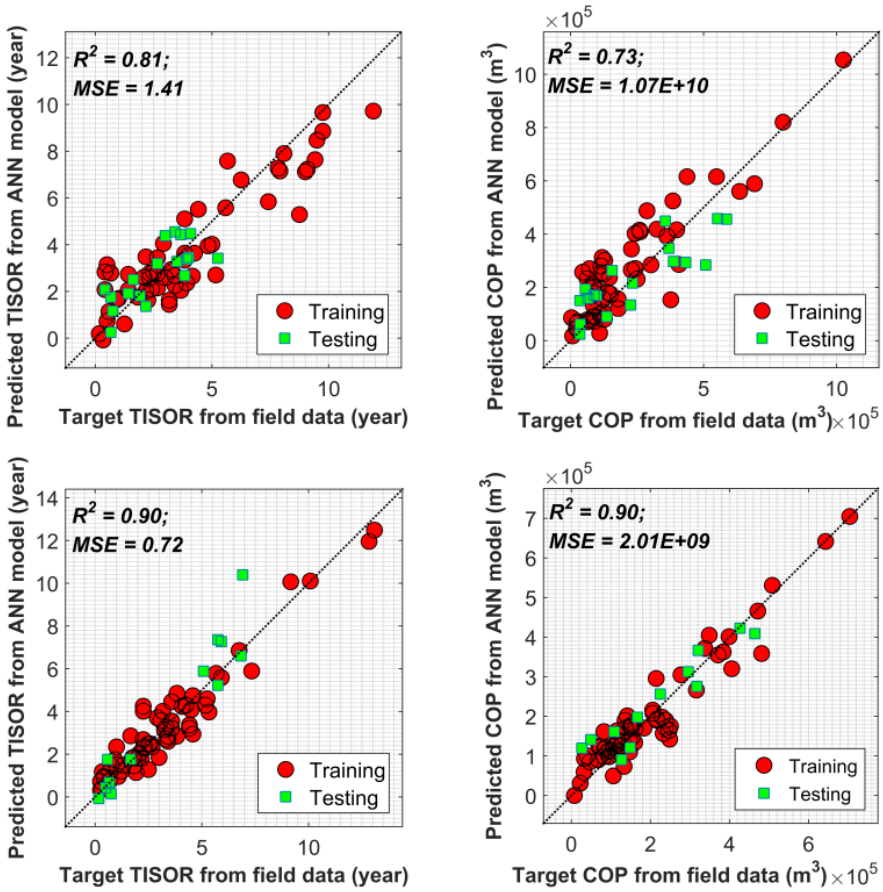


Fig. 4-10 Cross-plots of TISOR and COP from ANN prediction and target TISOR and COP from field data by using manually grouping based on the input parameter *d*: group 1 – data samples with *d* larger than the median (top); group 2 – data samples with *d* less than the median (bottom).

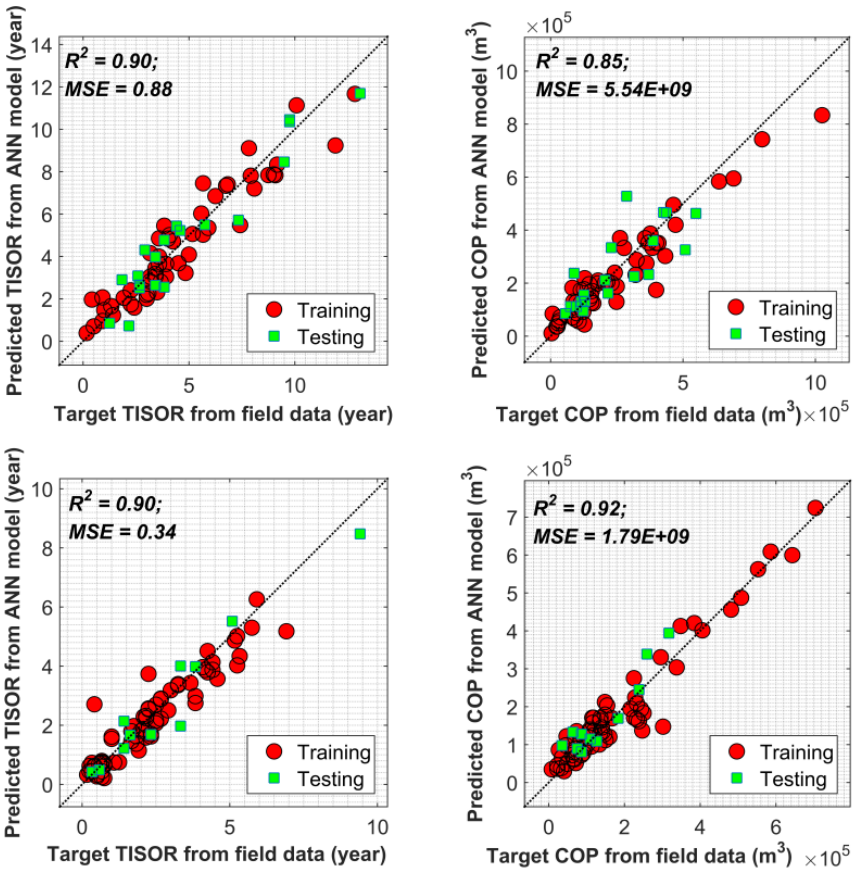


Fig. 4-11 Cross-plots of TISOR and COP from ANN prediction and target TISOR and COP from field data following k-mean clustering analysis: top – cluster 1; bottom – cluster 2.

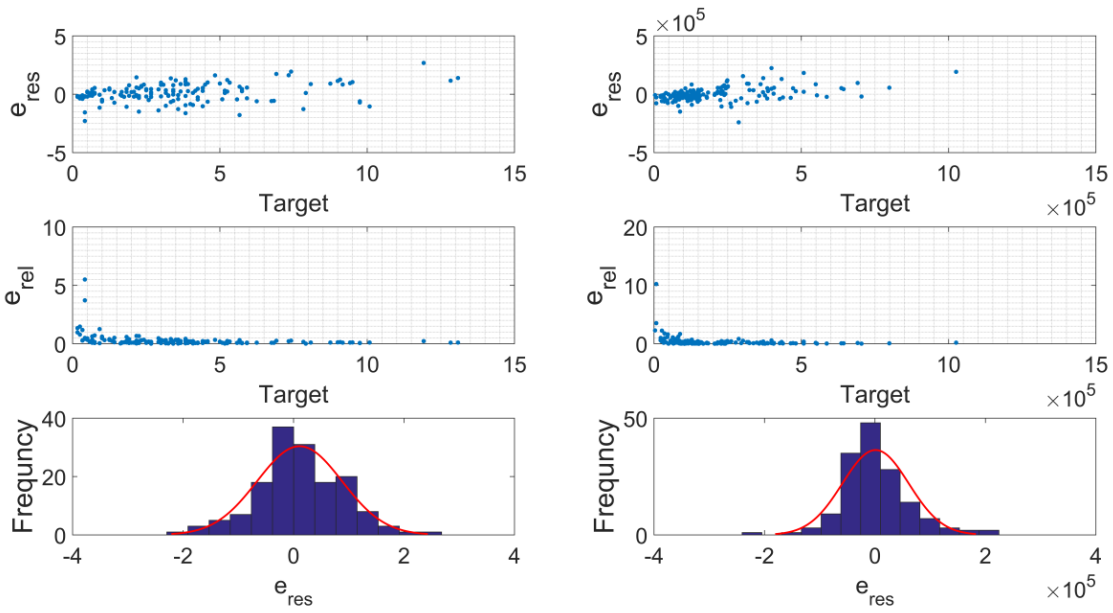


Fig. 4-12 Residual error analysis for the two outputs: TISOR (left) and COP (right). Top, middle, and bottom row represent the residual error, relative error, and distribution of corresponding output parameter, respectively.

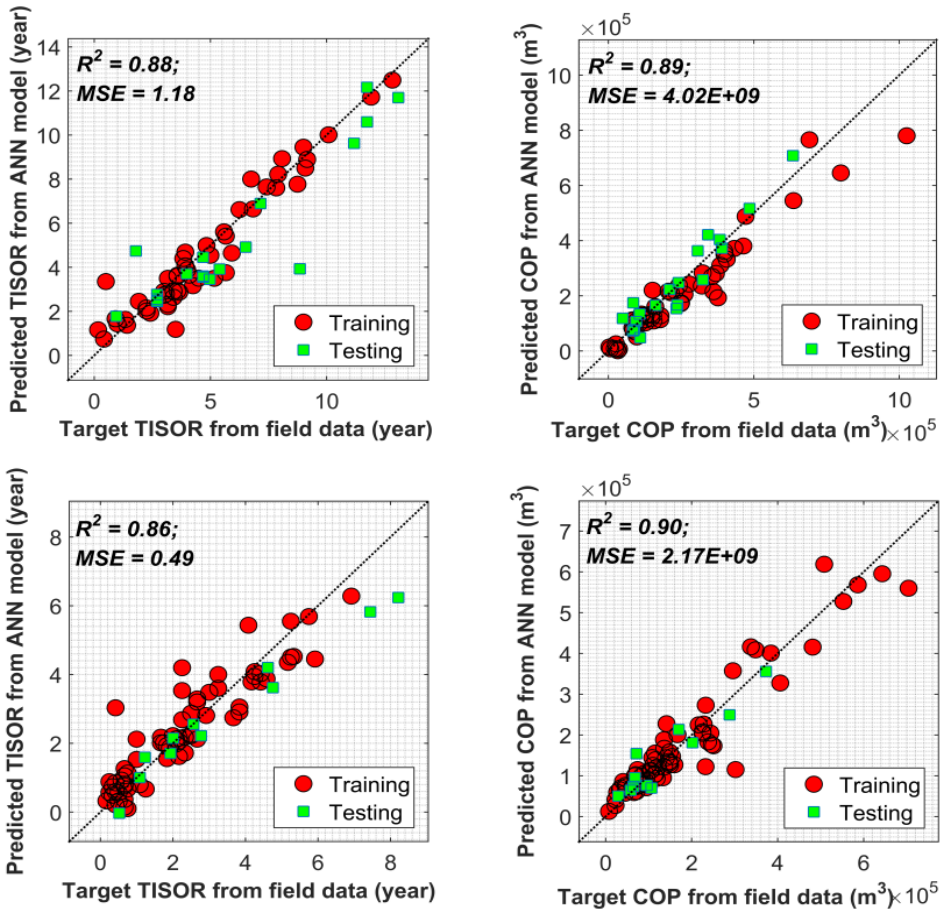


Fig. 4-13 Cross-plots of TISOR and COP from ANN prediction and target TISOR and COP from field data without PCA: top – cluster 1; bottom – cluster 2.

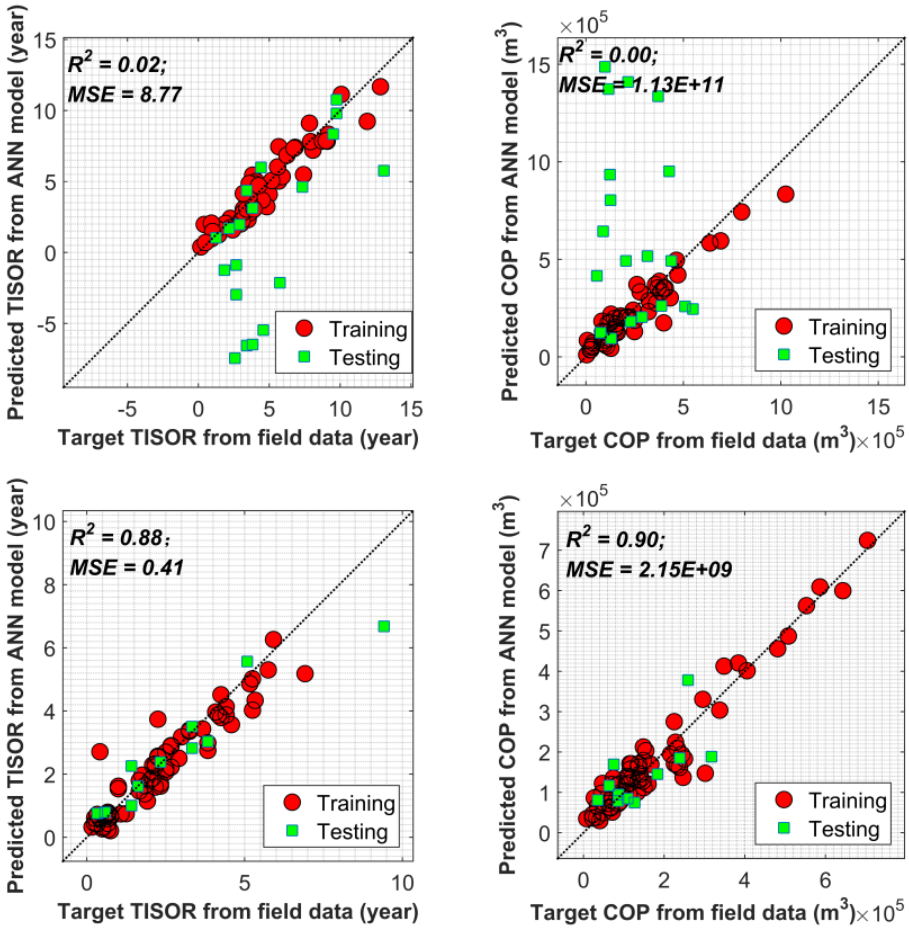


Fig. 4-14 Cross-plots of TISOR and COP from ANN prediction and target TISOR and COP from field data: switching the testing datasets between two clusters while keeping the training datasets unchanged: top – cluster 1; bottom – cluster 2.

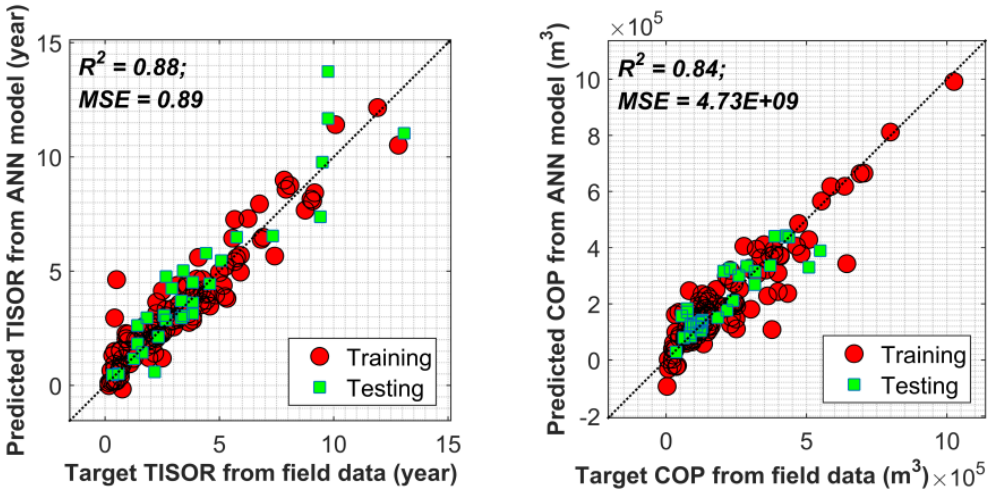


Fig. 4-15 Cross-plots of TISOR and COP from ANN prediction and target TISOR and COP from field data: without clustering analysis.

Chapter 5 Integration of Artificial Intelligence and Production Data Analysis for Shale Heterogeneity Characterization in SAGD Reservoirs[3]

Chapter Overview

SAGD operation is strongly impacted by distributions of heterogeneous shale barriers, which impede the vertical growth and lateral spread of a steam chamber and potentially reduce oil production. This study proposes a workflow integrating artificial intelligence (AI) in a model selection framework that aims to identify associated shale heterogeneities in SAGD reservoir based on extracted features from production time-series data.

Similar to chapter 3 and 4, where the forward correlation between reservoir characteristics and SAGD production performance has been constructed using field data, it is expected that shale heterogeneities can be characterized from field data in this chapter. However, due to some practical challenges associated with actual field data, such as missing important properties (e.g., bottom-hole pressure) or noise, it is difficult to infer shale heterogeneities directly from field data. To deal with this issue, a series of synthetic SAGD models based on typical Athabasca oil reservoir properties and operating conditions is constructed to test the proposed methodology with clean data. This workflow has a great potential to be extended to field data in the future.

Shale heterogeneities are modeled stochastically by sampling the location, lateral extent, and thickness from probabilistic distributions inferred from field data. Several types of input

[3] A version of this chapter has been published as:
Ma, Z., Leung, J. Y., & Zanon, S. (2018). Integration of artificial intelligence and production data analysis for shale heterogeneity characterization in steam-assisted gravity-drainage reservoirs, *Journal of Petroleum Science and Engineering*, *163*, 139–155.

feature extraction methods are introduced in this chapter: piecewise linear approximation, cubic spline interpolation, and discrete wavelet transform (DWT). ANN is constructed to calibrate a relationship between the retrieved production pattern parameters (inputs) and the corresponding geologic parameters describing shale heterogeneities (outputs). The final model is implemented in a novel characterization workflow to infer shale heterogeneities from production profiles. A number of realistic applications are presented to illustrate its utility.

This chapter presents a preliminary attempt in correlating stochastic shale parameters with observable features in production time-series data using AI techniques. The proposed method facilitates the selection of an ensemble of reservoir models that are consistent with the production history; these models can be subjected to further history-matching for a precise final match. The proposed methodology does not intend to replace traditional simulation and history-matching workflows, but it rather offers a complementary tool for extracting additional information from field data and incorporating AI-based models into practical reservoir modeling workflows.

5.1 Introduction

Steam-assisted gravity drainage (SAGD) is one of the proven thermal recovery techniques for bitumen production in Canada. Since its inception (Butler et al., 1981), the SAGD process has been widely employed for commercial heavy oil production. A pair of parallel horizontal wells, consisting of an injector and a producer that is located a few meters beneath the former, are placed near the bottom of the target formation. Steam with high temperature and pressure is injected continuously to form a steam chamber. As the temperature of the bitumen increases, its viscosity decreases. The heated crude oil would drain along the edge of the steam chamber to the producer due to gravitational force.

In a typical SAGD operation, the steam chamber would rise vertically and expand laterally away from the horizontal wellbore in a homogeneous reservoir. A primary challenge with SAGD process is that steam chamber development is highly sensitive to the underlying reservoir heterogeneities. Shale barriers with ultra-low permeability and high water saturation would often impede the steam chamber development, hindering proper contact between the injected steam and in-situ bitumen and posing negative impacts on the ensuing oil production and steam

efficiency (Le Ravalec et al., 2009). As discussed in Li et al., (2011), if a laterally-extensive shale barrier is located immediately above the injector, the vertical rise of a steam chamber is likely to halt at the bottom of this shale barrier. In most cases, it would be difficult for the steam to advance around the edge this shale barrier; the bitumen that is located above the shale barrier would, as a result, be bypassed and could be heated via conduction only. Interestingly, if a certain shale barrier is located further away from the injector, the steam chamber could possibly detour around its edge: after the initial vertical rise, the steam may spread laterally within the highly-permeable sand; as the heated oil at the edge of a shale barrier drains, a wider flow path may emerge, enabling the steam to advance around the barrier and contact the adjacent bitumen. This observation was also corroborated by the simulation work of Ito et al. (2001). Therefore, in theory, if a given shale barrier is limited in lateral extent and relatively discontinuous, its impact on the steam chamber development is minimized. Wang and Leung (2015) illustrated that, although shale barriers may impede the expansion of a steam chamber and give rise to an increased volume of bypassed oil in the steam chamber, their impacts could be subdued if the shale continuity is low. However, if the shale barriers are located in between the well pair, numerous studies have confirmed that even small and discourteous shale barriers could be highly detrimental (Le Ravalec et al., 2009).

Previous works have presented qualitative and quantitative studies regarding the impacts of shale barriers on SAGD production performance. Yang and Butler (1992) constructed a number of experimental models to simulate the effect of reservoir heterogeneities resulting from thin shale layers and reservoir layers of different permeability. Their results demonstrated that the production rate would depend on the locations of individual shale layers with respect to the horizontal well pair. The effects of shale barriers were also investigated in various numerical simulation studies. Chen et al. (2008) presented a stochastic model of shale distribution. They observed that the drainage and flow of hot fluid in the near-well region is highly sensitive to shale distribution, while the expansion of steam chamber away from the well pair is compromised by the presence of long, continuous shale or high shale proportion. Similar observations can also be found in Dang et al. (2013). Amirian et al. (2015) demonstrated that as the distance between a shale barrier and the horizontal well pair decreases, or as the volume or continuity of a shale barrier increases, SAGD recovery efficiency would decrease.

Characterization of shale distributions in heterogeneous SAGD reservoirs, which can be regarded as an "inverse" problem, remains challenging. In a typical reservoir modeling workflow, prior (initial) models of reservoir properties are constructed based on scarce direct measurements of static variables such as well logs or cores and indirect data such as seismic interpretation. The corresponding dynamic well variables, such as flowing pressures, oil rate, and steam injection rate, obtained from flow simulation often differ from actual observations at wells. Most conventional history-matching methods for integration of dynamic data in reservoir model updating (i.e., history-matching) can be classified as: manual adjustment (Williams et al., 1998), gradient-based approach (Zhang et al., 2003), Markov chain Monte Carlo (Pyrcz and Deutsch, 2014), stochastic probability perturbation (Caers and Hoffman, 2006), evolutionary method (Romero and Carter, 2001) and ensemble filtering (Gu and Oliver, 2005). The objective of history-matching is to condition these prior models to the observed dynamic data. With these approaches, the prior reservoir models are perturbed until the mismatch between the simulated production data and the actual production data is minimized below a certain tolerance.

Many of these techniques have been applied in SAGD reservoir characterization in recent years. In Jia et al. (2009), horizontal permeability, porosity and initial oil saturation in a 2D synthetic homogenous reservoir were estimated from saturation and temperature measurements, as well as production and steam-to-oil ratio (SOR) data. Mirzabozorg et al. (2013) performed a similar study, where a single shale layer located at a fixed depth was modeled. In their previous study, Mirzabozorg et al. (2012) argued that saturation and temperature data, in addition to production data, are needed for accurate future predictions. Hiebert et al. (2013) incorporated the observed shape and location of steam chamber during history-matching of a heterogeneous SAGD reservoir. A comprehensive history-matching routine consisting of geomodel construction and flow simulation for the Long Lake SAGD project were presented in Zhang et al. (2014) and Feizabadi et al. (2014). Although precise production matches can be obtained from traditional history-matching routines, these approaches are usually time-consuming (Mirzabozorg et al., 2013; Zhang et al., 2014). High computational demand as a result of the complex process physics deters detailed assisted history-matching at the field scale in a practical fashion. In addition, many assumptions related to the process physics and operating conditions must be invoked and assigned in the numerical simulation model. Non-linearity of the forward model also renders the inverse problem to be ill-posed with non-unique solutions. Consequently, integration of AI-based

approaches provides an attractive avenue to improve the current modeling workflow. In this work, AI techniques are integrated into a novel model-selection workflow that aims to identify the associated characteristics of shale distributions from observed production time-series data.

AI is a broad academic field of study focusing on the formulation of mathematical models that mimics the human neuronal structure and thinking (Haykin, 2008). Numerous applications of AI techniques can be found in natural language processing, automatic programming, robotics and intelligent data retrieval systems (Nilsson, 2014). One widely-adopted AI technique is the artificial neural network (ANN), which is employed as the main modeling method in this work. Since ANN was first proposed by McCulloch and Pitts (1943), it has achieved high popularity for tasks related to prediction and pattern recognition. For a given dataset consisting of a collection of data records (samples), where each of which is a vector comprised of both input and output attributes, ANN can identify and approximate the non-linear, complex and uncertain relationships that exist between its input and output variables. ANN has been adopted in reservoir characterization, production forecast, history-matching, production operation optimization and well design for many years (Al-Fattah and Startzman, 2001; An and Moon, 1993; Awoleke and Lane, 2011; Ayala H and Ertekin, 2007; Ramgulam, 2006; Stundner and Al-Thuwaini, 2001). ANN was also employed as a proxy model for SAGD production performance prediction in heterogeneous reservoirs by Ma et al. (2015; 2017). These previous works have illustrated the capability and versatility of AI approaches to engineering problems; however, direct application of available techniques to a given problem is not trivial. First, domain knowledge is needed to customize the techniques for the unique problem settings. Second, model predictability can be improved by considering and incorporating elements of the underlying physics. Therefore, despite AI techniques have been widely adopted in other areas of reservoir engineering, its immediate application in a history-matching framework for the characterization of shale barriers in SAGD reservoirs is rare. A number of major questions remain: "what schemes are appropriate for parameterizing the high-dimensional input/output vectors of shale barrier configuration and production time series?" or "how to construct a hierarchy of models for the stated purpose?"

As a result, the primary objectives and novelties of this work are (1) proposing a novel model-selection approach for characterization of shale barriers and (2) demonstrating the potential of integrating AI-based techniques in a practical production history-matching for shale

heterogeneities characterization in SAGD reservoirs. A comprehensive data set consisting of flow simulation results based on a wide range of heterogeneity configurations is assembled. Features are extracted from the production profiles as input variables, while the deterministic shale parameters that describe the location and geometry/size of a particular shale barrier are regarded as the output variables. Instead of capturing the entire heterogeneity distribution in a high-dimensional output vector and compromising the accuracy/predictability of the already ill-posed problem, a novel parameterization scheme is formulated to represent the model parameter space with a reduced dimension. A number of hybrid feature identification procedures involving piecewise linear approximation, cubic spline interpolation, and discrete wavelet transform (DWT) are formulated and examined, and their compatibility with different time-series data is tested. It is observed that DWT is often more robust in extracting features when the influence of an individual shale (often in the form of a decline pattern) cannot be readily detected with the other schemes. In addition to feature extraction, these techniques also serve to reduce the dimensionality of the time-series data for subsequent AI-based modeling. The final model is implemented in a novel characterization workflow to infer shale heterogeneities from production profiles. Unlike other history-matching studies, the proposed approach directly takes into account observable production patterns due to shale barriers and direct approximation of entire production data. It presents a preliminary attempt in correlating shale parameters with observable production patterns and wavelet coefficients using AI-based models. A number of realistic history-matching and production forecast applications are presented to illustrate its functionality; in particular, its utility for uncertainty quantification is highlighted.

A significant contribution is that, once the model is calibrated, it offers a computationally-efficient mechanism for selecting an ensemble of reservoir models, which closely honor the actual historical data. This ensemble of models can be further subjected to a robust assisted history-matching scheme to obtain a detailed final match if so desired. The workflow is formulated in a fashion that allows the models to be updated readily once new information becomes available, presenting important potential to facilitate efficient real-time reservoir management tasks. In addition, conventional reservoir modeling and flow simulation workflows are often burdened with extensive data requirement, the AI models presented in this work can analyze the relationships between the production time-series data and the corresponding shale heterogeneities description directly.

The chapter is organized as follows: in section 5.2, details of the model setup, feature extraction, data assembly and ANN modeling are explained; results of the ANN modeling and characterization workflow are presented and discussed in section 5.3; finally, major findings and conclusions are summarized in section 5.4.

5.2 Methods

5.2.1 Construction of Synthetic Datasets

A series of 2D models based on typical Athabasca oil reservoir properties and operating conditions are constructed. The homogeneous (base) model consisting of only oil sand and the heterogeneous model consisting of both oil sand and shale barriers are created as illustrated in chapter 3. In this chapter, the vertical and horizontal permeability is set as 3×10^{-8} and 5×10^{-8} Darcy, respectively, which are much smaller than that of clean oil sand specified in **Table 5-1**. A heterogeneous case is presented in **Fig. 5-1**.

From the constructed synthetic SAGD models, a detailed sensitivity study is conducted to parameterize unique patterns observable in the production response that are related to shale characteristics: a "decline" in the production rate is observed whenever the steam chamber encounters a shale barrier; this decline continues until the steam chamber has advanced beyond the shale barrier, and the production rate would rise again. In general, shale barriers located close to the well pair have much more pronounced impacts on the oil production profiles. Once the steam chamber has reached the top of the pay zone, the production rate starts to decline. Shale barriers that are located far away and encountered thereafter do not exact a noticeable production pattern.

These models are categorized into two datasets: those with a single shale barrier and others with multiple shale barriers. Results from the sensitivity analysis confirmed that in the single shale barrier case, the impacts of individual shale barrier on the production profiles can be detected more easily. These cases are used to assess and compare the performance of numerous feature extraction techniques, with respect to their ease of parameterization and ability to capture various salient features of the production patterns. This analysis helps to formulate an appropriate parameterization scheme for the more realistic cases with multiple shale barriers that may exist in the domain.

5.2.2 Parameterization of Shale Barrier Characteristics

Three parameters are formulated to uniquely describe the physical characteristics of a given shale barrier by considering its location, size, and geometry: the shortest distance between the shale barrier and the injector (D), the shortest horizontal distance between the shale barrier and the left model boundary (H), and the effective shale length (L). For the purpose of normalization, each of the three parameters is formulated as the following dimensionless variables:

$$D_d = \frac{D}{D_{\max}}; H_d = \frac{H}{0.5S}; L_d = \frac{L}{0.5S} \quad \text{..... (5-1)}$$

where each parameter is normalized against its maximum value: D_{max} is the maximum possible value for D (i.e., the distance from the injector to the top-right corner of the 2D model); the upper-bound values for H and L are set to be $0.5S$, which is ½ the spacing (S) between well pairs. This parameterization scheme is illustrated in **Fig. 5-2**.

The same parameterization scheme also applied for multiple shale barriers. The total number of parameters is now $N \times 3$, where N is the number of shale barriers. It is noted that if more than one shale barrier is present, the order of these parameters in the data record is important. A recommendation is to sort the shale barriers in the sequence of steam arrival (based on the change in temperature) and to arrange its parameters accordingly. A preprocessing step is implemented to facilitate this sorting. In the end, the three parameters (D_d, H_d, and L_d) for all N shale barriers are considered as output attributes for the AI-based model next. The thickness of each shale barrier is assumed to be 1 m, representing the maximum resolution of heterogeneity description; thicker shale barriers can be modeled by stacking multiple shale barriers. In this work, the maximum number of shale barriers is set as three. Therefore, the entire dataset is divided into three subsets, for which three different AI-based models are constructed.

It should be noted that, in addition to **Eq. 5-1**, other parameterization variables (e.g., the incident angle between the shale barrier and the injector) were tested. However, the corresponding prediction performance was not quite satisfactory; for instance, the non-unique mapping renders the inverse construction of shale barrier configurations from a given set of parameterized variables challenging.

5.2.3 Feature Extraction from Production Time-Series Data

The oil production profile corresponding to a single shale barrier model is presented in **Fig. 5-3**. Certain features are detected: (1) oil rate declines after the steam chamber has encountered a shale barrier at t_l; (2) this decline continues until the steam chamber has advanced beyond the shale barrier at t_r; (3) the oil rate subsequently increases thereafter. Also shown in **Fig. 5-3** is t_b, where a local minimum is observed. To illustrate the steam chamber advancement around a shale barrier, the distributions of oil saturation, temperature, and steam chamber location at t_l = 240 days, t_b = 1230 days and t_r = 1530 days are compared in **Fig. 5-4**. Sensitivity analysis using the suite of realizations constructed in section 2.1 confirms that the decline pattern, as characterized by t_l, t_b, and t_r, is highly sensitive to d, H, and L.

Three feature points can be retrieved directly from the production profile: (t_l, q_l, Q_l), (t_b, q_b, Q_b), and (t_r, q_r, Q_r), where t = cumulative production time, q = instantaneous oil rate and Q = cumulative oil production; as a result, a total of 9 parameters were used to represent a certain decline pattern, and they were considered as input attributes to the proposed AI-based models. Though the results are generally satisfactory, identifying the decline patterns directly from the production rate profiles has been challenging; in many cases, those feature points cannot be prominently detected (Ma et al., 2016).

In order to facilitate identification of these feature points, a more robust method is proposed here by examining the rate difference profile between a given heterogeneous model and the base model ($\Delta q = q_{\text{heterogeneous}} - q_{\text{base}}$). The detailed sensitivity analysis demonstrates that the Δq profile is capable of reflecting tiny decline patterns that are usually neglected by using the original production rate curves directly. An example of Δq profile is also presented in **Fig. 5-3**. In this study, profiles of Δq, instead of q, are analyzed. In particular, two feature extraction strategies are tested to extract the appropriate input parameters.

Input set #1: 60 decline pattern parameters: An automatic feature extraction workflow is adopted to detect the presence of a decline pattern and retrieve the corresponding feature points.

First, a piecewise linear approximation algorithm (Keogh et al., 2004) is implemented to approximate the Δq profile, as shown in **Fig. 5-3**. Piecewise linear approximation algorithm facilitates segmenting the time-series data into a reduced number of straight lines (linear

functions). The linear segments are used to approximately represent the original data. This step is needed to facilitate the handling of a large quantity of time-series data.

Next, the linear segments are analyzed to identify the presence of any decline feature, which is detected by identifying a local minimum at t_b and the two neighboring endpoints (t_l and t_r), at which Δq is close to zero. Once t_l, t_b, and t_r are determined, the following 9 variables can be readily extracted from the q and Δq profiles: q_l, q_b, q_r, Q_l, Q_b, Q_r, Δq_l, Δq_b, and Δq_r. Therefore, a total of 12 variables are extracted corresponding to every decline pattern.

Although piecewise linear approximation algorithm is useful for capturing the feature points associated with a given decline pattern; however, in order to represent the underlying curvature, the higher-order approximation is needed. To that end, cubic spline interpolation (McKinley and Levine, 1998), which is a piecewise third-order polynomial, is applied next to approximate the decline feature in the Δq profile. Therefore, once the decline pattern is identified, the original data corresponding to that particular decline portion is subdivided into 12 equal intervals, and each interval is approximated using cubic spline interpolation with 4 coefficients. Despite the incurred additional computational cost, the underlying rationale is that capturing the curvature in a decline pattern would enhance the final model accuracy. Details regarding the theory and implementation of cubic spline interpolation can be found elsewhere (McKinley and Levine 1998).

Finally, a total of 48 coefficients are extracted and considered, together with the 12 feature point variables, as input attributes for the AI-based model next. Sensitivity analysis reveals that retaining 48 coefficients would offer reasonable accuracy, without compromising the efficiency of the neural network modeling.

Input set #2: 34 DWT coefficients: Other automatic feature extraction methods that involve the transformation of a given time-series from the time domain into the frequency domain are also adopted. Wavelet transformation decomposes a function (or signal) into the shifted and scaled versions of the basic (mother) wavelet, $\Psi(t)$, which is a wave-shaped function with a zero mean and a limited length (Radunovic, 2009). It is capable of capturing both high- and low-frequency phenomena (Chen et al., 1999; Rioul and Vetterli, 1991). The wavelet function can be expressed as:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \quad \text{.....(5-2)}$$

where s and τ represents the scaling and translation parameter, respectively. The factor $1/\sqrt{s}$ is used to maintain constant energy at different values of scale. Discrete wavelet transform (DWT) provides a feasible and efficient workflow for the decomposition of a time-series data.

In this work, to extract the input features using DWT, the original Δq time-series is first subjected to a low-pass filter and a high-pass filter, where a Daubechies wavelet (db4) is employed. The components deriving from the low-pass filter are the approximation coefficients (cA), while the components deriving from the high-pass filter correspond to the detail coefficients (cD), which are usually regarded as noises and can be discarded.

Next, through a down-sampling procedure, the cA coefficients can be halved, allowing only half of the samples for decomposition in the next level.

Third, this procedure can be repeated for a number of levels, and the iterative DWT process is illustrated in **Fig. 5-5**. The level of decomposition may affect the subsequent data-driven modeling in the next step. A lower level of decomposition (i.e., retaining more coefficients) would result in more input variables, higher degrees of freedom and stronger nonlinearity (possibly overfitting) in the data-driven models. In this work, the original Δq profile is decomposed at level 7 to balance the accuracy of approximation and the modeling efficiency.

Finally, 34 cA coefficients from the final level are retained as input parameters for the AI-based modeling next.

Important steps in the data assembling procedure are summarized in **Fig. 5-3**. As an example, in the case of a single shale barrier, 60 input attributes (12 from piecewise linear approximation plus 48 from cubic spline interpolation) or 34 input attributes (34 DWT coefficients) and 3 output attributes (D_d, H_d, and L_d) are assembled into a single data record.

5.2.4 Artificial Neural Network Modeling

A multilayer perceptron (MLP) network is employed to construct a data-driven model that correlates the input and output attributes in sections 5.2.2-5.2.3. The basic MLP structure contains an input layer, an output layer and any number of hidden layers. The nodes at each network layer are called neurons. Each neuron consists of a bias term and is connected to

neurons of the neighboring layers by weights. Details of the MLP formulation can be found in Haykin (2008). **Fig. 5-6** shows an example of the MLP network configuration with only one hidden layer (10 hidden neurons), 7 input attributes and 3 output attributes. A widely-adopted MLP network, the back-propagation neural network (BPNN), is implemented in this work.

The back-propagation algorithm is a gradient-based supervised learning approach for estimating the unknown network parameters (weights and biases). The function signal transfers from the input layer to the output layer through the hidden layer(s). The error signal, calculated according to the mismatch between network outcomes and targets, transfers from the output layer to the input layer. During the signal transfer process, these weights and biases are updated by minimizing the mismatch using the gradient descent method (Haykin, 2008). The input signal x at a given neuron i is calculated according to **Eq. 5-3**:

$$x_i = B_i + \sum_{j=1}^{m} w_{ij} \cdot y_j \quad \text{.....(5-3)}$$

where x_i is the input signal of neuron i in the current layer; y_j represents the output signal of neuron j in preceding layer; m is number of neurons in the preceding layer; w_{ij} denotes the weight connecting neurons i and j; B_i is the bias term associated with neuron i. The output signal of neuron i is computed by subjecting x_i to a transfer (or activation) function, as shown in **Eq. 5-4**:

$$y_i = f(x_i) \quad \text{.....(5-4)}$$

In this chapter, a commonly-adopted sigmoid function, the hyperbolic tangent sigmoid function, is chosen. The hyperbolic tangent sigmoid function scales between -1 and 1 and is differentiable everywhere. Data normalization is needed to reduce large disparity in scales of different data sources and to alleviate bias in the minimized solution as a result of values with overwhelmingly large magnitude. In this work, all input/output attributes are normalized to vary between -1 and 1 prior to ANN modeling.

The network configuration (i.e., the number of hidden layers and the number of neurons in each hidden layer) should be optimized to achieve a balance between computational efficiency, overfitting and prediction accuracy. An n-fold cross-validation routine is implemented to select the optimal configuration based on the mean squared error (*MSE*) between the network outcomes and targets (Ma et al., 2015). The basis of this procedure is to test a number of possible ANN architectures and select the one that yields the smallest *MSE*. It entails randomly splitting the dataset into *n* subsets with equal size. For each ANN structure candidate (e.g., a single hidden layer with 10 nodes), one subset is designated as the testing dataset, while the remaining (*n-1*) subsets are used for training; this process is repeated *n* times, such that a new testing dataset is selected each time. At the end, an average *MSE* over all *n* folds is computed for this particular candidate. This technique is useful for reducing the potential over-fitting and bias from the training process (Singh & Panda, 2011). Due to the relatively small size of the original dataset, *n* is selected to be 5. It should be noted that experiments with $n = 10$ have been tested, and the resultant network structures are similar to those obtained with $n = 5$. In this study, the 'Neural Network Toolbox^{TM}' (Beale et al., 1992) in MATLAB R2015a is employed to construct all the ANN models. The corresponding computations are performed using a personal computer with an Intel(R) Core (TM) i7-4770 CPU (3.4 GHz) and 12 GB of RAM.

5.2.5 AI-Based Reservoir Characterization Workflow

The ANN models from section 5.2.4 are now incorporated in a history-matching framework to facilitate the integration of production data for shale heterogeneity characterization. The key idea is to select a suite of plausible realizations of shale distribution (location, orientation, geometry, and size), whose production characteristics are consistent with the actual history. The steps are outlined here:

1) If a single distinct decline pattern can be observed from the production history, extract feature points corresponding to the given decline pattern and assemble them into an input vector. Use the calibrated ANN models developed in section 5.2.4 to predict the corresponding output vector for a single shale barrier (D_d, H_d, and L_d).
2) If a single distinct decline pattern cannot be observed from the production history, apply DWT to extract the corresponding cA coefficients and assemble them into an input vector. Use the calibrated ANN models developed in section 2.4 to predict the corresponding

output vector for each of the N cases (e.g., D_d^1, H_d^1, D_d^2, and H_d^2 for two shale barriers case). Given that the exact number of shale barriers is uncertain, estimating different sets of shale configurations from the N ANN models would facilitate the integration of uncertainty in this characterization workflow. An example is illustrated in **Fig. 5-7**, in which the production profiles corresponding to two shale configurations are different. The impact of two smaller shale barriers that are located very close together is essentially indistinguishable from of that of a large shale barrier.

3) Uncertainties due to limited data size and model parameter uncertainty due to training algorithm and initialization are accounted for by estimating a likelihood function for the network predictions and performing parametric bootstrapping of this likelihood to assess the associated uncertainty in the shale distribution parameters. Therefore, multiple sets of shale parameters are sampled from probability distributions, whose means are the network predictions from steps #1-2. Bootstrapping, which is a statistical resampling procedure with replacement for calculation of uncertainty (Secchi et al., 2008), is applied to sample multiple sets of shale parameters; next, the Monte Carlo simulation technique is employed to construct several realizations of shale barrier configuration corresponding to each set of sampled shale parameters.

5.3 Results and Discussions

5.3.1 ANN Modeling for Single Shale Barrier

To examine the various aspects of the proposed methodology in a more controlled setting, a total of 400 heterogeneous reservoir cases with a single rectangular shale barrier are utilized. The decline pattern in the rate (q) profile for many cases is quite small or even invisible; however, if the rate difference (Δq) profile is analyzed, instead of the rate (q) profile alone, a decline pattern is observed in a total of 224 cases [instead of 179 in Ma et al. (2016)] for this dataset. For the remaining models, the shale barrier is either too small and/or located too far away from the horizontal well pair; hence, their production responses are essentially similar to that of the homogeneous model, suggesting that the impacts of the shale barrier can be ignored.

A location map of the shale barriers for all 400 cases is shown in **Fig. 5-8(a)**, where the position of the lower-left corner (or D) of each barrier is presented. There is a clear distinction

between the 224 reservoir models, where the shale barriers are located relatively close to the well pair, and the remaining 176 models. This observation would suggest that beyond a certain maximum distance to the well pair, the impact of any shale barrier on production performance is not evident. However, it is also easy to notice that several cases located close to the well pair don't present an observable decline pattern in oil rate curve. That is because the lengths of shale barriers in these heterogeneous cases are too short (around 5 m). Therefore, this observation indicates that the appearance of a decline pattern is also impacted by the shale length. For instance, a shale barrier cannot cause a detectable rate decline if it is very short, even though it locates closer to the well pair, since the steam chamber can quickly advance it.

In order to examine these 224 models in details, they are subdivided into 3 categories according to L: (1) $L \leq 8$ m, (2) 8 m $< L \leq 11$ m and $L > 11$ m. The location maps corresponding to the three categories are illustrated in **Fig. 5-8(b-d)**. The color scale represents the ratio of cumulative oil productions (Q_r) at t_r of the homogeneous model to the heterogeneous model. This ratio is greater than one, as oil production decreases when shale barriers are present. For all three groups, the impacts of shale heterogeneity on oil production increase dramatically as d decreases; a strong positive correlation between L and a reduction in Q_r can also be detected.

Next, shale barrier with arbitrary thickness and shape is studied. A total of 396 cases heterogeneous reservoir cases with a single shale barrier are modeled and subjected to flow simulation. Among all these cases, a decline pattern is observable in 290 models. 60 input attributes and 3 output attributes for the 290 samples are extracted. The dimensionality of this entire dataset is 290 × 63. A total of 50 records (~17.24%) are randomly selected from the entire dataset and assigned as testing data, while the remaining 240 records are regarded as training data. Sensitivity analysis has revealed that the prediction performance for the output attributes pertinent to the position of a given shale barrier (i.e., D_d and H_d) is superior in comparison to the output attribute representing the size of the shale barrier (i.e., L_d) (Ma et al., 2016). This observation would suggest that the production response is relatively less sensitive to the size of individual shale barrier. It is postulated that given the relatively limited lateral extent of the shale barriers in this synthetic dataset, sensitivity due to L_d would be less apparent, as compared to D_d and H_d. In an attempt to improve the overall prediction performance of all three output attributes, two ANN models, one with two outputs (D_d and H_d) and another one with one output (L_d), are constructed. The input parameters remain unchanged. Considering a relatively small training

dataset, 5 folds are used in the n-fold cross-validation procedure to assess the performance of two network configurations: (1) single hidden layer (with the number of neurons varying from 3 to 20) and (2) two hidden layers (with the number of neurons for each layer varying from 3 to 15).

Two optimized ANN configurations are employed: 8 × 5 for the first model with D_d and H_d as output attributes and 15 × 8 for the second model with L_d as a single output attribute. Network parameters corresponding to the optimum network structure are estimated using the training subset. The training process is repeated for numerous times to avoid being trapped at the local minima. Prediction capability of the trained model is assessed using the testing subset. The ANN prediction performance is shown in **Fig. 5-9**. The 45º line represents a perfect correlation between the network prediction and target values. The overall performance is deemed to be reliable, as most points are located close to the 45º line (R^2 = 1, MSE = 0, where R^2 is the coefficient of determination). This observation is corroborated by the values of R^2 and MSE in each figure. The production profile is more sensitive to the location, instead of the size, of an individual shale barrier; hence, the network prediction of D_d and H_d would be superior than that of L_d (as evidenced by the increased scattering for L_d). To understand this discrepancy, one should consider the advancement of steam chamber around a given shale barrier. Common to all data-driven approaches, model performance depends on the quality of the dataset. In this case, it is related to how precise these decline patterns can be identified. In most cases, the beginning of a decline pattern, which corresponds to the location of a shale barrier, is relatively easy to detect. On the other hand, the definition of the remaining portion of the decline pattern (e.g., duration, endpoint and curvature) is much less precise. These additional features are correlated to L_d.

Finally, parameterization of the Δq profile with DWT is examined. As explained in section 5.2.3, a total of 34 cA coefficients are retrieved and considered as input parameters for the ANN modeling. The results are shown in **Fig. 5-10**. It is interesting to note the significant improvement in the predictions of D_d and H_d, as compared to those resented in **Fig. 5-9**, while similar results for the output variable of L_d is obtained.

It should be mentioned that other types of machine learning algorithms can also be employed. The goal of this paper is to propose a general framework and parameterization scheme to integrate time-series production data for inference of shale barriers. It is certainly possible to integrate other machine learning algorithms in the proposed method; therefore, the random forest technique is also tested. In particular, the random forests (RF) regression

algorithm in scikit-learn Python package (Pedregosa et al., 2011) is employed to train RF models, which use the same inputs as for the ANN models. The results are shown in **Fig. 5-11** and **5-12**. By comparing the estimation presented in **Fig. 5-9** and **Fig. 5-11**, as well as in **Fig. 5-10** and **Fig. 5-**12, one may conclude that both algorithms offer reasonable estimates of the shale parameters. The training performance of RF is slightly better than that of ANN, while ANN is slightly better than RF in terms of testing performance. This comparison would support the conclusion that the overall workflow and parameterization scheme proposed in this paper are quite robust. Given that ANN slightly outperforms RF with the testing dataset, only results of the ANN models would be presented in the remaining sections.

5.3.2 ANN Modeling for Multiple Shale Barriers

Results in the previous section have illustrated the feasibility of the parameterization and feature extraction strategies. However, its direct application to the multiple shale barriers scenario proves challenging. This is because the compounding effect of multiple shale barriers is not a linear sum of individual shale barriers; for example, impact of a shale barrier located far away may be masked by those that are located closer to the well pair; as a result, the number of decline patterns observable in q or Δq is often less than the number of individual shale barriers; this observation is illustrated in **Fig. 5-13**. Another challenge is related to the parameterization of output attributes: given the difficulty in inferring L (as shown in the single shale barrier case), this issue would be further exacerbated when multiple interfering shale barriers are present. Therefore, a few modifications to the parameterization and feature extraction strategies are proposed. In this case, only D_d and H_d are considered as output parameters. A total number of 453 models are constructed: 254 cases consisting two shale barriers (group #1) and 199 cases consisting of three shale barriers (group #2). Similar to the single shale barrier case, a decline pattern is observed in only 443 cases: 245 for group #1 and 198 for group #2. For the remaining models, the shale barriers are either too small and/or located too far away from the horizontal well pair; hence, their production responses are essentially similar to that of the homogeneous model, suggesting that the impacts of the shale barrier can be ignored.

The input parameters are 34 DWT coefficients, while the output parameters are D_d and H_d for each shale barrier: D_d^1, H_d^1, D_d^2, and H_d^2 in group #1 and D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, and H_d^3 in group #2. Separate ANN model corresponding to each group is constructed. The optimal network

architecture is determined as a single hidden layer with 18 nodes (group #1) and two hidden layers with 15 × 9 nodes (group #2). The results are presented in **Fig. 5-14**. The corresponding R^2 value is close to one for most of the training and testing datasets. It is noted that the model performance is better for group #1 with fewer shale barriers. It is expected that, as the number of shale barriers increases, the number of output variables would increase, introducing more degrees of freedom in the ANN model. Given the training dataset for group #2 is actually smaller than that for group #1, it is reasonable to observe a reduction in prediction accuracy corresponding to group #2. In addition, influences from multiple shale barriers tend to overlap, as the number of shale barriers increases; this, in turn, compromises the resolution of individual model parameter from the aggregated production profile and obscures the inference of the respective contribution of each shale barrier.

Two more remarks should be made. First, the prediction of D_d is slightly superior to that of H_d. Second, for both groups, a higher R^2 value is obtained corresponding to parameters of a shale barrier that is first encountered by the steam chamber. This observation corroborates with the trend in **Fig. 5-8**: as the distance to the well pair increases, the impact of any shale barrier on production performance is less evident. The implication is that the impact of a shale barrier that is located further away from the well pair is often masked by other shale barriers that are physically closer to the well pair. A particular example is demonstrated in **Fig. 5-15**; the third shale barrier in the model on the right essentially has no observable impact on the overall production performance.

5.3.3 Production History-Matching

A case study is presented next to illustrate the proposed AI-based reservoir characterization workflow presented in section 2.5. A model consisting of two shale barriers, as shown in **Fig. 5-16**, is selected to be the true case, and the corresponding production profile is shown in **Fig. 5-17**. The three trained ANN models corresponding to single shale barrier (section 5.3.1) and multiple shale barriers (section 5.3.2) are employed to estimate the unknown shale parameters. Uncertainty in each shale barrier parameter (L and D) is quantified using a likelihood function (whose mean is the ANN prediction, while the variance is assumed to be 20% of the mean). Bootstrapping is subsequently performed to sample multiple sets of shale parameters from these distributions. Given the difficulty in inferring the shale barrier length, L is instead sampled

randomly from a uniform probability distribution (minimum = 10 m; maximum = 16 m). In this case study, an ensemble of shale distributions with a varying number of shale barriers is constructed. Ten realizations are randomly selected from this ensemble and shown in **Fig. 5-16**. It is clear that all 10 models are visually consistent with the true case. The production profiles corresponding to this ensemble of characterized models have bracketed the response of the true model, as depicted in **Fig. 5-17**. It demonstrates the capacity of the proposed workflow in identifying a number of reservoir models that are consistent with the production history. These models can be further subjected to a more detailed history-matching scheme to obtain a thorough final match. However, in many cases, the ensemble of reservoir models would have already provided valuable insights regarding the heterogeneity uncertainty. Moreover, this modeling framework is flexible such that the networks can be updated periodically as more production history becomes available.

An obvious limitation of this workflow is that the maximum number of shale barriers is set as 3. It is practically impossible to define the number of shale barriers precisely in real applications. Petrophysical and well data should be consulted to estimate the expected ranges in proportion, size, and number of shale barriers. The workflow can certainly be applied to construct additional ANN models if more shale barriers are considered. However, the results presented thus far have demonstrated that the inverse history-matching problem is inherently ill-posed with non-unique solutions (e.g., **Fig. 5-16**). The aggregated impact of multiple shale barriers near the well pair can generally be replaced by a reduced number of shale barriers with bigger size, while, in other scenarios, shale barriers that are located far away would exact negligible impact on the production. Therefore, it is argued that incorporating a large number of shale barriers is probably unnecessary. An example is shown in **Fig. 5-18**, where the true model (shown in the red solid frame) is consisting of stochastically-distributed shale barriers generated via sequential indicator simulation, as implemented in GSLIB (Deutsch and Journel 1998). The workflow is applied to construct an ensemble of history-matched models (shown in the black dashed frames). It is clear that a reduced number of shale barriers (≤ 3) is sufficient to represent the large number of smaller shale barriers in the true model. This ensemble of models may serve as initial guesses and be subjected to further history-matching for a precise final match. In the end, a sensitivity analysis integrating actual characteristics of shale barriers is recommended

when constructing these ANN models. A balance between model efficiency and estimation accuracy should be considered.

Another interesting application of this workflow is illustrated next. Consider a scenario where the production profile is experiencing an early decline, indicating that the steam chamber has encountered at least one shale barrier. However, the entire decline pattern may not be observed for another few months. Nevertheless, it is often desirable to obtain an early assessment of the potential impact on ultimate recovery (or reserves), even if only a portion of the decline pattern is observed. One approach would be to extrapolate the decline pattern empirically; however, this method suffers two main deficiencies: (1) the extrapolation criteria appear arbitrary or ad-hoc; (2) the impact on the production profile is not tied to the actual shale distribution; in other words, it is not possible to quantify the relative impact on production for different shale barrier configurations. The characterization workflow presented in this work, however, can be adopted to generate a suite of possible realizations of shale distribution and the corresponding production forecasts.

An example is presented in **Fig. 5-19**. In (**a**), the q profiles corresponding to a single shale barrier (true case) and a base case with homogeneous properties (i.e., no shale barrier) are shown. If the base case is considered for the original forecast, a reduction in ultimate recovery should be expected once a portion of the decline pattern is detected. As explained in section 2.3, a given decline pattern can be parameterized by 3 feature points: (t_l and Δq_l), (t_b and Δq_b), and (t_r and Δq_r), as presented in **Fig. 19**(**b**). For illustrative purposes, it is assumed that only the first feature point has been detected (**Fig. 19**(**c**)), the remaining decline pattern can be extrapolated by sampling the remaining two feature points: t_b, Δq_b, t_r, and Δq_r. A simple uniform distribution is assigned and 100 sets of possible feature points are sampled, as shown in **Fig. 19**(**d**). It should be noted that an average of zero is assumed for Δq, $t \geq t_r$, in all 100 cases; the justification for this choice is that Δq is > 0 for a period of time immediately after the steam chamber has advanced beyond the shale barrier prior to dropping to < 0 eventually, as evidenced in **Fig. 19**(**b**). The q profiles at a given time (t) corresponding to the 100 possible (estimated) cases, as shown in **Fig. 19**(**e**) can be estimated as:

$$q^t_{\text{heterogeneous}} = \Delta q^t + q^t_{base} \quad \text{.....(5-5)}$$

The cumulative profiles (Q) can be obtained by integrating the rate profile (q) over time numerically. Finally, each profile is subjected to the characterization workflow to infer a corresponding shale barrier configuration. For this particular example, the estimated Q ranges between 4.69×10^5 to 4.93×10^5 m^3, with an average of 4.83×10^5 m^3, representing a 7-13% reduction in ultimate recovery as compared to the base case. Instead of arbitrarily extrapolating the production profile for forecast purposes, this workflow facilitates the construction of probable production profiles that are linked to specific shale barrier configurations.

5.4 Conclusion

This work presents a practical implementation of AI-based models for characterizing shale heterogeneities by correlating heterogeneity parameters with production data. A set of novel parameterization schemes is implemented to represent shale barrier configurations with reduced dimensional vectors, as well as to identify and parameterize particular patterns observable in the production response that are related to shale characteristics. Input feature points are extracted from the time-series production data using the piecewise linear approximation function, cubic spline interpolation, and discrete wavelet transform, and they are employed in ANN modeling to calibrate a relationship between the retrieved production pattern parameters and the corresponding shale heterogeneities. Results of the ANN model are promising and satisfactory. It is observed that, in comparison to the shale length, higher prediction fidelity is obtained with the location parameters. Model accuracy also decreases with the larger number of shale barriers. Generally speaking, slightly superior characterization performance can be observed for shale barriers that are located closer to the well pair.

These calibrated ANN models are integrated into a model-selection workflow to infer shale distribution from actual production history. The proposed characterization workflow is applied in a number of case studies, where the number of shale barriers is unknown. The outcome of this workflow is an ensemble of reservoir models that are consistent with the production history. A major advantage of this workflow is that a set of reservoir models that are consistent with the production history can be identified reliably and quickly. It offers a viable and complementary alternative to the conventional history-matching characterization routines. These models may serve as initial guesses and be subjected to more rigorous history-matching for a precise final

match; however, it is highly probable that they contain sufficient information regarding the shale barrier distribution for most practical decision-making purposes.

Future work will involve integrating other data-mining approaches such as dimensionality-reduction techniques and clustering analysis. Other machine learning algorithms, such as deep neural networks and support vector regression, should also be explored. The sensitivity of the modeling dataset size should be explored to examine issues related to extrapolation, overfitting and enhancing model predictability. Although initial attempts have involved models derived from synthetic data, subsequent efforts would integrate models calibrated from field data.

5.5 Reference

Al-Fattah, S. M., & Startzman, R. A. (2001). Neural network approach predicts U.S. natural gas production, Paper presented at the *SPE Production and Operations Symposium*, Oklahoma City, Oklahoma, US.

Amirian, E., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Integrated cluster analysis and artificial neural network modeling for steam-assisted gravity drainage performance prediction in heterogeneous reservoirs. *Expert Systems with Applications, 42*(2), 723-740.

An, P., & Moon, W. (1993). Reservoir characterization using feedforward neural networks. *In SEG Technical Program Expanded Abstracts 1993*, 258-262.

Awoleke, O., & Lane, R. (2011). Analysis of data from the barnett shale using conventional statistical and virtual intelligence techniques. *SPE Reservoir Evaluation & Engineering, 14*(05), 544-556.

Ayala H, L. F., & Ertekin, T. (2007). Neuro-simulation analysis of pressure maintenance operations in gas condensate reservoirs. *Journal of Petroleum Science and Engineering, 58*(1), 207-226.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (1992). *Neural network toolbox™ user's guide*. The Mathworks Inc.

Butler, R., McNab, G., & Lo, H. (1981). Theoretical studies on the gravity drainage of heavy oil during in‐situ steam heating. *The Canadian Journal of Chemical Engineering, 59*(4), 455-460.

Caers, J., & Hoffman, T. (2006). The probability perturbation method: A new look at bayesian inverse modeling. *Mathematical Geology, 38*(1), 81-100.

Chen, B., Wang, X., Yang, S., & McGreavy, C. (1999). Application of wavelets and neural networks to diagnostic system development, 1, feature extraction. *Computers & Chemical Engineering, 23*(7), 899-906.

Chen, Q., Gerritsen, M. G., & Kovscek, A. R. (2008). Effects of reservoir heterogeneities on the steam-assisted gravity-drainage process. *SPE Reservoir Evaluation & Engineering, 11*(05), 921-932.

CMG, 2015. STARS: Users' Guide, advanced processes & thermal reservoir simulator (Version 2015), Calgary, Alberta, Canada: Computer Modeling Group Ltd.

Dang, T. Q. C., Chen, Z., Nguyen, T. B. N., Bae, W., & Mai, C. L. (2013). Numerical simulation of SAGD recovery process in presence of shale barriers, thief zones, and fracture system. *Petroleum Science and Technology, 31*(14), 1454-1470.

Deutsch, C. V., & Journel, A. G. (1998). *GSLIB geostatistical software library and User's guide (2nd ed.)*. New York: Oxford University Press, Inc.

Feizabadi, S. A., Zhang, X. K., & Yang, P. (2014). An integrated approach to building history-matched geomodels to understand complex long lake oil sands reservoirs, part 2: Simulation. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Canada.

Gu, Y., & Oliver, D. S. (2005). History matching of the PUNQ-S3 reservoir model using the ensemble kalman filter. *SPE Journal, 10*(02), 217-224.

Haykin, S.S. (2008). *Neural networks and learning machines* (3rd ed.), Upper Saddle River, NJ, USA: Pearson.

Hiebert, A. D., Morrish, I. C., Card, C., Ha, H., Porter, S., Kumar, A., . . . , & Close, J. C. (2013). Incorporating 4D seismic steam chamber location information into assisted history

matching for A SAGD simulation. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Canada.

Ito, Y., Hirata, T., & Ichikawa, M. (2001). The growth of the steam chamber during the early period of the UTF phase B and hangingstone phase I projects. *Journal of Canadian Petroleum Technology, 40*(09).

Jia, X., Cunha, L., & Deutsch, C. (2009). Investigation of a stochastic optimization method for automatic history matching of SAGD processes. *Journal of Canadian Petroleum Technology, 48*(01), 14-18.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). Segmenting time series: A survey and novel approach. *Data Mining in Time Series Databases*, 57, 1-22.

Le Ravalec, M., Morlot, C., Marmier, R., & Foulon, D. (2009). Heterogeneity impact on SAGD process performance in mobile heavy oil reservoirs. *Oil & Gas Science and Technology-Revue De l'IFP, 64*(4), 469-476.

Li, W., Mamora, D., Li, Y., & Qiu, F. (2011). Numerical investigation of potential injection strategies to reduce shale barrier impacts on SAGD process. *Journal of Canadian Petroleum Technology, 50*(03), 57-64.

Ma, Z., Leung, J. Y., & Zanon, S. (2016). Integration of artificial intelligence and production data analysis for shale heterogeneity characterization in SAGD reservoirs. Paper presented at the *SPE Canada Heavy Oil Technical Conference*, Calgary, Alberta, Canada.

Ma, Z., Leung, J. Y., & Zanon, S. (2017). Practical data mining and artificial neural network modeling for SAGD production analysis. *Journal of Energy Resources Technology, 139*(3), 032909.

Ma, Z., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Practical implementation of knowledge-based approaches for steam-assisted gravity drainage production analysis. *Expert Systems with Applications, 42*(21), 7326-7343.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

McKinley, S., & Levine, M. (1998). Cubic spline interpolation. *College of the Redwoods, 45*(1), 1049-1060.

Mirzabozorg, A., Nghiem, L., Chen, Z., & Yang, C. (2013). Differential evolution for assisted history matching process: SAGD case study. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Mirzabozorg, A., Nghiem, L., Chen, Z., Yang, C., & Hajizadeh, Y. (2012). History matching saturation and temperature fronts with adjustments of petro-physical properties; SAGD case study. Paper presented at the *SPE Kuwait International Petroleum Conference and Exhibition*, Kuwait City, Kuwait.

Nilsson, N. J. (2014). *Principles of artificial intelligence*, Morgan Kaufmann.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830.

Pyrcz, M. J., & Deutsch, C. V. (2014). Geostatistical reservoir modeling, Oxford University Press.

Radunovic, D. P. (2009). *Wavelets: From math to practice (1st Ed.)*, Springer Publishing Company, Incorporated.

Ramgulam, A. (2006). Utilization of artificial neural networks in the optimization of history matching (Doctoral dissertation, the Pennsylvania State University).

Rioul, O., & Vetterli, M. (1991). Wavelets and signal processing. *IEEE Signal Processing Magazine, 8*(LCAV-ARTICLE-1991-005), 14-38.

Romero, C., & Carter, J. (2001). Using genetic algorithms for reservoir characterization. *Journal of Petroleum Science and Engineering, 31*(2), 113-123.

Secchi, P., Zio, E., & Di Maio, F. (2008). Quantifying uncertainties in the estimation of safety parameters by using bootstrapped artificial neural networks. *Annals of Nuclear Energy, 35*(12), 2338-2350.

Sereda, J. N., & James, B. R. (2014). A case study in the application of bitumen geochemistry for reservoir characterization in SAGD development. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Singh, G., & Panda, R. K. (2011). Daily sediment yield modeling with artificial neural network using 10-fold cross validation method: A small agricultural watershed, kapgari, india. *International Journal of Earth Sciences and Engineering, 4*(6), 443-450.

Stundner, M., & Al-Thuwaini, J. S. (2001). How data-driven modeling methods like neural networks can help to integrate different types of data into reservoir management, Paper presented at the *SPE Middle East Oil Show*, Manama, Bahrain.

Wang, C., & Leung, J. (2015). Characterizing the effects of lean zones and shale distribution in steam-assisted-gravity-drainage recovery performance. *SPE Reservoir Evaluation & Engineering, 18*(03), 329-345.

Williams, M., Keating, J., & Barghouty, M. (1998). The stratigraphic method: A structured approach to history matching complex simulation models. *SPE Reservoir Evaluation & Engineering, 1*(02), 169-176.

Yang, G., & Butler, R. (1992). Effects of reservoir heterogeneities on heavy oil recovery by steam-assisted gravity drainage. *Journal of Canadian Petroleum Technology, 31*(08), 37-43.

Zhang, F., Reynolds, A. C., & Oliver, D. S. (2003). An initial guess for the Levenberg–Marquardt algorithm for conditioning a stochastic channel to pressure data. *Mathematical Geology, 35*(1), 67-88.

Zhang, X. K., Feizabadi, S. A., & Yang, P. (2014). An integrated approach to building history-matched geomoels to understand complex long lake oil sands reservoirs, part 1: Geomodeling. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Table

Table 5-1 Reservoir properties and operating constraints for the base and heterogeneous models.

Reservoir depth (m)	200
Reservoir thickness (m)	32
Reservoir size in X direction (m)	51
Injector depth (m)	225
Producer depth (m)	230
Initial temperature (°C)	15
Initial reservoir pressure (kPa)	1088 @ 230m
Initial oil viscosity (cp)	592000
Oil sand porosity (fraction)	0.32
Oil sand horizontal permeability (D)	2.5
Oil sand vertical permeability (D)	1.5
Shale barrier porosity (fraction)	0.32
Shale barrier horizontal permeability (D)	5×10^{-8}
Shale barrier vertical permeability (D)	3×10^{-8}
Molar fraction of methane (%)	5
Rock compressibility (kPa^{-1})	2.0×10^{-6}
Rock heat capacity (J/m3*°C)	2.35×10^{6}
Thermal conductivity of matrix (J/m*°C)	1.468×10^{5}
Thermal conductivity of oil (J/m*°C)	1.15×10^{4}
Thermal conductivity of gas (J/m*°C)	1.3997×10^{2}
Thermal conductivity of water (J/m*°C)	5.35×10^{4}
Initial oil saturation (fraction)	0.8
Initial water saturation (fraction)	0.2
Number of production wells	1
Number of injection wells	1

Steam temperature (°C)	214.86
Steam quality (%)	95
Injection pressure (kPa)	2100
Preheating period (day)	60
Total production time (day)	3652

Figures

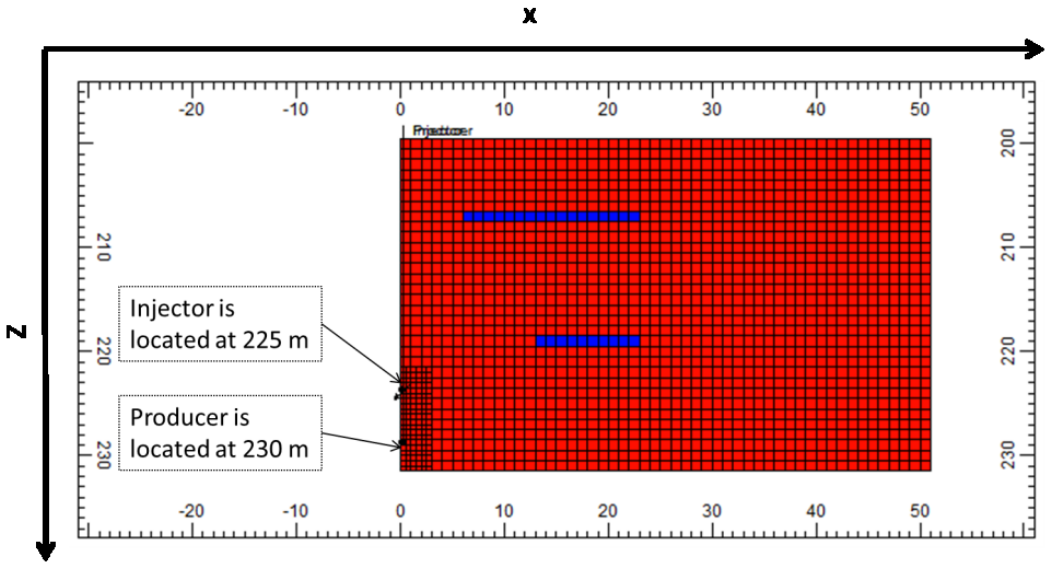


Fig. 5-1 Illustration of the 2D SAGD models (only half of the distance between neighboring well pairs is incorporated): the horizontal production well pair is located at the left.

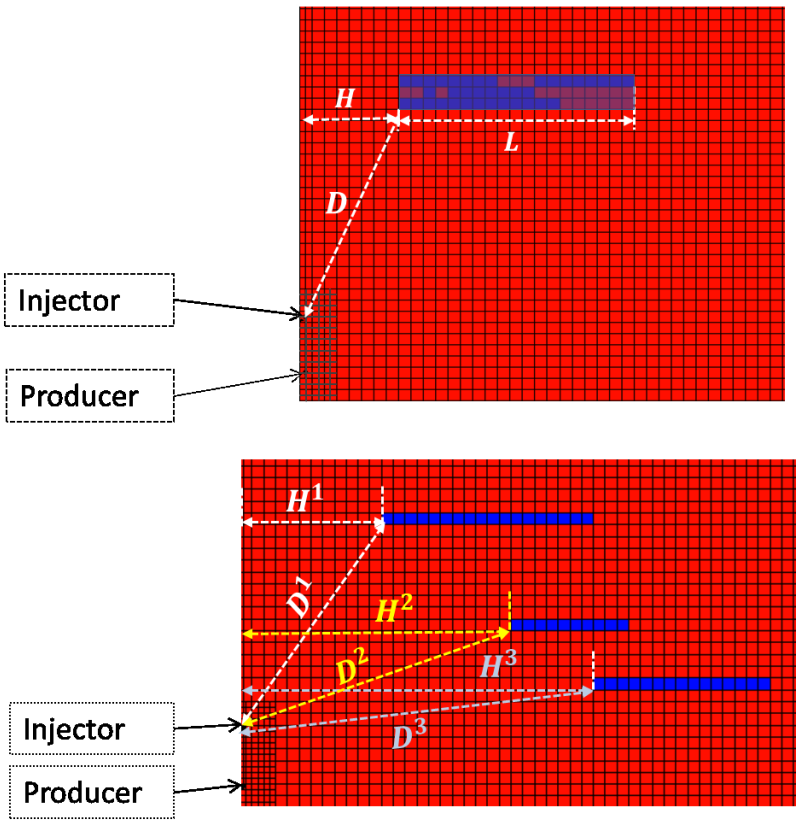


Fig. 5-2 Parameterization of shale barrier(s): top – single shale barrier; bottom – multiple shale barriers.

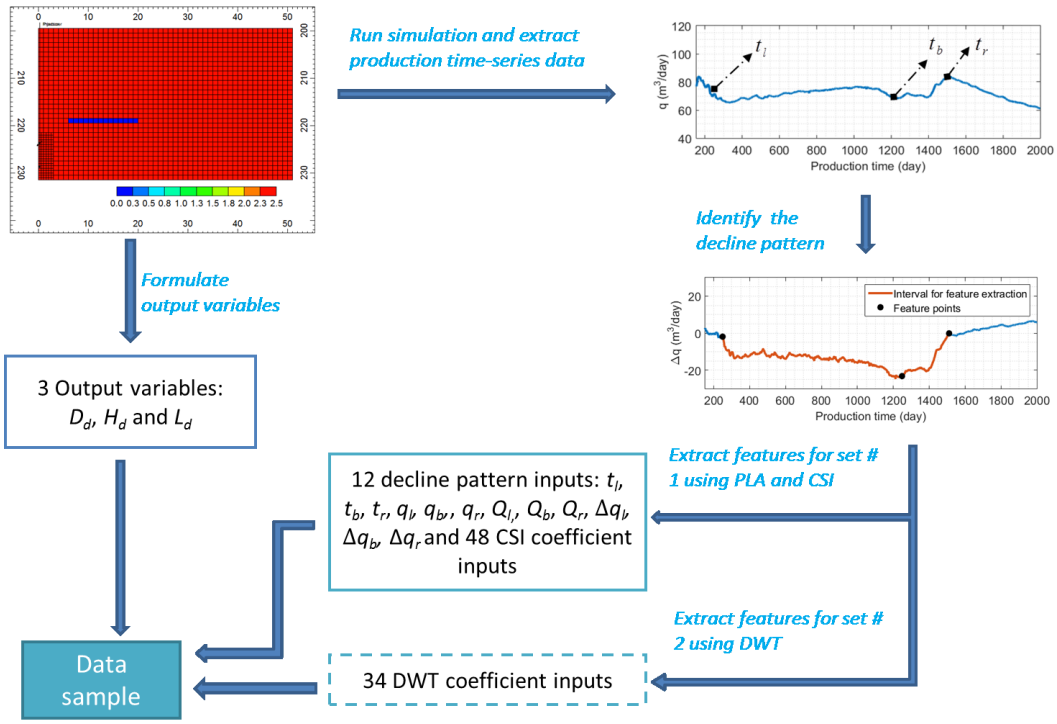


Fig. 5-3 Illustration of the data assembling procedure for single shale barrier case.

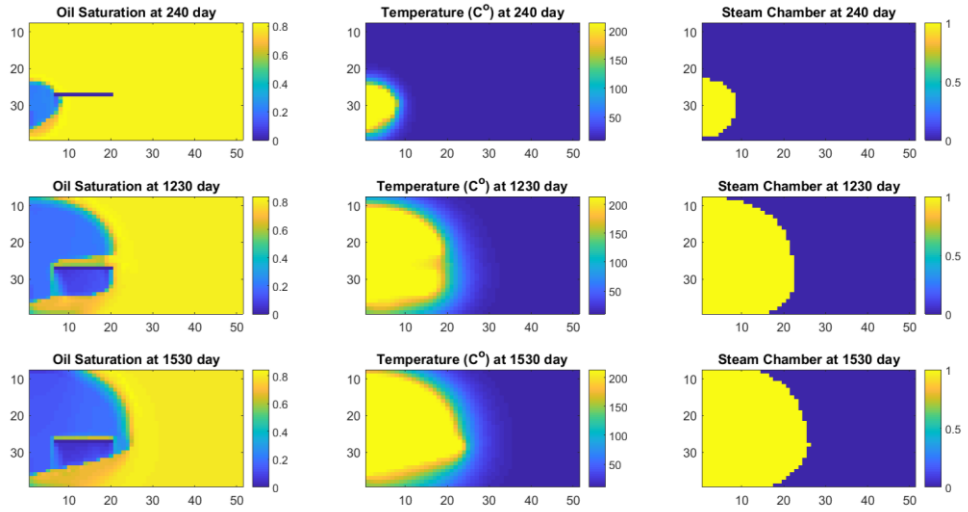


Fig. 5-4 Evolution of oil saturation, temperature (°C), and steam chamber location (where temperature > 80 °C) with time. The unit for the x-axis and y-axis is in m.

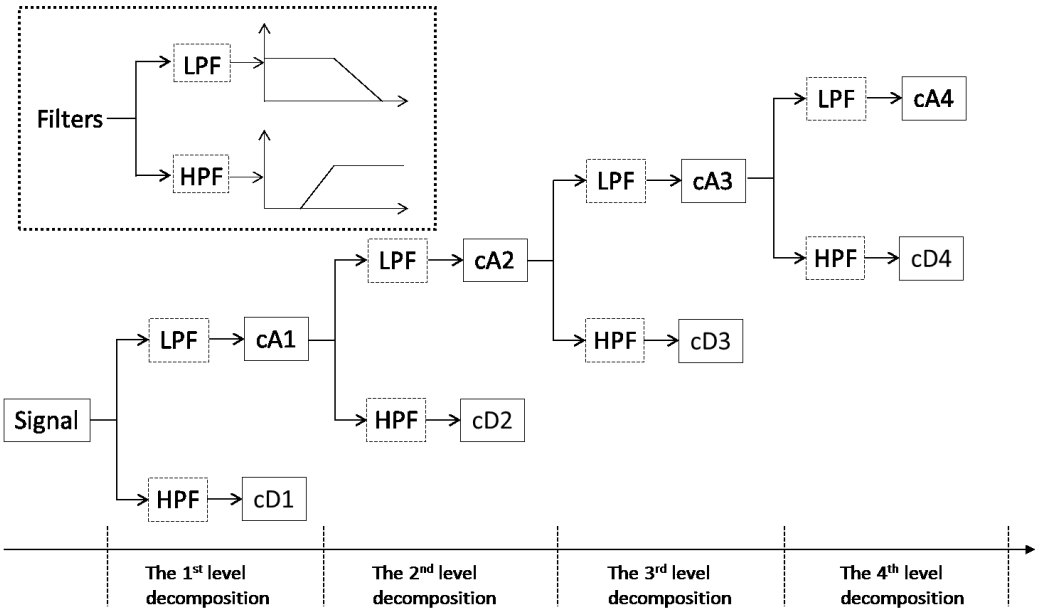


Fig. 5-5 Example of a time-series decomposition (i.e., feature extraction) using DWT involving a 4-level decomposition.

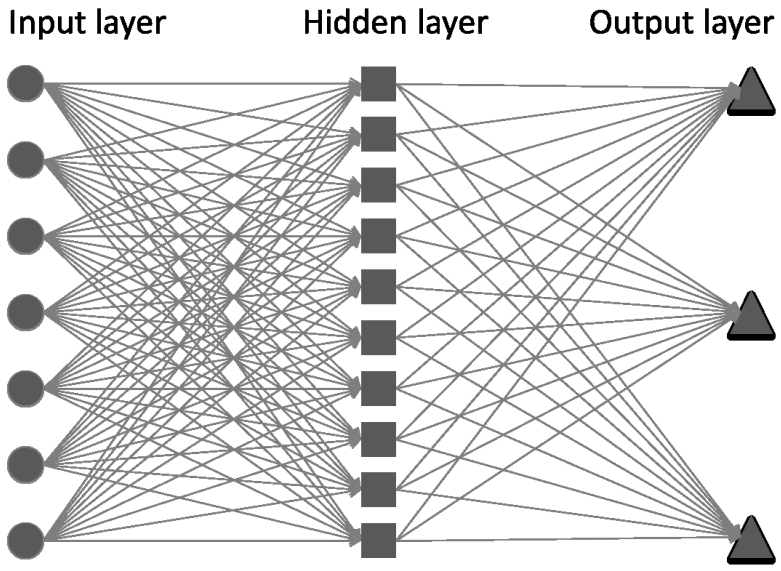


Fig. 5-6 An example of a single hidden layer ANN structure with 7 input variables, 10 hidden neurons, and 3 output variables.

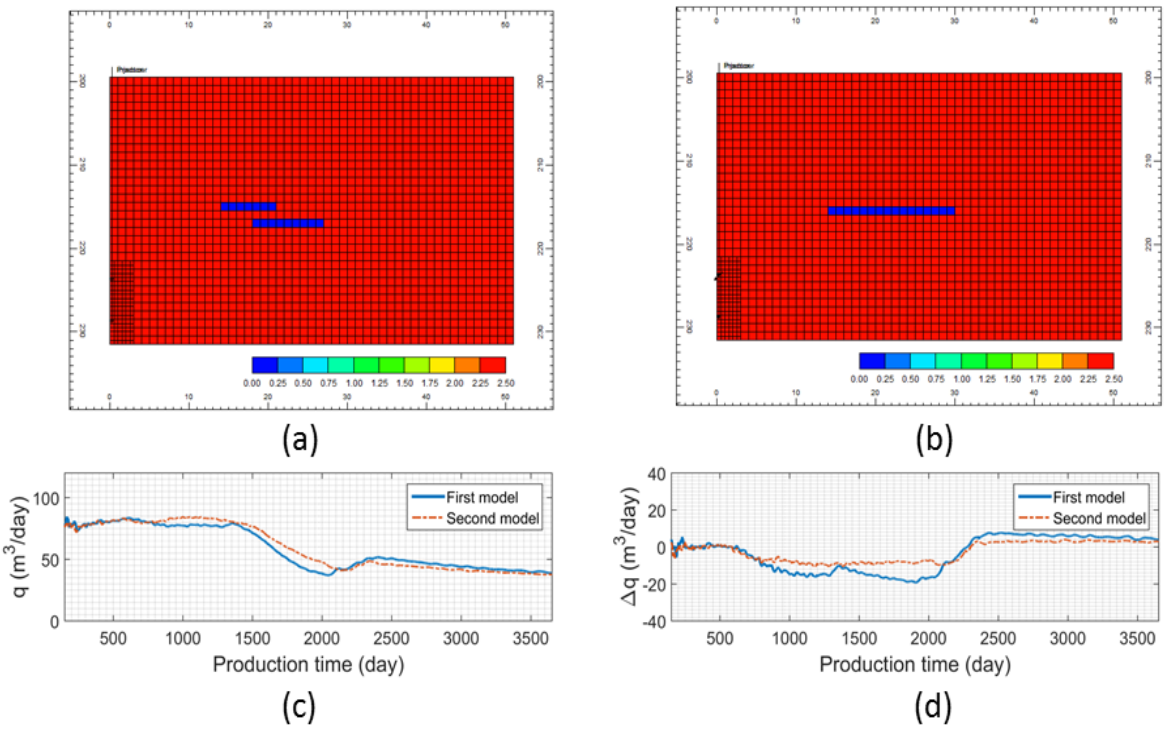


Fig. 5-7 Uncertainty in reservoir characterization: similar production profiles in terms of q and Δq can be obtained from two different reservoir models: (a) – the first model with two shale barriers; (b) – the second reservoir model with single shale barrier.

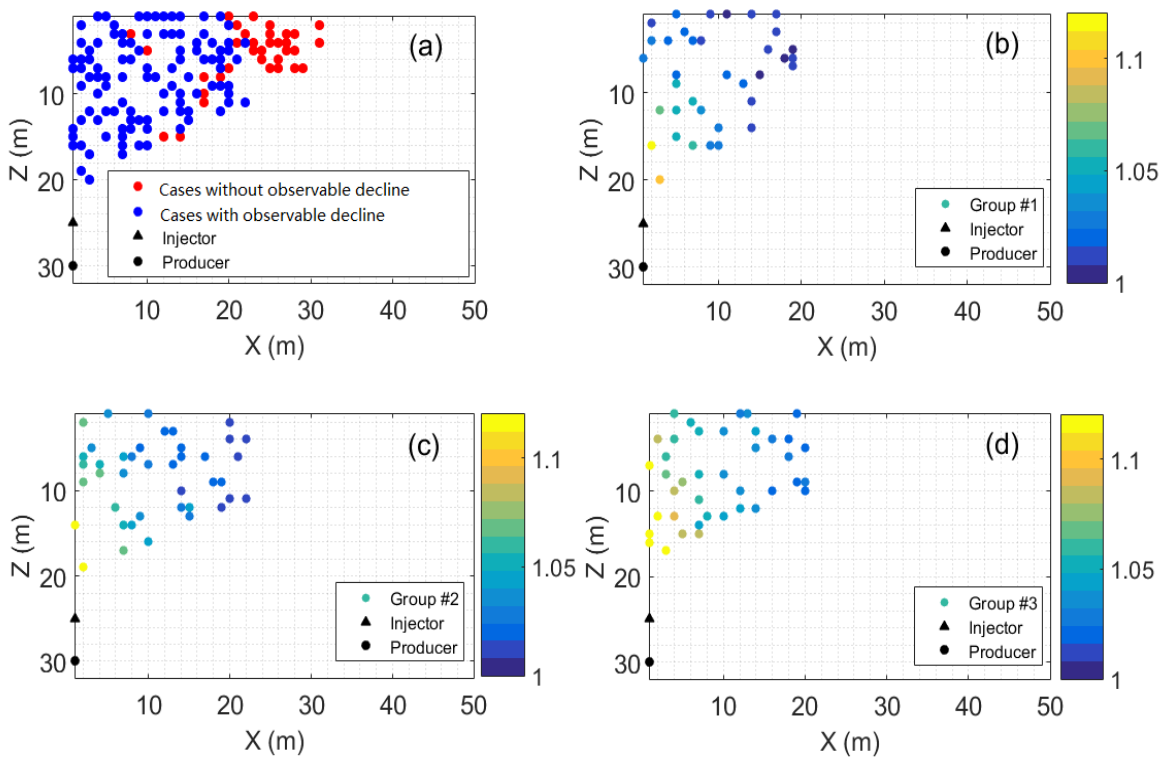


Fig. 5-8 Location maps of shale barriers – position of the lower-left corner of each shale barrier is indicated: (a) – all 400 models; (b) – group # 1 ($L \leq 8$ m); (c) – group # 2 (8 m $< L \leq 11$ m); (d) – group # 3 ($L > 11$ m). The color scale represents the ratio of Q_r at t_r of the homogeneous model to the heterogeneous model.

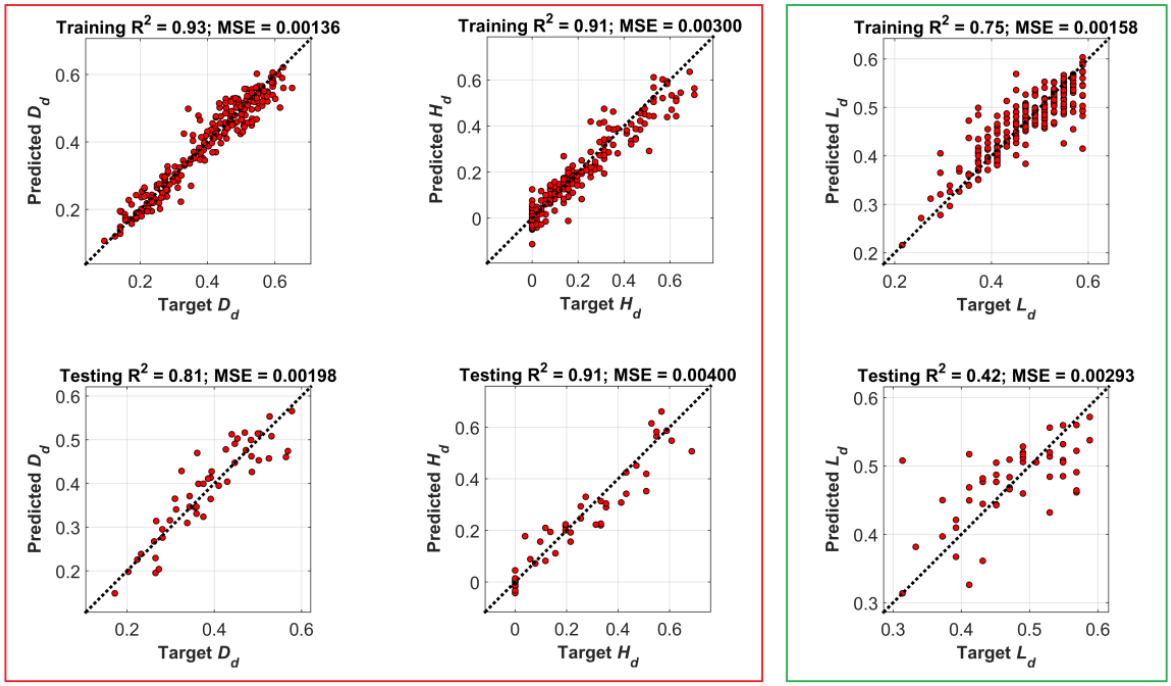


Fig. 5-9 Results of shale characterization (D_d, H_d, and L_d) from two ANN models in the single shale barrier case using piecewise linear approximation and cubic spline interpolation coefficients as inputs: top row – training dataset; bottom row – testing dataset.

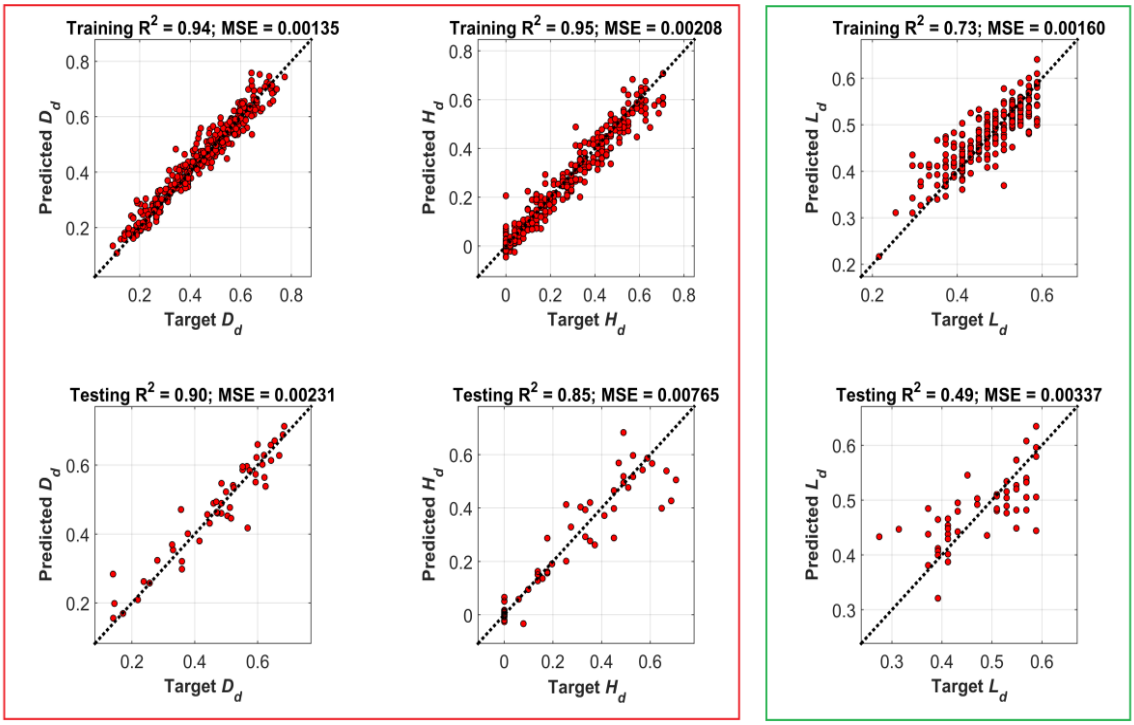


Fig. 5-10 Results of shale characterization (D_d, H_d, and L_d) from two ANN models in the single shale barrier case using DWT coefficients as inputs: top row – training dataset; bottom row – testing dataset.

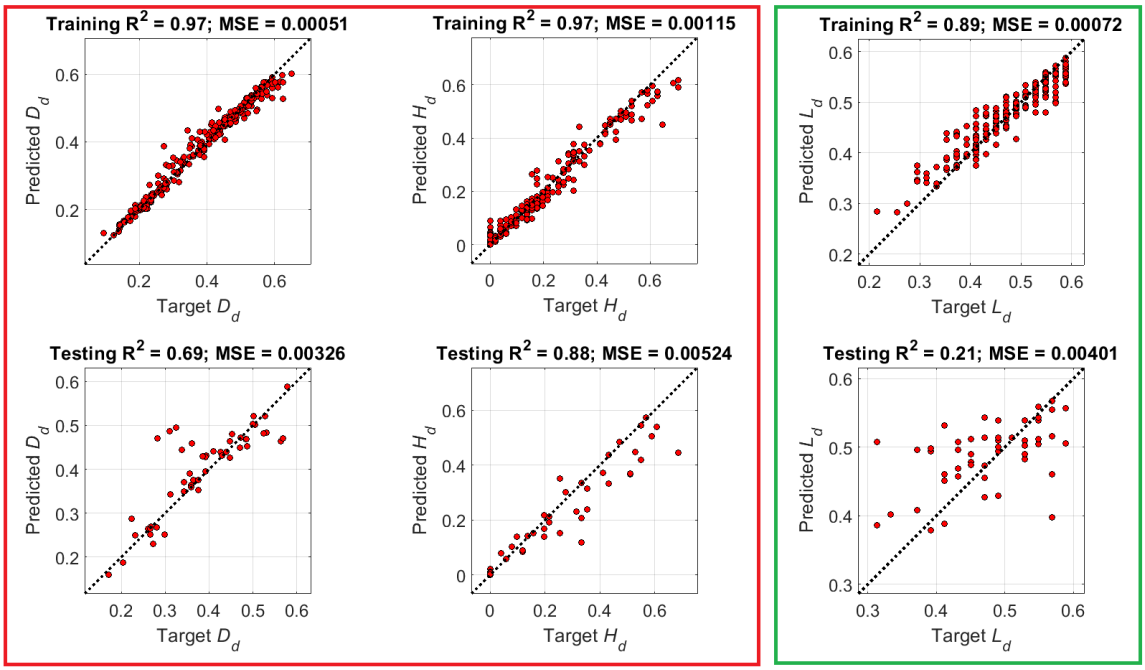


Fig. 5-11 Results of shale characterization (D_d, H_d, and L_d) from two RF models in the single shale barrier case using piecewise linear approximation and cubic spline interpolation coefficients as inputs: top row – training dataset; bottom row – testing dataset.

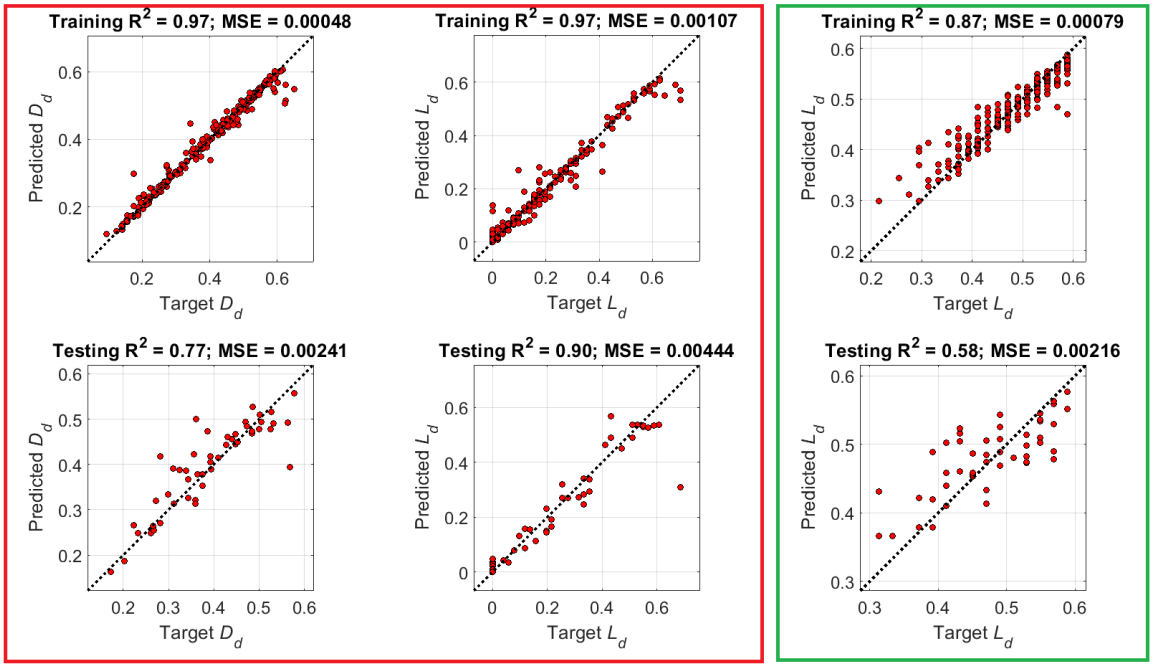


Fig. 5-12 Results of shale characterization (D_d, H_d, and L_d) from two RF models in the single shale barrier case using DWT coefficients as inputs: top row – training dataset; bottom row – testing dataset.

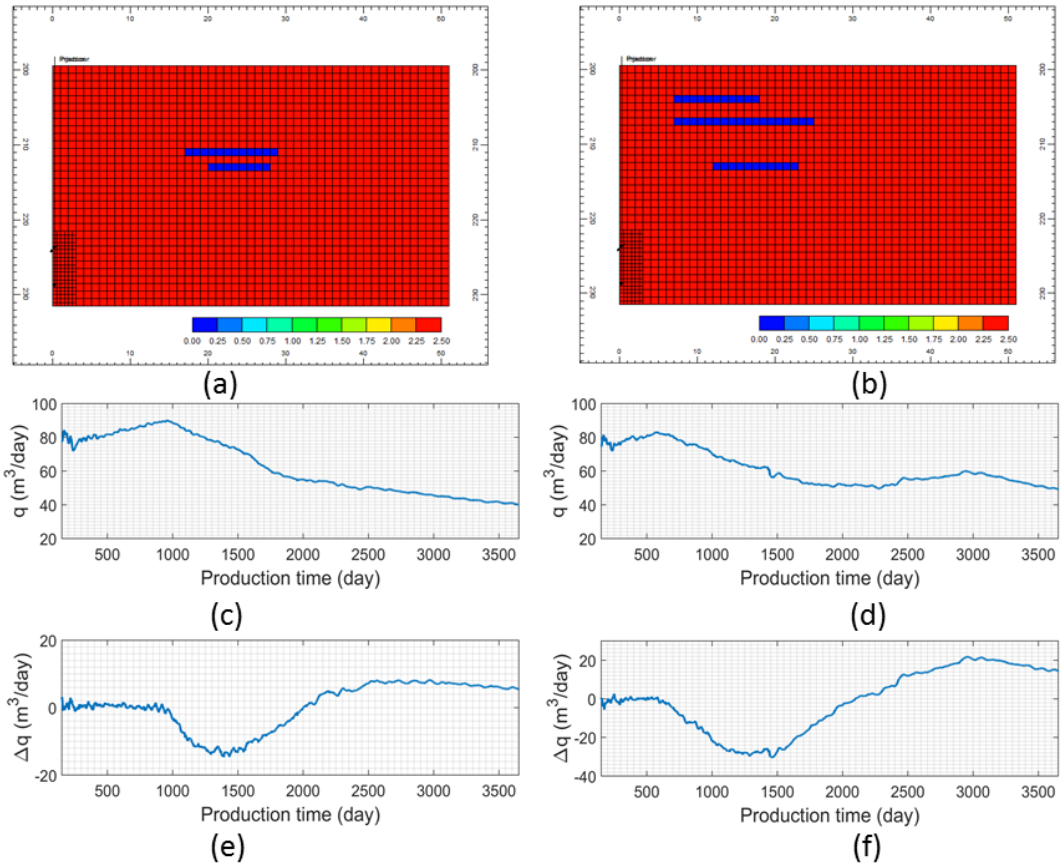


Fig. 5-13 Examples of multiple shale barriers configurations that exhibit only one decline pattern. Top row – permeability distribution in Darcy; middle row – q profiles; bottom row – Δq profiles.

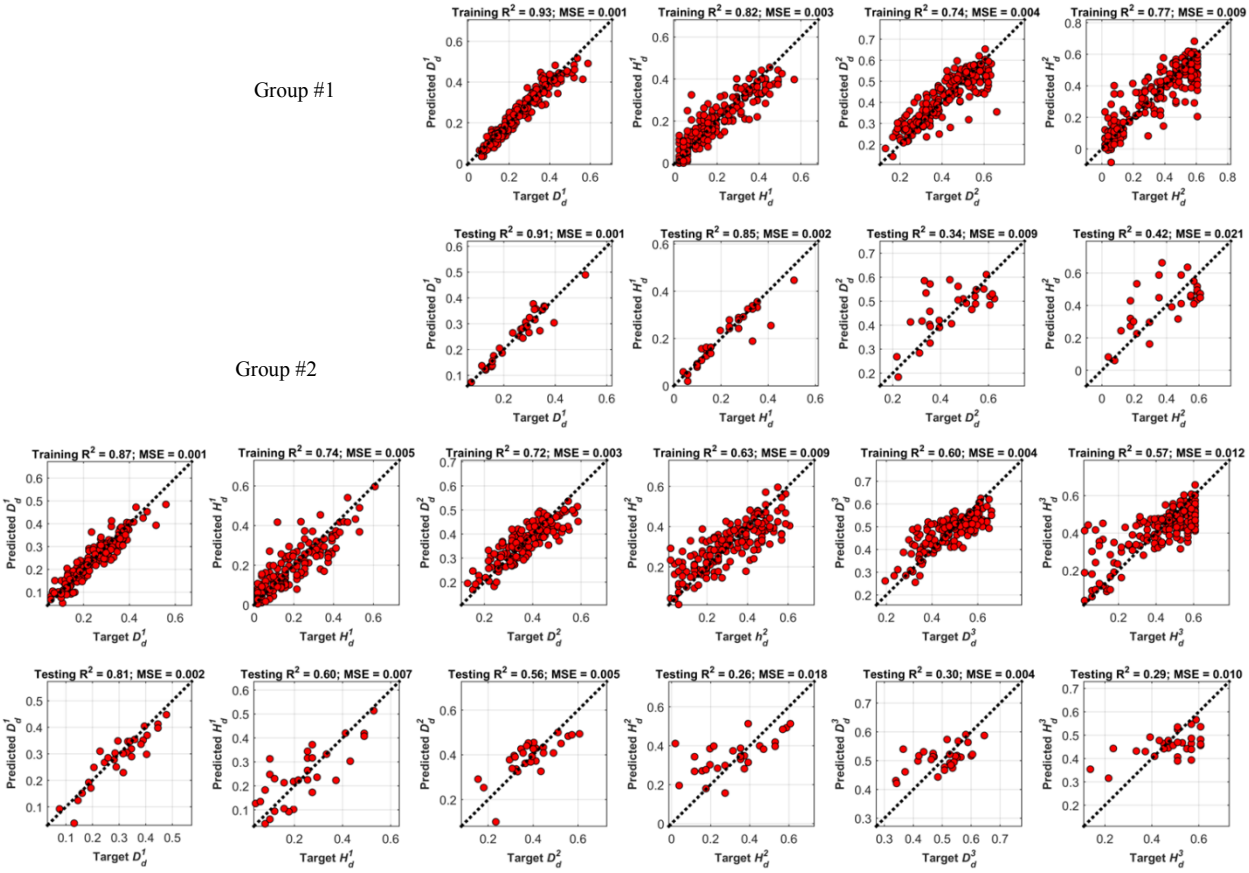


Fig. 5-14 Results of shale characterization (D_d and H_d) from ANN models in multiple shale barriers case: for each group, top row – training dataset; bottom row – testing dataset.

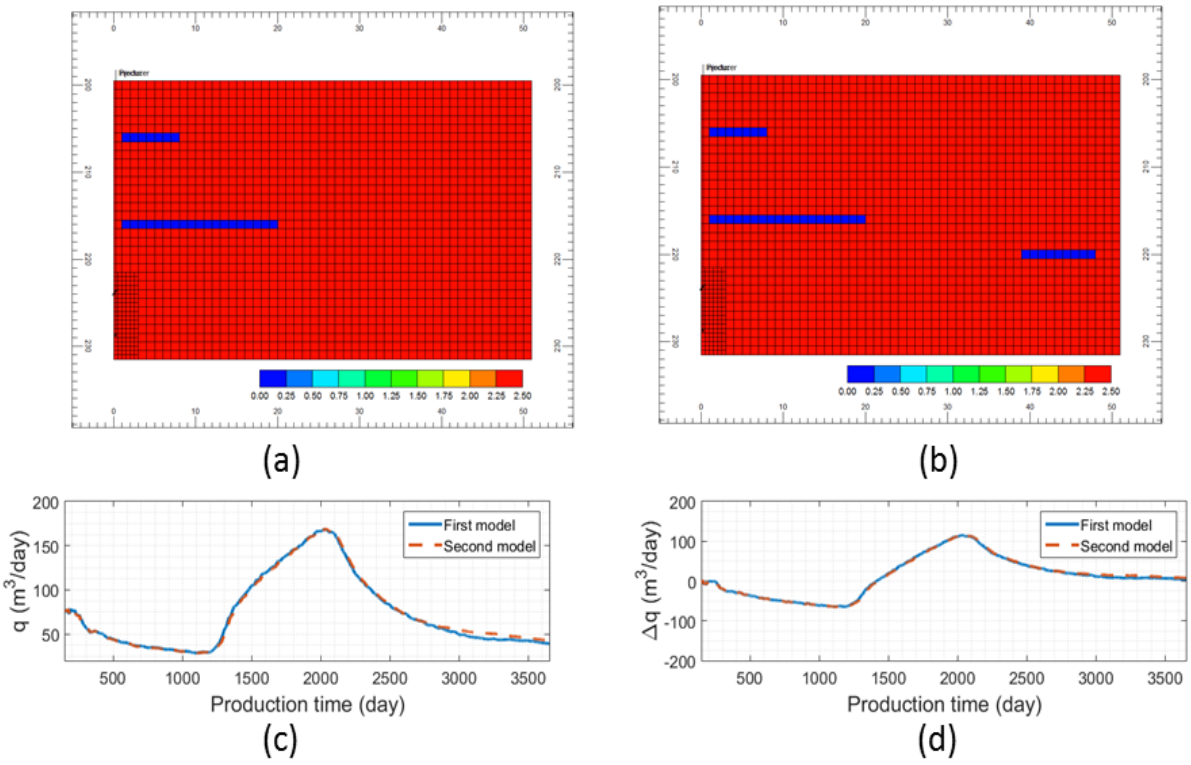


Fig. 5-15 Examples of multiple shale barriers configurations that exhibit similar production behavior. Top row – permeability distribution in Darcy; bottom row – q and Δq profiles.

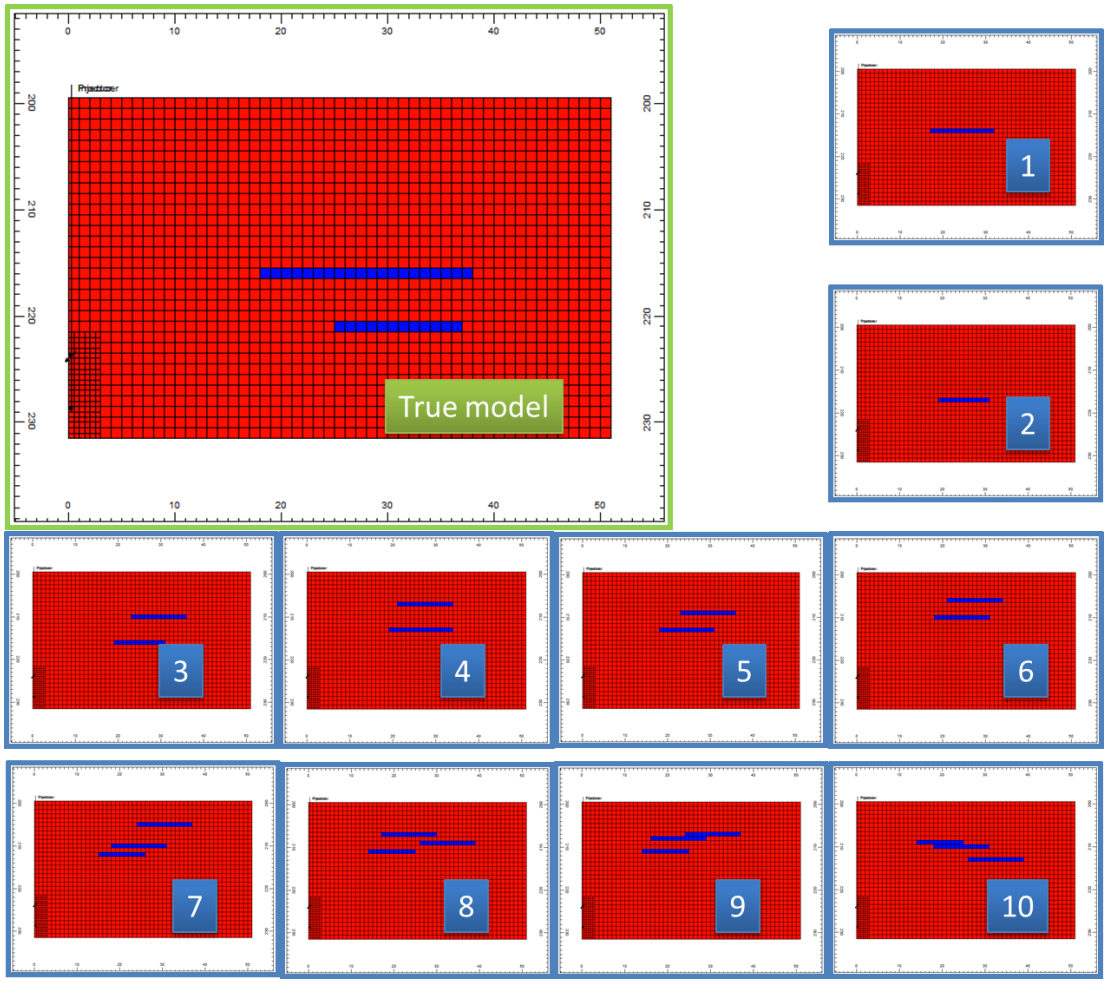


Fig. 5-16 Comparison of shale barrier distribution between the true model and 10 randomly-selected history-matched models

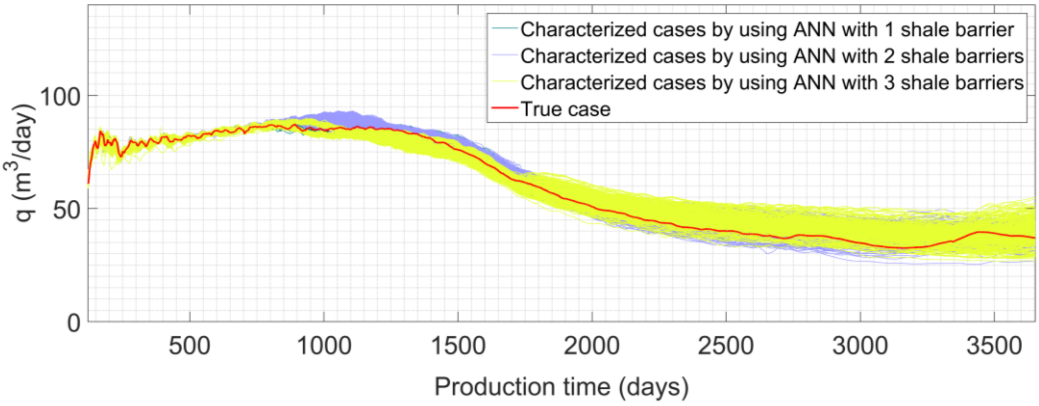


Fig. 5-17 Comparison of q profiles between the true case and an ensemble of history-matched models.

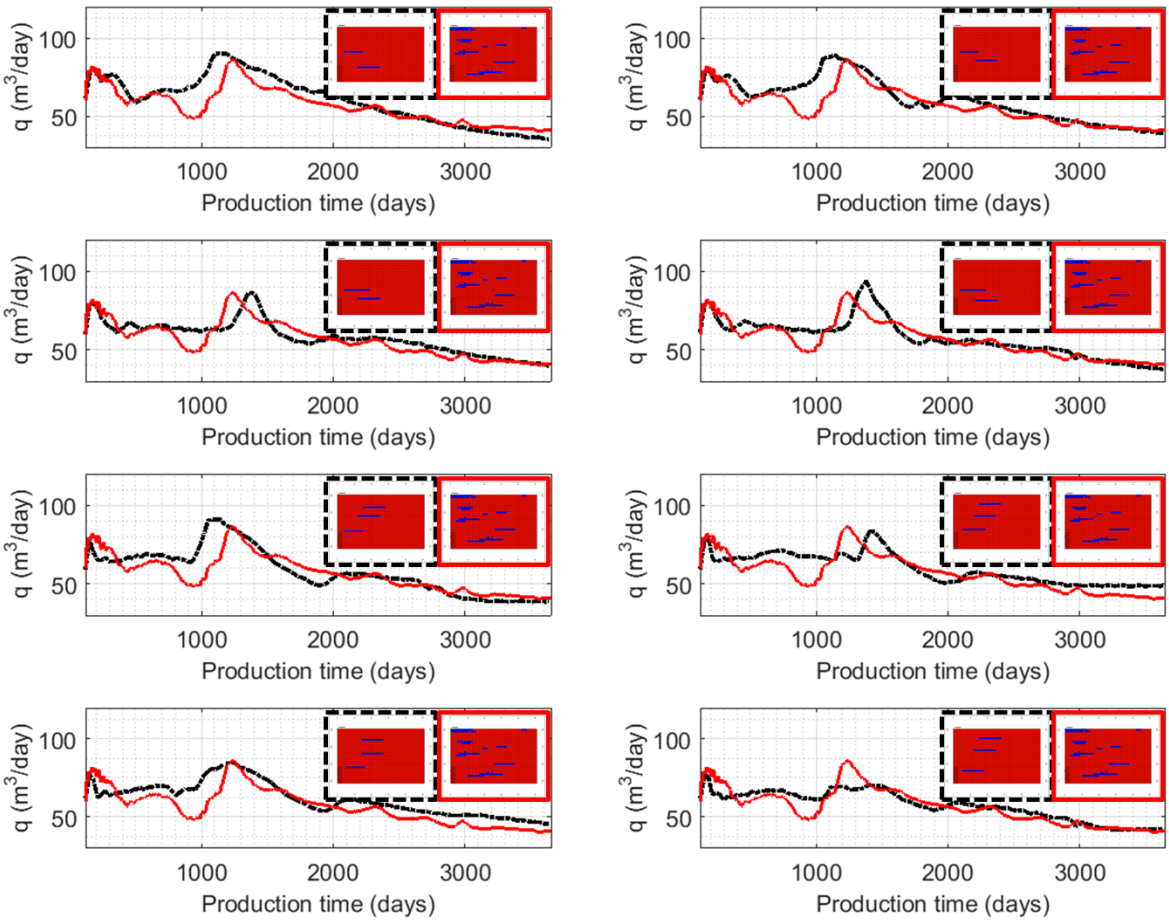


Fig. 5-18 Comparison of q profiles between the case with stochastically-distributed shale barriers (red solid line) and a few history-matched models (black dashed line): the inset figure in each subplot compares the shale distributions from the stochastically- distributed shale barriers model (red frame) and the corresponding history-matched model (black dashed frame).

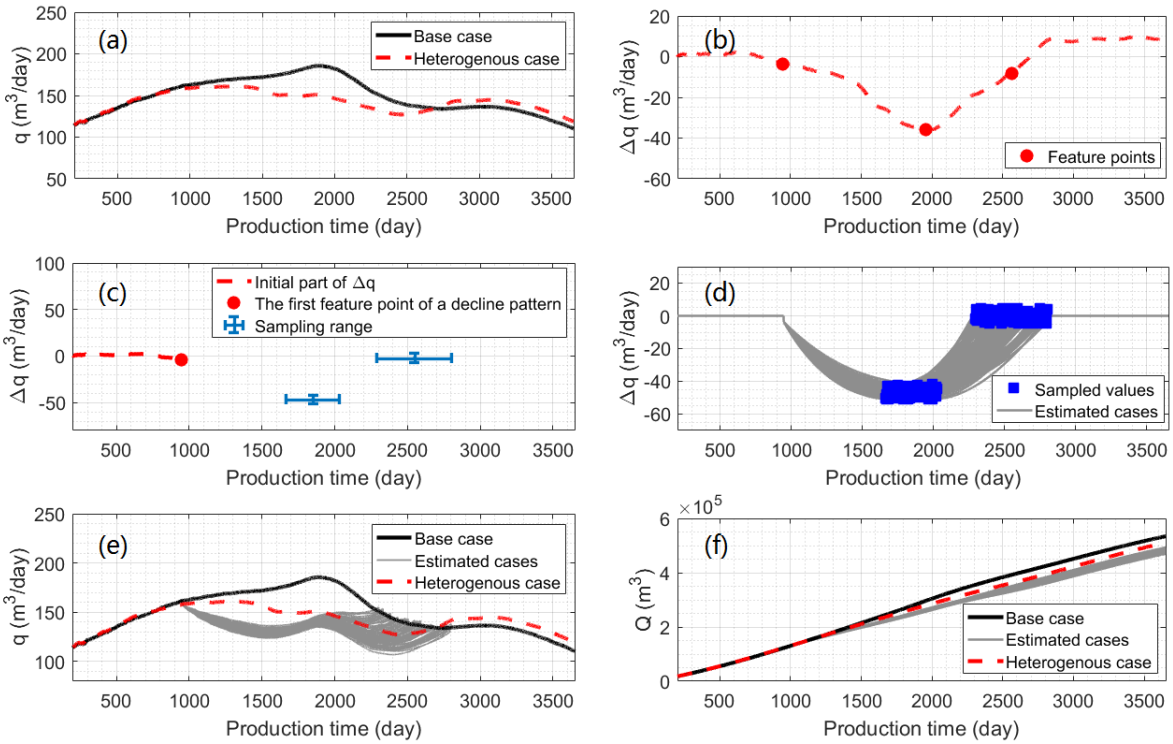


Fig. 5-19 A production history-matching example where a partial production profile is known (a) – q profiles of the base (homogenous) case and the true heterogeneous case; (b) – Δq profile and the three corresponding features points of the true heterogeneous case; (c) – partial production profile and sampling ranges for two unknown feature points; (d) – sampled feature points and the fitted Δq profiles for 100 cases; (e and f) – q and Q profiles of the 100 estimated cases, the heterogeneous case, and the base case.

Chapter 6 Characterization of Lean Zone and Shale Barrier in SAGD Reservoirs Using Data-Driven Modeling Techniques[4]

Chapter Overview

High water saturation zone, which is also known as lean zone, is another type of heterogeneous feature in SAGD reservoirs. Similar to shale barrier, lean zone also poses a detrimental influence on conventional SAGD operations, as it causes steam utilization efficiency to decrease and increases SOR. Although shale heterogeneities are studied and characterized by the proposed data-driven modeling workflow in the previous chapter, the effects of lean zone regions have not been studied.

In this chapter, both types of heterogeneities are investigated using synthetic models, which are built based on field data extracted from several existing SAGD projects. Lean zones and shale barriers with varying distribution, volume, and orientation are studied. From the corresponding production time-series data, a set of input features are identified through discrete wavelet transform (DWT) and principal component analysis (PCA), while the output parameters are formulated to describe the actual number and geological parameters of two types of heterogeneities. A two-level data-driven model is employed to characterize heterogeneities characteristics. Finally, this calibrated model is integrated into a novel characterization workflow

[4] A version of this chapter will be submitted for publication as:
Ma, Z. & Leung, J. Y. Integration of data-driven modeling techniques for lean zone and shale barrier characterization in SAGD reservoirs.

to infer an ensemble of probable realizations of heterogeneities distributions that are conditioned to a given production historical profile.

This chapter demonstrates the potential of practical application of data-driven models in correlating complex reservoir heterogeneity properties and production time-series data. Results from the case study illustrate the utility of the proposed workflow in facilitating the efficient identification of heterogeneous features from SAGD profiles.

6.1 Introduction

Heavy oil or bitumen, with extremely high viscosity and density, is an important unconventional crude oil in Canada. According to the Alberta Energy Regular (2016a), the bitumen reserve in Alberta is approximately 165 billion barrels, as of 2015. However, the total extracted volume is less than 10% (only 11 billion barrels) since the late 1960s. SAGD is an effective thermal recovery technique and was originally invented by Dr. Butler and his colleagues in Imperial Oil in the 1970s. As shown in **Fig. 6-1**, two parallel horizontal wells are drilled into the bottom of the target formation. One well is used as the injection well, while the other is used as the production well placed at approximately 5 m lower than the injector. The concept of SAGD involves 1) injection of high-temperature steam into an expanding steam chamber to heat the bitumen and reduce its viscosity; 2) the heated oil and condensate would drain down to the production well along the edge of the steam chamber by gravity (Butler et al., 1981). Though the process is highly effective in most cases, its performance can be severely hampered by the reservoir heterogeneities present along the path of the expanding steam chamber (Chen et al. 2008; Sheng, 2013).

Extensive research has been conducted in the past to investigate the influence of heterogeneous features, including lean zone and shale barrier, on SAGD performance through experiments and numerical simulations. In the experimental study of Yang and Butler (1992), they introduced reservoir heterogeneities by placing thin shale layers and varying permeability in reservoir layers. The main finding from that work was that the location of individual shale layer is significant. A stochastic model of shale distribution was used in numerical simulation study by Chen et al. (2008). Their results showed that the presence and distribution of shale in the near-well region has a bigger impact on the drainage pattern and ensuing production. Laterally-

continuous shale barrier or a high proportion of shale in the above-well region would be particularly detrimental to SAGD performance. Similar findings were also reported by others (Amirian et al., 2015; Wang and Leung, 2015).

Lean zones, such as those in Nexen's Long Lake and Suncor's Firebag projects (Xu et al., 2014), also pose some practical challenges to steam utilization in SAGD operations by increasing the steam-oil ratio (SOR). They often act as thief zones, and in order to minimize its impacts, the operating pressure should be kept below a certain threshold. According to the simulation results in Xu et al. (2014) and Doan et al. (2003), lean zones that are located below the producer (e.g., bottom water) would have a negligible influence. On the other hand, lean zones that are located at depths above the injector and in between the injector and producer may act as thief zones, resulting in a higher cumulative steam-oil ratio (cSOR) (Wang and Leung 2015). Pooladi-Darvish and Mattar (2002) also discussed the issue of steam chamber collapse due to the presence of water zones located above the pay interval. Similar studies were presented by Law et al. (2003) to investigate the impacts of confined and non-confined over- and under-laying water zones. In the end, works focusing on intra-formational water zones are rare (Fairbridge et al., 2012), which illustrated that the spatial distribution of water zones would contribute to the non-uniform growth of steam chamber with uneven drainage. The effects of lean zones were also studied in the works of Harding et al. (2016), Hocking et al. (2013), and Yang et al. (2016).

Therefore, it is necessary to identify the presence of reservoir heterogeneities and quantitatively determine their characteristics (e.g., locations, distributions, and vertical and areal extents). Inference of reservoir (petrophysical) properties from production data (flow rate and pressure measurements) is often posed as an inverse problem referred to history-matching. The inverse problem is usually ill-posed with non-unique solutions, such that multiple possible combinations of model parameters may produce reservoir responses that are consistent with the actual production data (Oliver and Chen, 2011). Common assisted history-matching techniques include gradient-based, Markov chain Monte Carlo, stochastic probability perturbation, evolutionary methods, and ensemble filtering techniques (Caers and Hoffman, 2006; Gu and Oliver, 2005; Pyrcz and Deutsch, 2014; Williams et al., 1998; Zhang et al., 2003).

These techniques were applied to characterize SAGD reservoir properties in the past. A stochastic optimization approach was implemented to estimate horizontal permeability, initial oil situation, and porosity for a homogeneous synthetic model in Jia et al. (2009). Mirzabozorg et al.

(2013) applied an evolutionary algorithm to estimate reservoir properties including the length of a single shale layer located at a fixed depth. To represent the complex shale barrier configurations in reservoir models, various model parameterization schemes have been adopted. In Hiebert et al., (2013), steam chamber inferred from seismic data was used in a history-matching process to identify geo-bodies. In Panwar et al. (2012), formation facies were perturbed during history-matching of well temperature data, during which the Ensemble Kalman Filter (EnKF) and discrete cosine transform were implemented. Although these assisted history-match techniques are quite robust for integrating a wide range of production (rate, temperature, saturation etc.), they are computationally intensive (Mirzabozorg et al., 2013; Zhang et al., 2014). Additionally, certain assumptions and simplifications regarding the operational constraints, governing equations, and process physics must be invoked in the simulation model. Previous works that focus specifically on the inference of stochastic heterogeneous distributions of shale barrier and lean zone have been limited.

Therefore, the motivation of this work is to propose alternative frameworks for inference of lean zones and shale barriers from production data. The integration of data-driven modeling techniques, which are useful for analyzing relevant data and detecting relationships and patterns between the system variables (Solomatine et al., 2009), for the purpose of heterogeneity inference is explored in this work. An important advantage is that explicit knowledge of the physical system is not required. In particular, artificial neural network (ANN), which is one of the most widely-applied data-driven modeling techniques (Agatonovic-Kustrin and Beresford, 2000), is adopted to approximate the non-linear relationship between production profiles and heterogeneity parameters. Since its proposal by McCulloch and Pitts (1943), ANN has been successfully adopted in many fields of study. Recent practical application of ANN in the petroleum engineering literature can be found in Adibifard et al., (2014), Ma et al. (2015; 2017; 2018) and Ulker and Sorgun (2016). In particular, a characterization workflow involving a number of data analytics approaches is developed. The main objectives are: 1) proposing a robust scheme for parameterization of heterogeneous lean zones and shale barriers; 2) extracting features of production data using methods associated with time-series analysis; and 3) developing data-driven models to approximate the relationship between heterogeneity parameters and production data features. A major advantage of the proposed workflow is that, once the model is calibrated, it can be used to quickly identify a suite of probable realizations of

heterogeneity distributions that are consistent with the production history. As will be shown in the case study, these realizations often reveal significant detail pertinent to large-scale heterogeneity description. They could also serve as reliable initial guesses for further history-matching analysis.

This chapter is organized as follows: first, construction of the synthetic data set is described in detail, which includes model setup, output parameterization, input feature extraction, and data-driven modeling schemes are presented in section 6.2; results for the data-driven modeling and characterization workflow are discussed in section 6.3; finally, in section 6.4, major conclusions and findings from this study are summarized.

6.2 Methods

6.2.1 Heterogeneous Model Setup

A base SAGD reservoir model, as shown in **Fig. 6-2**, is constructed based on typical Athabasca oil sands properties. In particular, representative values corresponding to Nexen's Long Lake and Suncor's Firebag projects are extracted from the public domain (Alberta Energy Regulator, 2015; 2016b). For symmetric development and growth of steam chamber, which is a common assumption in most numerical simulation studies, only a half of the distance between two well pairs is modeled. The SAGD model is 51 m × 900 m × 32 m, and the grid size is 1 m, 900 m, and 1 m in the X, Y, and Z direction, respectively. A horizontal well pair of 900 m is located at the left boundary (perpendicular to the X-Z plane). Heterogeneous features are superimposed onto this base model: characteristics of individual shale barrier and lean zone, such as the location, length, thickness, permeability, and water saturation (S_w), are assigned according to those described in Ma et al. (2017). Values of oil saturation (S_o) in sand, ranging between 0.75 and 0.85, are sampled from a uniform distribution. The detailed reservoir, fluid, and operational parameters can be found in **Table 6-1**. In this study, shale barriers have low horizontal and vertical permeabilities, with S_w of 1, while lean zones have same porosity and anisotropic permeability as the sand, but with a higher S_w of 0.9.

6.2.2 Parameterization of Heterogeneous Features

In order to construct data-driven models for the inference of heterogeneous features from production profiles, it is important to formulate a proper scheme for parameterizing the location and size of each distinct feature.

6.2.2.1 Number of Heterogeneous Features

The number of lean zones and shale barriers is denoted by N_l and N_s, respectively. These parameters represent "regions" of lean zones and shale barriers, and they may not necessarily correspond to the exact number of individual features. An example is shown in **Fig. 6-3**(**a**), where N_l = 3, despite the total number of lean zone features is 5. The rationale is that two lean zones near the top are merely too small to instigate any noticeable impacts on the production pattern. Similarly, numerous smaller features can be grouped into 3 consolidated regions of lean zone in **Fig. 6-3**(**b**).

6.2.2.2 Properties of Heterogeneous Features

A novel parameterization scheme is proposed to capture key geometrical aspects of a heterogeneous feature that could impact the development of a steam chamber, as shown in **Fig. 6-2**(**c**). *D* refers to the shortest distance between the bottom-left corner of a given feature and the injector; *H* represents the shortest horizontal distance between the feature and the left model boundary, and *L* denotes the lateral extent of the feature. They are presented in dimensionless form (subscript *d*) as:

$$\begin{aligned} D_d &= \frac{D}{D_{\max}} \\ H_d &= \frac{H}{0.5S} \\ L_d &= \frac{L}{0.5S}, \end{aligned} \quad \text{.....(6-1)}$$

where D_{max} refers to a maximum value of *D*, which is the distance between the injector and the top-right corner of the model; *S* denotes the surface distance between two adjacent well pairs.

Considering that only half of that length is modeled in this work, a factor of 0.5 is included, and $0.5S = 51$ m.

6.2.3 Parameterization of Production Profiles

6.2.3.1 Production Profiles Analysis

The constructed heterogeneous SAGD models are subjected to numerical flow simulator, i.e., STARS (CMG, 2015), and the corresponding production profiles are analyzed. Four examples are shown in **Fig. 6-4**, and the corresponding oil production rate (q) and instantaneous steam-oil-ratio ($iSOR$) are compared with that of the homogeneous base case in **Fig. 6-5**(**a** and **b**). Case #1 illustrates the impact of a single lean zone, both q and $iSOR$ will increase rapidly once the steam chamber has encountered a portion of the lean zone. As the steam chamber advances past the lean zone, the impact of the lean zone would gradually dissipate. The impact of a single shale barrier is illustrated in case #2. An instantaneous reduction in q is noticeable when the steam chamber makes contact with the shale barrier, and q continues to decline until the steam chamber bypasses the entire shale barrier; there is also a slight increase in $iSOR$.

In the presence of multiple heterogeneous features (case #3 and #4), the changes in production profiles are more dramatic and complex. Features may either exacerbate or offset one another. Therefore, instead of identifying an individual pattern corresponding to every heterogeneous feature, a more robust input feature extraction technique, such as discrete wavelet transform (DWT), is applied.

Another strategy for amplifying the minute variations in the production time-series is proposed, in which the Δq and $\Delta iSOR$ are computed:

$$\begin{aligned} \Delta q_t &= q_t^{\text{heterogeneous}} - q_t^{base} \\ \Delta iSOR_t &= iSOR_t^{\text{heterogeneous}} - iSOR_t^{base} \end{aligned} \quad \text{..... (6-2)}$$

where the subscript t represents a certain production time. The profiles of Δq and $\Delta iSOR$ of cases #1-4 are presented in **Fig. 6-5**(**c** and **d**). A sensitivity analysis confirms that profiles of cumulative quantity are generally too smooth for variation detection; hence, they are not suitable for feature extraction in this context.

6.2.3.2 Time-Series Features Extraction Using Discrete Wavelet Transform

DWT is based on a sub-band coding scheme, which is capable of providing a fast wavelet transform with minimal computational effort (Azim et al, 2010). After applying DWT, a coarse approximation of the original series is represented by cA, while high-frequency detail (or noise) is retained in cD (Yohanes et al, 2012). A schematic of DWT decomposition of a time-series data (Δq or $\Delta iSOR$ in this chapter) is presented in **Fig. 6-6**, where the cA coefficients are retained for subsequent modeling.

The production time-series data are decomposed to the 7th level using the Daubechies wavelet (db4). A sensitivity analysis was performed to determine the optimal decomposition by considering the performance the ensuing data-driven models and the quality of the reconstructed time series. Finally, 34 cA coefficients are retained. A comparison between the original and reconstructed Δq and $\Delta iSOR$ profiles, as shown in **Fig. 6-7**, reveals acceptable accuracy with the 7^{th} decomposition level. Finally, two sets of cA coefficients from Δq and $\Delta iSOR$ are incorporated into an input vector (contains 68 input features).

PCA (Jolliffe, 2005; Smith, 200) is also applied to reduce the input dimensionality; the resultant principle scores are considered as input features for the data-driven modeling. It should be noted that several other techniques for time-series parameterization were explored; it was observed that techniques such as cubic spline interpolation and cosine interpolation would often yield a feature vector that is too large in dimensionality; the issue becomes that reducing its dimensionality would compromise the ability to capture the essential characteristics of the time-series that are sensitive to the heterogeneity parameters. Therefore, they are not employed in this chapter.

6.2.4 Construction of Data-Driven Models

6.2.4.1 Correlation between Inputs and Outputs Using ANN

ANN is used to correlate the time-series features (inputs) the heterogeneous features (outputs). A typical 3-layer ANN structure is shown in **Fig. 6-8**. Two nodes or neurons between adjacent layers are connected by a weight (w_{ij}). In addition, a bias term (b) that is associated with each

neuron in hidden and output layer, which serves as an adjustable offset for the activation function (Behler and Parrinello, 2007). A common gradient descent training technique called backpropagation algorithm is employed to calibrate w_{ij} and b (Haykin, 2008). The output signal y of a given hidden or output node is computed as:

$$y_i = f\left(b_i + \sum_{j=1}^{m} w_{ij} \cdot x_j \right) \quad \text{(6-3)}$$

where y_i denotes the output from the node i of the current layer; b_i is the bias term for node i; node i is connected to the node j in the preceding layer by weight w_{ij}, and there are m nodes in the preceding layer; f is the activation or transfer function; in this work, the hyperbolic tangent function, which is a form of sigmoid function and with values ranging from -1 to 1, is employed for f. In order to alleviate the influence of large disparity among different input or output variables due to different data scales or ranges, it is important to normalize the input and output parameters. Detailed theories and mathematical derivation for the backpropagation method can be found elsewhere (Bishop, 1995; Haykin, 2008). The optimal ANN architecture is determined using an n-fold cross-validation workflow, which identifies the network configuration with the smallest mean squared error (*MSE*) between the predictions and targets, as illustrated in Ma et al. (2015). In this chapter, the MATLAB Neural Network ToolboxTM (Beale et al., 1992) is employed to construct ANN models.

6.2.4.2 Construction of a Two-Level Heterogeneity Characterization Workflow in SAGD Reservoirs

To construct a training data set, 15 heterogeneity scenarios, where the number of regions of shale barriers and lean zones varies between zero and three, are modeled. All 2800 cases and their corresponding scenarios are summarized in **Table 6-2.** For a given heterogeneity scenario, N_h is defined as the sum of N_l and N_s (i.e., total number of lean zone and shale regions). The output vector is formulated as: [N_l, N_s, D_d^j, H_d^j, L_d^j, ...], j = 1, …, N_h. A preprocessing step is required to rearrange the heterogeneous features in the output vector in an increasing order of steam chamber arrival. The rationale is that the influence of a particular feature diminishes as the steam chamber advances away from the well pair. In addition, it is observed from the training data set

that as N_h increases, the production response becomes less sensitive to the parameter L_d; therefore, for scenarios with $N_h \geq 3$, L_d is omitted from the output vector for all heterogeneity features.

A two-level modeling procedure, as shown in **Fig. 6-9(a)**, is implemented. For the first level, a screening model (referred to as screening-ANN) is used to predict the appropriate heterogeneity scenario. The input vector is composed of 25 principal scores (capturing over 97% of the variance in the original 68 DWT coefficients), while the output vector consists of N_l and N_s. **Fig. 6-10** shows a scree plot after PCA. The n-fold (n = 10) cross-validation routine is applied to explore the optimum ANN structure; structures with single and two hidden layers are tested, and the number of hidden nodes in each layer is varied between 3 and 20. The optimal ANN architecture is determined to include two hidden layers with 20 × 3 hidden nodes. Among all 2800 samples, 420 samples (15%) are randomly chosen for model testing. A summary of the screening-ANN model construction is presented in **Table 6-3**.

In the second level, a series of models (referred to as sub-ANNs) are constructed to predict the specific parameters of the heterogeneous features corresponding to each scenario. For these models, the input vector is composed of the 25 principal scores (same as the screening-ANN model); however, the output vector consists of D_d, H_d, and L_d. If a particular scenario involves two types of heterogeneity, a separate ANN model is needed for each heterogeneity type: one for the lean zones and another one for the shale barriers. Hence, the total number of sub-ANN models in this step is 24. Details are summarized in **Table 6-2**.

Several remarks regarding the proposed two-level modeling approach should be made here. A few alternative hierarchical workflow and machine learning techniques were tested, but the results were inferior. First, classification techniques, such as support vector machine (SVM), were integrated as a model selection tool to identify individual heterogeneity scenario (in place of the proposed screening-ANN step). However, the overall classification accuracy is 0.33, which is very low. Alternatively, a multi-level classification routine, which involved grouping the 15 scenarios into several sub-groups, was also explored. The routine is capable of identifying scenarios that are distinctly different from the rest (e.g., cases with only shale barriers (scenario #13, 14, and 15) can be easily differentiated from the other 12 scenarios). However, at each subsequent level, the classification accuracy deteriorates, as the similarities in the production profiles among these scenarios increase.

A second attempt was to introduce a heterogeneity indicator (defined as the permeability ratio between a given heterogeneity and the background oil sand) as an additional output variable; the goal was to eliminate the screening/classification step and use a single ANN model to predict the heterogeneity type directly. The issue is that, even for a small number of heterogeneous features (e.g., fewer than 6), the procedure fails to characterize heterogeneity features effectively. It is believed that introducing an additional output variable would increase the dimensionality of the unknown model parameters and further exacerbating the high degree of nonlinearity between the production profiles and heterogeneity parameters. A similar conclusion was also reached when incorporating the value of S_w as an additional output variable. In that case, the minor change in S_w between lean zones and shale barriers is not very useful for differentiating between the two heterogeneity types.

6.2.5 Data-Driven-Based Heterogeneities Characterization Workflow

Once the two-level data-driven model is successfully constructed, it is employed to facilitate heterogeneity characterization for any given production time-series data, as described in **Fig. 6-9**(**b**). Following the approach in Ma et al. (2015), to approximate the uncertainty in the model outputs, a Gaussian likelihood function is proposed: the mean corresponds to the predicted value from the sub-ANN models and the variance corresponds to 20% of the mean. Bootstrapping is performed to sample a set of possible values for each parameter and used to construct realizations of the heterogeneity model. As explained in section 6.2.4.2, for the scenarios with $N_h \geq 3$, where L_d is absent from the output vector, the corresponding values are sampled from uniform probability distributions of [19, 29] and [17, 25] for the lean zone and shale barrier, respectively. These ranges are computed from the built 2800 cases.

6.3 Results and Discussion

6.3.1 Results of Data-Driven Models Construction

6.3.1.1 Screening-ANN Model

The training and testing performance of the screening-ANN model is presented in **Fig. 6-11**, where predictions of N_l and N_s over all scenarios and cases are presented. The four distributions

in each subplot refer to the predictions corresponding to each of the four targets: 0, 1, 2, and 3. For example, in the top-left histogram of **Fig. 6-11(a)**, the target value of N_l is zero, and the histogram represents the distribution of predicted N_l values based on the training set. Overall, reasonable agreement between the target values and predictions is achieved, as the prediction accuracy for N_l and N_s is 0.90 and 0.65, respectively, and these values are much higher than the SVM model, previously described in section 6.2.4.2. Although 100% prediction is not achieved, as evidenced by the non-zero variance in each of these histograms, the model results are still considered reliable because the predicted value with the highest frequency is the same as the target value.

It should be noted that the estimation of N_l is superior to that of N_s, and this difference can be attributed to the fact that the size of individual lean zone unit is thicker than that for the shale barrier; thus, the production profiles are more sensitive to the number of lean zones. The implication is that the predicted N_l can be used reliably to select the appropriate sub-ANN models that correspond to N_l shale barriers. However, due to the large uncertainty in the predicted N_s, it is necessary to consider the uncertainty associated with the distribution of predicted N_s. As an example, in **Fig. 6-11(b)**, for the case where the majority of predicted N_s values is 2, uncertainty should be taken into account by considering N_s = 1, 2, and 3, with 12.3%, 66.4%, and 21.3% of probability, respectively. Consequently three sub-ANN models corresponding to different N_s values should be used for the next step.

6.3.1.2 Sub-ANN Models

To illustrate the results of these sub-ANN models, performance corresponding to scenario #2, #7 and #12 is presented in **Figs. 6-12** to **6-14**. The values of R^2 (coefficient of determination) and *MSE* of the training and testing cases for all 24 sub-ANN models are shown in **Table 6-4** and **6-5**, respectively. Most points in the cross-plots would follow the 45-degree splitting line, which indicates a perfect estimation. In particular, for the case with N_l = N_s = 1 (**Fig. 6-12**), performance for both sub-ANN model corresponding each heterogeneity type is good. Satisfactory results (small values in *MSE*) can also be observed in scenario #7 and #12.

However, it should also be noted that as N_h increases, the overall performance of all sub-ANN models would worsen. The reason is that impacts of individual heterogeneous feature on the overall production response may overlap, rendering isolating the contribution from each

individual heterogeneity to the aggregated production profiles to be extremely difficult. In other words, the solution to this characterization problem is non-unique; multiple configurations of heterogeneity may instigate similar impacts on the production profiles. Therefore, the goal is to identify a set of possible realizations or configurations that are consistent with the production profiles. The results in **Table 6-4** and **6-5** would also suggest that better prediction performance is achieved when D is low: a small value of D implies that the heterogeneous feature is located close to well pair; therefore, its impact would be readily reflected in the production profiles.

6.3.2 Application of Characterization Workflow

In this section, the two-level data-driven model is employed to characterize lean zones and shale barriers for a set of unknown production profiles. First, a set of profiles corresponding to heterogeneities with simplified (e.g., rectangular) geometries is studied; next, a set of stochastic cases with more realistic and irregularly-shaped heterogeneities is tested.

6.3.2.1 Cases with Simply-Shaped Heterogeneities

All heterogeneities are defined as rectangles with a thickness of 4 m and 1 m for the lean zone and shale barrier, respectively. Three cases with varying combinations of N_l and N_s are tested. An ensemble of realizations, which are consistent with the production profiles, is obtained following the proposed workflow: (1) multiple probable scenarios with different number of lean zones are sampled from the probability distribution of N_s (e.g., **Fig. 6-11**); (2) bootstrapping is performed to generate multiple heterogeneous realizations by regarding each set of estimated parameters as the mean and 20% of the mean as the variances as illustrated in **Fig. 6-9**(**b**) and section 6.2.5. Profiles of q and *iSOR* of these realizations are compared with that of the true model in **Fig. 6-15**(**b**, **d**, and **f**), while the heterogeneity configurations for the true models are shown in **Fig. 6-15**(**a**, **c**, and **e**). Reasonable agreement is observed for all three cases. Production profiles of the characterized models closely resemble those of the true model.

6.3.2.2 Cases with Irregularly-Shaped Heterogeneities

In this section, the workflow is studied using models with more complex heterogeneity distributions. As shown in **Fig. 6-16(a**, **c**, and **e)**, the location, size, and shape of individual heterogeneity feature are random. It is true that the proposed workflow could not resolve the complete detail pertinent to each and every heterogeneity feature; the idea, however, is to infer a

set of parameters that capture the aggregated impact of several heterogeneity regions. Therefore, instead of attempting to identify many small-scale heterogeneities with highly irregular geometry, the workflow could identify a few major regions of simply-shaped heterogeneities that have instigated most significant impact on the production profiles.

Once again, multiple realizations are inferred from the production profiles for each of the 3 testing cases. Profiles of *q* and *iSOR* of these realizations are consistent with those of the true model, as shown in **Fig. 16**(**b**, **d**, and **f**). This result illustrates the utility of the proposed workflow for inference of distributions or configurations of lean zones and shale barriers with complex geometry. Occasionally, several of these sampled realizations may deviate substantially from the true model; it is recommended that a post-processing procedure is applied to screen out the realizations with mismatches in production profiles exceeding a pre-defined error tolerance.

6.4 Conclusions

Two types of reservoir heterogeneity features, namely shale barrier and lean zone, which are common in SAGD reservoirs, are studied. A novel two-level data-driven modeling workflow is proposed to characterize such reservoir heterogeneities by correlating production time-series data and heterogeneity parameters using artificial neural networks. A synthetic training data set consisting of a total number of 2800 numerical SAGD simulations, which reflect representative petrophysical and operational parameters assembled from the field data, are constructed. The first-level model is used to predict the number of lean zone and shale barrier regions, while the second-level models are used to estimate specific properties corresponding to individual heterogeneous feature. Two case studies are used to examine the prediction performance of the proposed workflow. An ensemble of probable realizations of lean zone and shale barrier configurations, which are consistent with the actual production time-series data, are inferred. The proposed workflow may not resolve all the detail corresponding to every heterogeneity feature; however, it is capable of inferring approximately a set of regions of simply-shaped heterogeneity that have influenced the production profiles most significantly.

This work presents a practical workflow that integrates production time-series data analysis and data-driven modeling techniques for complex SAGD reservoir heterogeneity characterization. One limitation is that the maximum number of identifiable heterogeneity

scenarios is fixed, depending on the training data set and number of second-level models. However, in most cases, the ability to infer approximate location and geometry (e.g., lateral extent) of major heterogeneous regions would offer valuable insight for practical geo-modeling and decision-making purposes. The objective is to explore the potential of employing data-driven modeling techniques to analyze production data and characterize heterogeneous features. The workflow can be considered as a complementary tool to conventional reservoir modeling routines. Future work should extend the technique to 3D models and involve a field training data set.

6.5 Reference

Adibifard, M., Tabatabaei-Nejad, S., & Khodapanah, E. (2014). Artificial neural network (ANN) to estimate reservoir parameters in naturally fractured reservoirs using well test data. *Journal of Petroleum Science and Engineering*, 122, 585-594.

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis, 22*(5), 717-727.

Alberta Energy Regulator (2015), *AER Annual Performance Presentation*. Prepared by Nexen Long Lake Project.

Alberta Energy Regulator (2016a), *Alberta energy regulator 2015/16 annual report*. Report prepared by Alberta Energy Regulator, Alberta, Canada.

Alberta Energy Regulator (2016b), *AER Annual Performance Presentation*. Prepared by Suncor Firebag Projects.

Amirian, E., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Integrated cluster analysis and artificial neural network modeling for steam-assisted gravity drainage performance prediction in heterogeneous reservoirs. *Expert Systems with Applications, 42*(2), 723-740.

Azim, M. R., Amin, M. S., Haque, S. A., Ambia, M. N., & Shoeb, M. A. (2010). Feature extraction of human sleep EEG signals using wavelet transform and fourier transform.

Paper presented at the *2010 2nd International Conference on Signal Processing Systems*, 3 V3-701-V3-705, Dalian, China.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (1992). *Neural network toolbox™ user's guide*. The Mathworks Inc.

Behler, J., & Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14), 146401.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University press.

Butler, R., McNab, G., & Lo, H. (1981). Theoretical studies on the gravity drainage of heavy oil during in‐situ steam heating. *The Canadian Journal of Chemical Engineering, 59*(4), 455-460.

Caers, J., & Hoffman, T. (2006). The probability perturbation method: A new look at bayesian inverse modeling. *Mathematical Geology, 38*(1), 81-100.

Chen, Q., Gerritsen, M. G., & Kovscek, A. R. (2008). Effects of reservoir heterogeneities on the steam-assisted gravity-drainage process. *SPE Reservoir Evaluation & Engineering, 11*(05), 921-932.

CMG, 2015. STARS: Users' Guide, advanced processes & thermal reservoir simulator (Version 2015), Calgary, Alberta, Canada: Computer Modeling Group Ltd.

Doan, L., Baird, H., Doan, Q., & Ali, S. (2003). Performance of the SAGD process in the presence of a water sand-a preliminary investigation. *Journal of Canadian Petroleum Technology, 42*(01).

Fairbridge, J. K., Cey, E., & Gates, I. D. (2012). Impact of intraformational water zones on SAGD performance. *Journal of Petroleum Science and Engineering, 82*, 187-197.

Gu, Y., & Oliver, D. S. (2005). History matching of the PUNQ-S3 reservoir model using the ensemble kalman filter. *SPE Journal, 10*(02), 217-224.

Harding, T. G., Zanon, S., Imran, M., & Kerr, R. K. (2016). In-situ reflux: An improved in-situ recovery method for oil sands. Paper presented at the *SPE Canada Heavy Oil Technical Conference*, Calgary, Alberta, Canada.

Haykin, S.S. (2008). *Neural networks and learning machines* (3rd ed.), Upper Saddle River, NJ, USA: Pearson.

Hiebert, A. D., Morrish, I. C., Card, C., Ha, H., Porter, S., Kumar, A., & Close, J. C. (2013). Incorporating 4D seismic steam chamber location information into assisted history matching for A SAGD simulation. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Hocking, G., Cavender, T. W., & Person, J. (2011). Single-well SAGD: Overcoming permeable lean zones and barriers. Paper presented at the *Canadian Unconventional Resources Conference*, Calgary, Alberta, Canada.

Jolliffe, I. (2005). *Principal component analysis*, Wiley Online Library.

Kempeneers, P., De Backer, S., Debruyn, W., & Scheunders, P. (2004). Wavelet-based feature extraction for hyperspectral vegetation monitoring. Paper presented at the *Proceedings of SPIE*, Barcelona, Spain, 5238 297-305.

Law, D. H., Nasr, T. N., & Good, W. K. (2003). Field-scale numerical simulation of SAGD process with top-water thief zone. *Journal of Canadian Petroleum Technology, 42*(08).

Ma, Z., Leung, J. Y., & Zanon, S. (2018). Integration of artificial intelligence and production data analysis for shale heterogeneity characterization in steam-assisted gravity-drainage reservoirs, *Journal of Petroleum Science and Engineering, 163*, 139–155.

Ma, Z., Leung, J. Y., & Zanon, S. (2017). Practical data mining and artificial neural network modeling for SAGD production analysis. *Journal of Energy Resources Technology, 139*(3), 032909.

Ma, Z., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Practical implementation of knowledge-based approaches for steam-assisted gravity drainage production analysis. *Expert Systems with Applications, 42*(21), 7326-7343.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics. 5* (4): 115–133.

Mirzabozorg, A., Nghiem, L., Chen, Z., & Yang, C. (2013). Differential evolution for assisted history matching process: SAGD case study. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Oliver, D. S., & Chen, Y. (2011). Recent progress on reservoir history matching: A review. *Computational Geosciences, 15*(1), 185-221.

Panwar, A., Trivedi, J. J., & Nejadi, S. (2012). Importance of distributed temperature sensor (DTS) placement for SAGD reservoir characterization and history matching within ensemble kalman filter (EnKF) framework. Paper presented at the *SPE Latin America and Caribbean Petroleum Engineering Conference*, Mexico City, Mexico.

Pooladi-Darvish, M., & Mattar, L. (2002). SAGD operations in the presence of overlying gas cap and water layer-effect of shale layers. *Journal of Canadian Petroleum Technology, 41*(06).

Pyrcz, M. J., & Deutsch, C. V. (2014). Geostatistical reservoir modeling, Oxford University Press.

Sereda, J. N., & James, B. R. (2014). A case study in the application of bitumen geochemistry for reservoir characterization in SAGD development. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Sheng, J. (2013). *Enhanced oil recovery field case studies*, Gulf Professional Publishing.

Smith, L. I. (2002). A tutorial on principal components analysis. Cornell University, USA, 51(52), 65.

Solomatine, D., See, L. M., Abrahart, R. (2009). Data-driven modeling: Concepts, approaches and experiences. *Practical hydroinformatics* (pp. 17-30), Springer.

Ulker, E., & Sorgun, M. (2016). Comparison of computational intelligence models for cuttings transport in horizontal and deviated wells. *Journal of Petroleum Science and Engineering, 146*, 832-837.

Wang, C., & Leung, J. (2015). Characterizing the effects of lean zones and shale distribution in steam-assisted-gravity-drainage recovery performance. *SPE Reservoir Evaluation & Engineering, 18*: 329-345.

Williams, M., Keating, J., & Barghouty, M. (1998). The stratigraphic method: A structured approach to history matching complex simulation models. *SPE Reservoir Evaluation & Engineering, 1*(02), 169-176.

Xu, J., Chen, Z. J., Cao, J., & Li, R. (2014). Numerical study of the effects of lean zones on SAGD performance in periodically heterogeneous media. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Yang, G., & Butler, R. (1992). Effects of reservoir heterogeneities on heavy oil recovery by steam-assisted gravity drainage. *Journal of Canadian Petroleum Technology, 31*(08), 37-43.

Yang, R., Zhang, J., Yang, L., Chen, H., & Tang, S. (2016). Performance and calculation method of steam chamber overcoming high water saturation intervals during SAGD process. Paper presented at the *International Petroleum Technology Conference*, Bangkok, Thailand.

Yohanes, R. E., Ser, W., & Huang, G. (2012). Discrete wavelet transform coefficients for emotion recognition from EEG signals. Paper presented at the *Engineering in Medicine and Biology Society (EMBC), 2012* Annual International Conference of the IEEE, 2251-2254, San Diego, CA, USA.

Zhang, F., Reynolds, A. C., & Oliver, D. S. (2003). An initial guess for the Levenberg–Marquardt algorithm for conditioning a stochastic channel to pressure data. *Mathematical Geology, 35*(1), 67-88.

Zhang, X. K., Feizabadi, S. A., & Yang, P. (2014). An integrated approach to building history-matched geomoels to understand complex long lake oil sands reservoirs, part 1: Geomodeling. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Tables

Table 6-1 Reservoir properties and operating constraints for the base and heterogeneous models.

Reservoir depth (m)	200	Lean zone horizontal permeability (D)	2.5
Reservoir thickness (m)	32	Shale barrier horizontal permeability (D)	5×10^{-8}
Reservoir size in X direction (m)	51	Oil sand vertical permeability (D)	1.5
Injector depth (m)	225	Lean zone vertical permeability (D)	1.5
Producer depth (m)	230	Shale barrier vertical permeability (D)	3×10^{-8}
Length of horizontal well pair (m)	900	Oil sand porosity (fraction)	0.32
Initial temperature (°C)	15	Lean zone porosity (fraction)	0.32
Initial reservoir pressure (kPa)	1088 @ 230m	Shale barrier porosity (fraction)	0.25
Initial oil viscosity (cp)	592000	Range of initial S_o of oil sand (fraction)	[0.75-0.85]
Number of production wells	1	Initial S_o of lean zone(fraction)	0.1
Number of injection wells	1	Initial S_o of shale barrier (fraction)	0.0
Molar fraction of methane (%)	5	Range of S_w of oil sand (fraction)	[0.15-0.25]
Rock compressibility (kPa^{-1})	2.0×10^{-6}	Initial S_w of lean zone (fraction)	0.9
Rock heat capacity (J/m3*°C)	2.35×10^{6}	Initial S_w of shale barrier (fraction)	1.0
Thermal conductivity of matrix (J/m*°C)	1.468×10^{5}	Injection pressure (kPa)	2100
Thermal conductivity of oil (J/m*°C)	1.15×10^{4}	Preheating period (day)	60
Thermal conductivity of gas (J/m*°C)	1.3997×10^{2}	Total production time (day)	3652

Thermal conductivity of water (J/m*°C)	5.35×10^4	Thickness of lean zone (m)	4
Oil sand horizontal permeability (D)	2.5	Thickness of shale barrier (m)	1

Table 6-2 Summary of heterogeneity scenarios, number of cases in the training dataset, and optimized ANN model structures.

Scenario	N_h	N_l	N_s	# of cases in the training dataset	# of sub-ANN models	# of ANN outputs		Optimal ANN architecture	
						lean zone sub-model	shale sub-model	lean zone sub-model	shale sub-model
1	1	1	0	100	1	3	NA	[3]	NA
2	2	1	1	200	2	3	3	[20, 5]	[11, 5]
3	3	1	2	200	2	2	4	[7, 7]	[10, 6]
4	4	1	3	200	2	2	6	[4, 5]	[4, 3]
5	2	2	0	200	1	6	NA	[8]	NA
6	3	2	1	200	2	4	2	[3]	[5, 3]
7	4	2	2	200	2	4	4	[3, 4]	[3, 5]
8	5	2	3	200	2	4	6	[5, 3]	[3]
9	3	3	0	200	1	6	NA	[4]	NA
10	4	3	1	200	2	6	2	[3]	[13, 9]
11	5	3	2	200	2	6	4	[4, 4]	[18, 3]
12	6	3	3	200	2	6	6	[3, 4]	[4]
13	1	0	1	100	1	NA	3	NA	[18,3]

14	2	0	2	200	1	NA	6	NA	[3]
15	3	0	3	200	1	NA	6	NA	[5, 4]

Table 6-3 Summary for the screening-ANN model construction.

# of cases	# of training samples	# of testing samples	ANN architecture	# of inputs	# of outputs
2800	2380	420	[20, 3]	25	2

Table 6-4 Sub-ANN performances (R^2) for 24 sub-ANN models.

Scenario	R^2		
		lean zone sub-model	shale sub-model
1	Outputs	$[D_d^1, H_d^1, L_d^1]$	NA
	Train	[0.99 0.95 0.96]	
	Test	[0.96 0.92 0.82]	
2	Outputs	$[D_d^1, H_d^1, L_d^1]$	$[D_d^1, H_d^1, L_d^1]$
	Train	[0.96 0.89 0.91]	[0.84 0.83 0.63]
	Test	[0.93 0.80 0.71]	[0.77 0.81 0.16]
3	Outputs	$[D_d^1, H_d^1]$	$[D_d^1, H_d^1, D_d^2, H_d^2]$
	Train	[0.97 0.93]	[0.82 0.68 0.50 0.69]
	Test	[0.95 0.75]	[0.27 0.12 0.23 0.28]
4	Outputs	$[D_d^1, H_d^1]$	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$

	Train	[0.88 0.70]	[0.76 0.55 0.49 0.31 0.32 0.24]
	Test	[0.82 0.52]	[0.55 0.41 0.28 0.10 0.00 0.17]
5	Outputs	$[D_d^1, H_d^1, L_d^2, H_d^2, D_d^2, L_d^2]$	NA
	Train	[0.93 0.81 0.64 0.66 0.78 0.62]	
	Test	[0.90 0.10 0.32 0.33 0.46 0.47]	
6	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2]$	$[D_d^1, H_d^1]$
	Train	[0.83 0.56 0.39 0.49]	[0.80 0.82]
	Test	[0.68 0.32 0.44 0.26]	[0.72 0.76]
7	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2]$	$[D_d^1, H_d^1, D_d^2, H_d^2]$
	Train	[0.83 0.66 0.39 0.44]	[0.92 0.85 0.87]
	Test	[0.67 0.03 0.00 0.24]	[0.89 0.58 0.77]
8	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2]$	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$
	Train	[0.78 0.64 0.55 0.47]	[0.77 0.55 0.40 0.20 0.22 0.17]
	Test	[0.64 0.58 0.21 0.26]	[0.64 0.47 0.23 0.20 0.00 0.00]
9	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$	NA
	Train	[0.89 0.67 0.55 0.49 0.40 0.53]	
	Test	[0.54 0.07 0.39 0.44 0.47 0.46]	
10	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$	$[D_d^1, H_d^1]$
	Train	[0.83 0.51 0.37 0.54 0.26 0.49]	[0.78 0.89]
	Test	[0.50 0.17 0.13 0.17 0.19 0.15]	[0.64 0.80]

11	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$	$[D_d^1, H_d^1, D_d^2, H_d^2]$
	Train	[0.81 0.58 0.43 0.40 0.17 0.27]	[0.72 0.57 0.39 0.37]
	Test	[0.49 0.43 0.40 0.51 0.01 0.37]	[0.36 0.34 0.30 0.35]
12	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$	$[D_d^1, H_d^1, L_d^2, H_d^2, D_d^2, L_d^2]$
	Train	[0.83 0.49 0.27 0.40 0.15 0.20]	[0.68 0.52 0.25 0.28 0.39 0.49]
	Test	[0.63 0.35 0.32 0.21 0.00 0.12]	[0.43 0.53 0.22 0.02 0.00 0.00]
13	Outputs	NA	$[D_d^1, H_d^1, L_d^1]$
	Train		[0.92 0.89 0.72]
	Test		[0.85 0.80 0.22]
14	Outputs	NA	$[D_d^1, H_d^1, L_d^2, H_d^2, D_d^2, L_d^2]$
	Train		[0.84 0.59 0.43 0.48 0.53 0.19]
	Test		[0.89 0.68 0.19 0.18 0.13 0.00]
15	Outputs	NA	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$
	Train		[0.88 0.67 0.54 0.44 0.35 0.13]
	Test		[0.80 0.64 0.40 0.03 0.17 0.21]

Table 6-5 Sub-ANN performances (*MSE*) for 24 sub-ANN models.

Scenario		*MSE* ($\times 10^{-4}$)	
		lean zone sub-model	shale sub-model
1	Outputs	$[D_d^1, H_d^1, L_d^1]$	NA

	Train	[1.04 5.42 3.53]	
	Test	[4.50 10.15 14.37]	
2	Outputs	$[D_d^1, H_d^1, L_d^1]$	$[D_d^1, H_d^1, L_d^1]$
	Train	[5.52 13.27 6.90]	[20.47 36.07 13.30]
	Test	[6.47 22.78 29.59]	[23.88 37.71 30.31]
3	Outputs	$[D_d^1, H_d^1]$	$[D_d^1, H_d^1, D_d^2, H_d^2]$
	Train	[4.52 8.92]	[17.73 46.36 38.93 609.80]
	Test	[9.03 40.33]	[51.42 109.70 59.25 129.30]
4	Outputs	$[D_d^1, H_d^1]$	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$
	Train	[12.41 3.12]	[21.11 59.73 37.65 127.7 48.50 129.5]
	Test	[15.23 5.65]	[35.81 66.77 75.16 128.1 49.32 132.3]
5	Outputs	$[D_d^1, H_d^1, L_d^2, H_d^2, D_d^2, L_d^2]$	NA
	Train	[5.99 20.99 28.44 18.12 26.32 29.61]	
	Test	[8.91 100.00 52.68 27.75 5.78 3.59]	
6	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2]$	$[D_d^1, H_d^1]$
	Train	[15.00 46.25 28.78 63.28]	[23.50 33.60]
	Test	[16.08 55.91 39.85 87.06]	[43.86 47.12]
7	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2]$	$[D_d^1, H_d^1, D_d^2, H_d^2]$
	Train	[24.43 54.20 43.84 75.45]	[10.05 18.13 9.65]
	Test	[39.67 132.70 60.29 105.20]	[10.45 48.25 23.28]

8	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2]$	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$
	Train	[23.70 41.68 32.00 60.51]	[17.91 40.78 34.09 11.28 38.41 136.20]
	Test	[30.04 59.75 69.64 110.50]	[28.62 32.78 49.27 132.80 46.58 137.20]
9	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$	NA
	Train	[8.00 26.76 27.18 61.19 31.66 52.16]	
	Test	[1.71 55.22 40.04 70.42 34.640 59.55]	
10	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$	$[D_d^1, H_d^1]$
	Train	[13.27 44.44 51.16 65.34 44.79 52.60]	[26.77 22.39]
	Test	[36.47 68.32 66.24 120.60 40.23 69.81]	[38.46 37.39]
11	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$	$[D_d^1, H_d^1, D_d^2, H_d^2]$
	Train	[17.82 45.84 43.47 78.49 46.90 83.22]	[29.52 50.47 42.75 103.20]
	Test	[36.52 51.40 33.63 65.24 72.29 79.54]	[62.97 84.93 30.36 106.20]
12	Outputs	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$	$[D_d^1, H_d^1, L_d^2, H_d^2, D_d^2, L_d^2]$
	Train	[16.73 51.43 49.04 98.65 47.77 84.13]	[24.58 36.14 510.40 104.60 39.91 72.77]
	Test	[40.77 66.71 65.77 138.60 38.79 84.37]	[45.77 41.46 38.23 108.00 59.83 155.50]
13	Outputs	NA	$[D_d^1, H_d^1, L_d^1]$
	Train		[9.690 16.60 9.68]
	Test		[18.91 41.57 37.07]
14	Outputs	NA	$[D_d^1, H_d^1, L_d^2, H_d^2, D_d^2, L_d^2]$
	Train		[12.45 46.91 20.57 32.45 90.48 35.59]

	Test		[8.58 36.96 25.53 41.48 141.20 49.28]
15	Outputs	NA	$[D_d^1, H_d^1, D_d^2, H_d^2, D_d^3, H_d^3]$
	Train		[8.05 33.17 23.69 79.64 30.33 138.60]
	Test		[17.03 45.99 35.56 110.60 52.06 137.90]

Figures

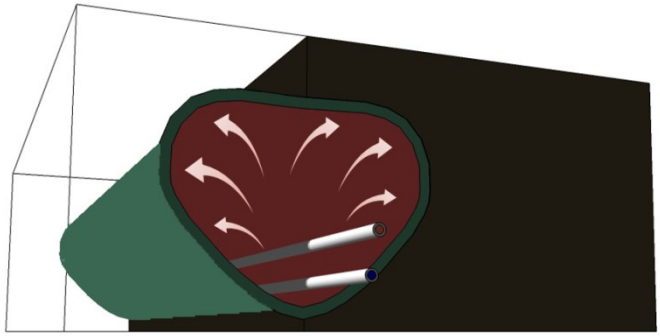


Fig. 6-1 A schematic of the typical SAGD operation in 3D view.

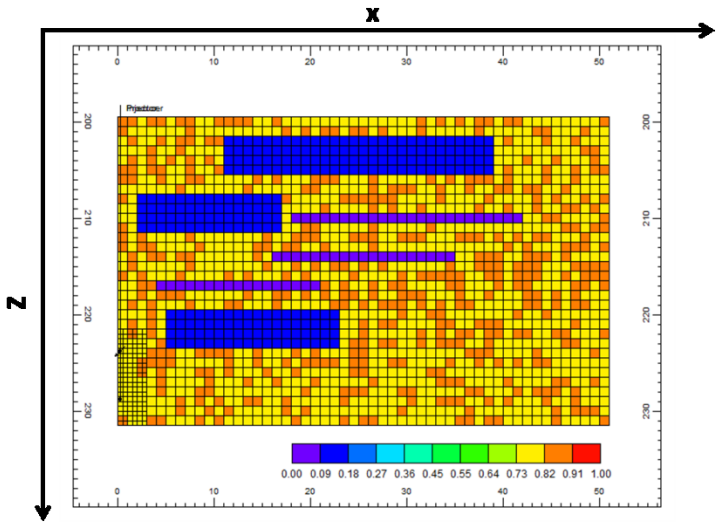


Fig. 6-2 Oil saturation (S_o) distribution for one of the models in the study: the background represents oil sands, the thin purple layers represent shale barriers (S_o = 0.0); the thick blue layers represent lean zones (S_o = 0.1). The well pair is located at the left boundary of the reservoir. The injector is located at a distance of 5 m above the producer.

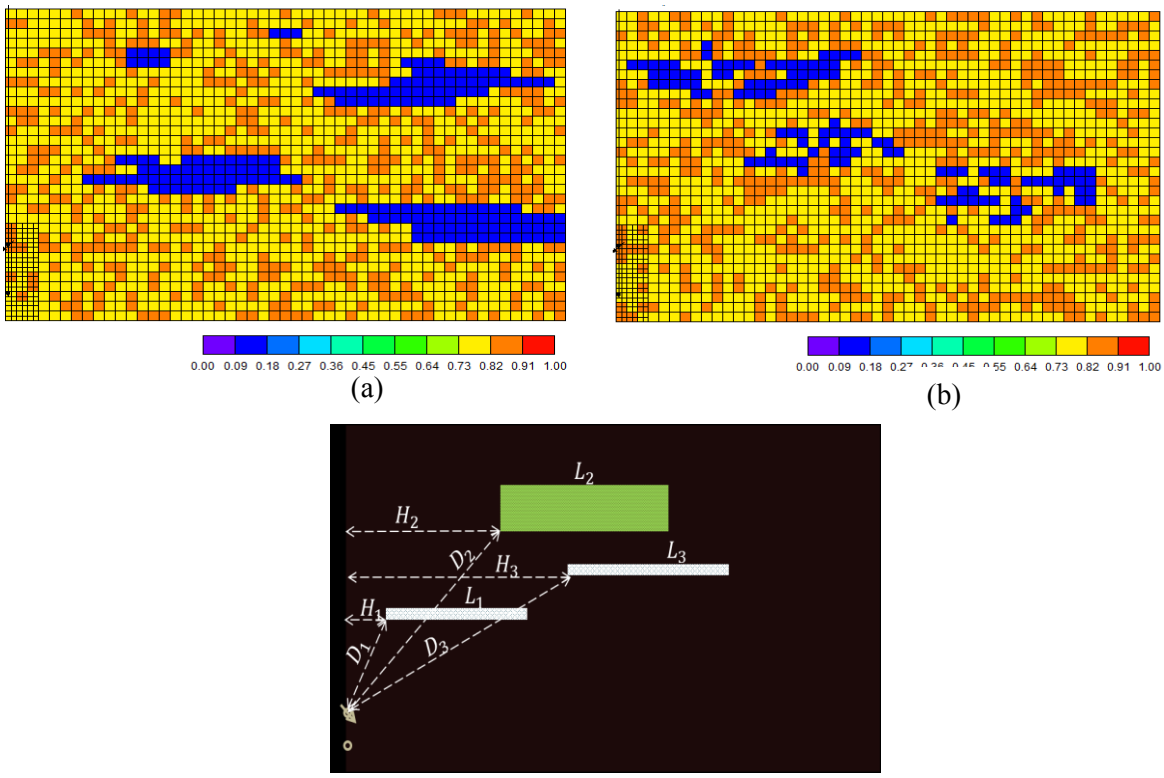


Fig. 6-3 Formulation of heterogeneous features: (a and b) – examples where the number of lean zone regions (N_l) = 3; (c) – geometrical properties of each heterogeneous feature.

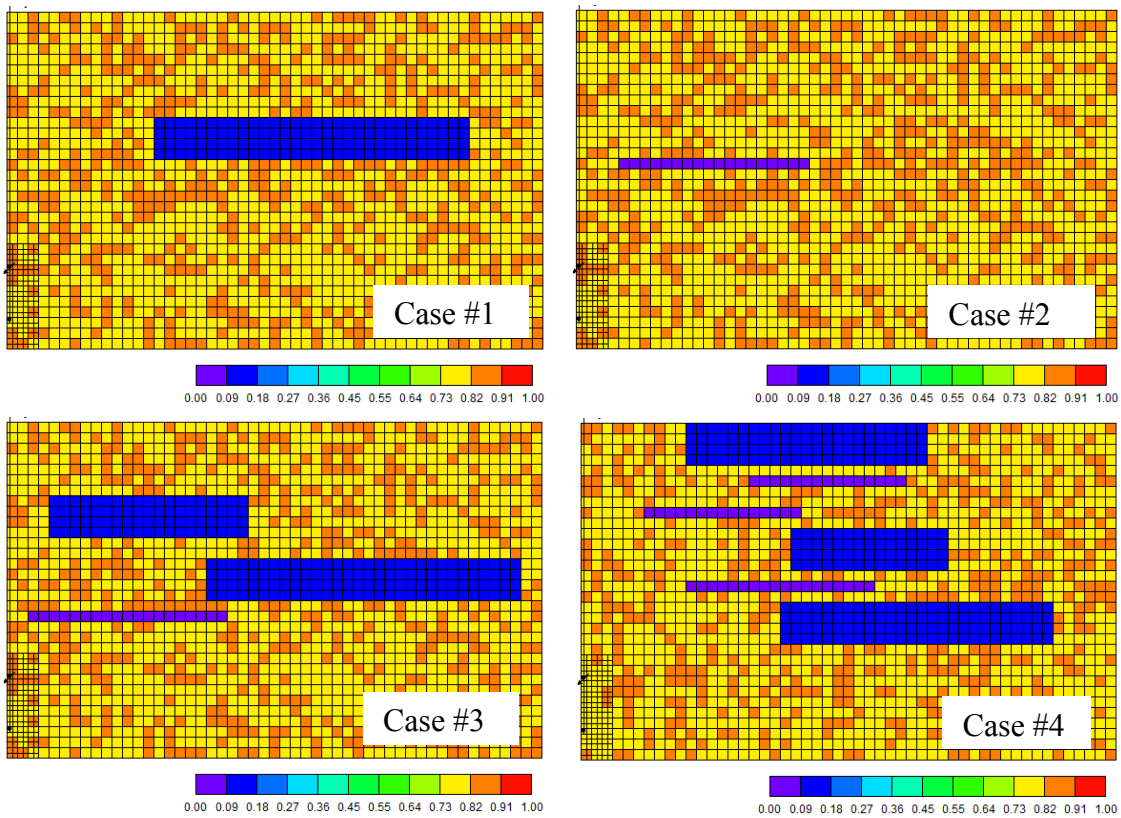


Fig. 6-4 Four examples of heterogeneous models (colorscale denotes S_o): blue features represent lean zones, while purple features represent shale barriers.

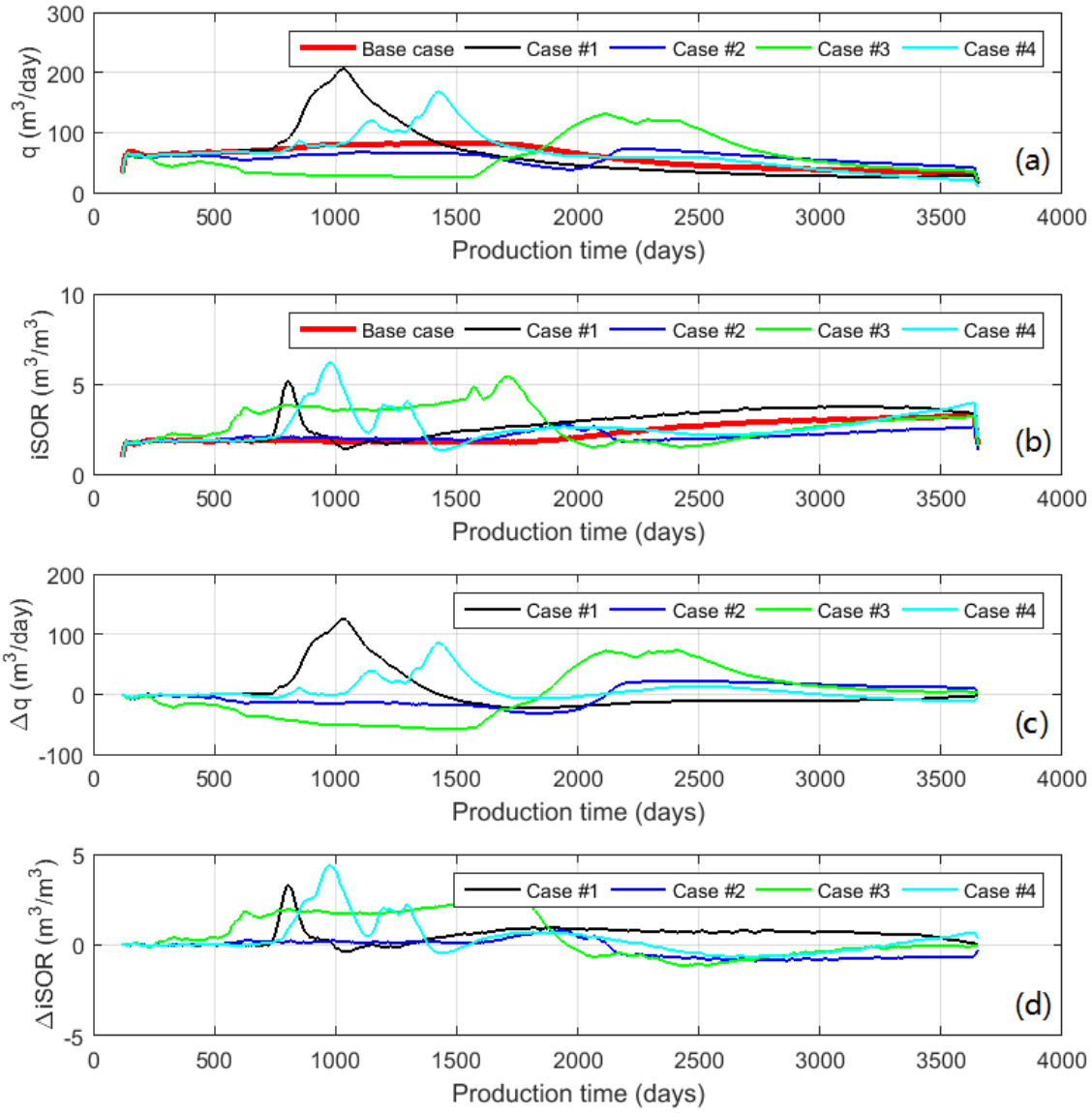


Fig. 6-5 Comparison of production profiles (q, $iSOR$, Δq, and $\Delta iSOR$) between the cases #1-4 in Fig. 6-3 and the base homogeneous case.

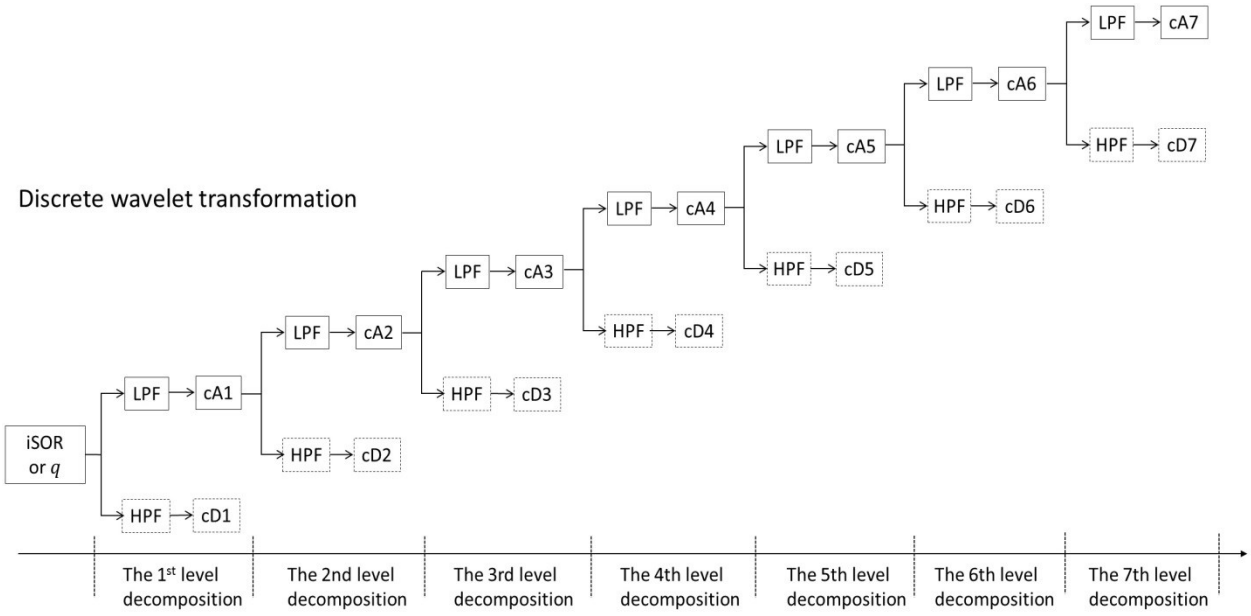


Fig. 6-6 Application of DWT in time-series decomposition

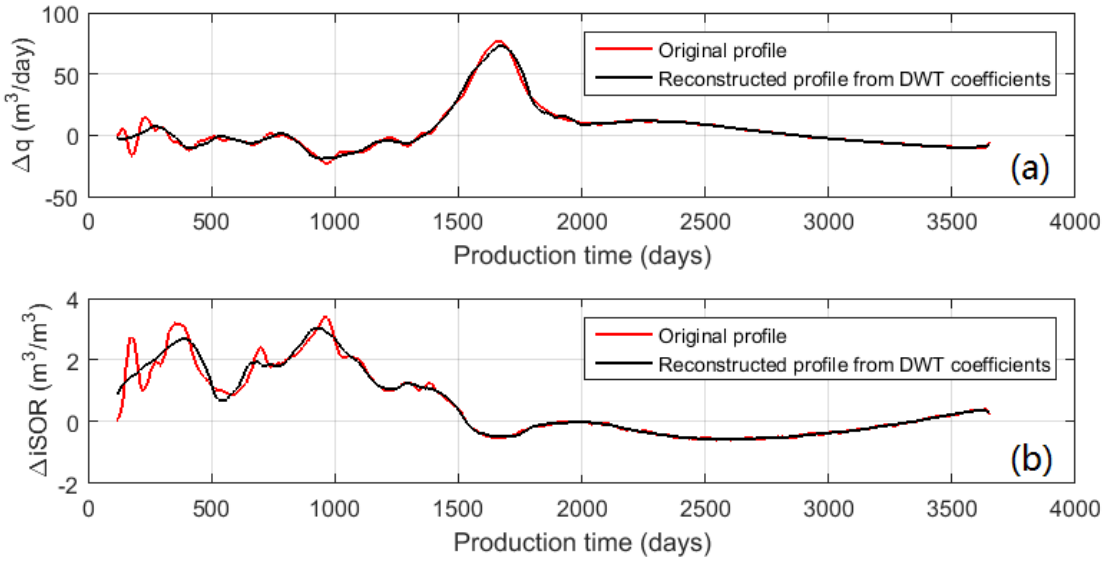


Fig. 6-7 Original and reconstructed Δq and Δ$iSOR$ profiles for the heterogeneous model shown in Fig. 6-2.

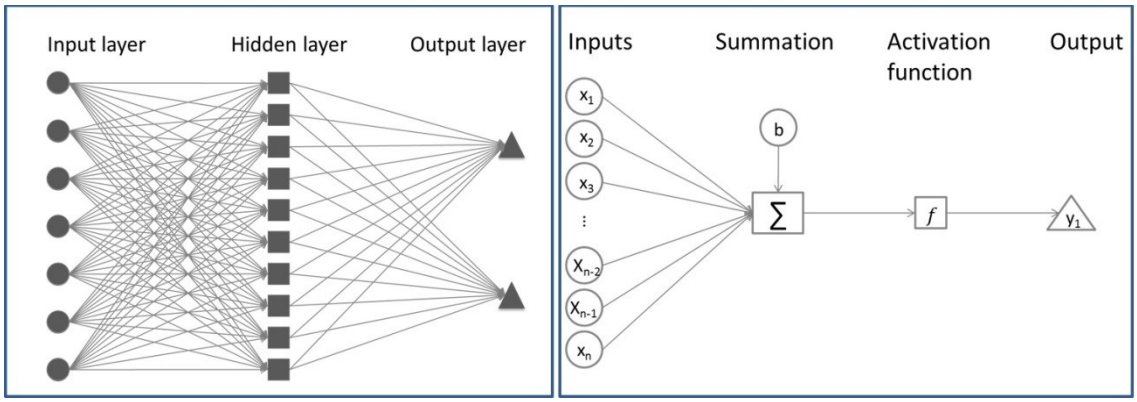


Fig. 6-8 A typical ANN architecture: left – a single hidden layer configuration with 7 input variables, 2 output variables, and 10 hidden nodes; right – schematic of the forward ANN computation.

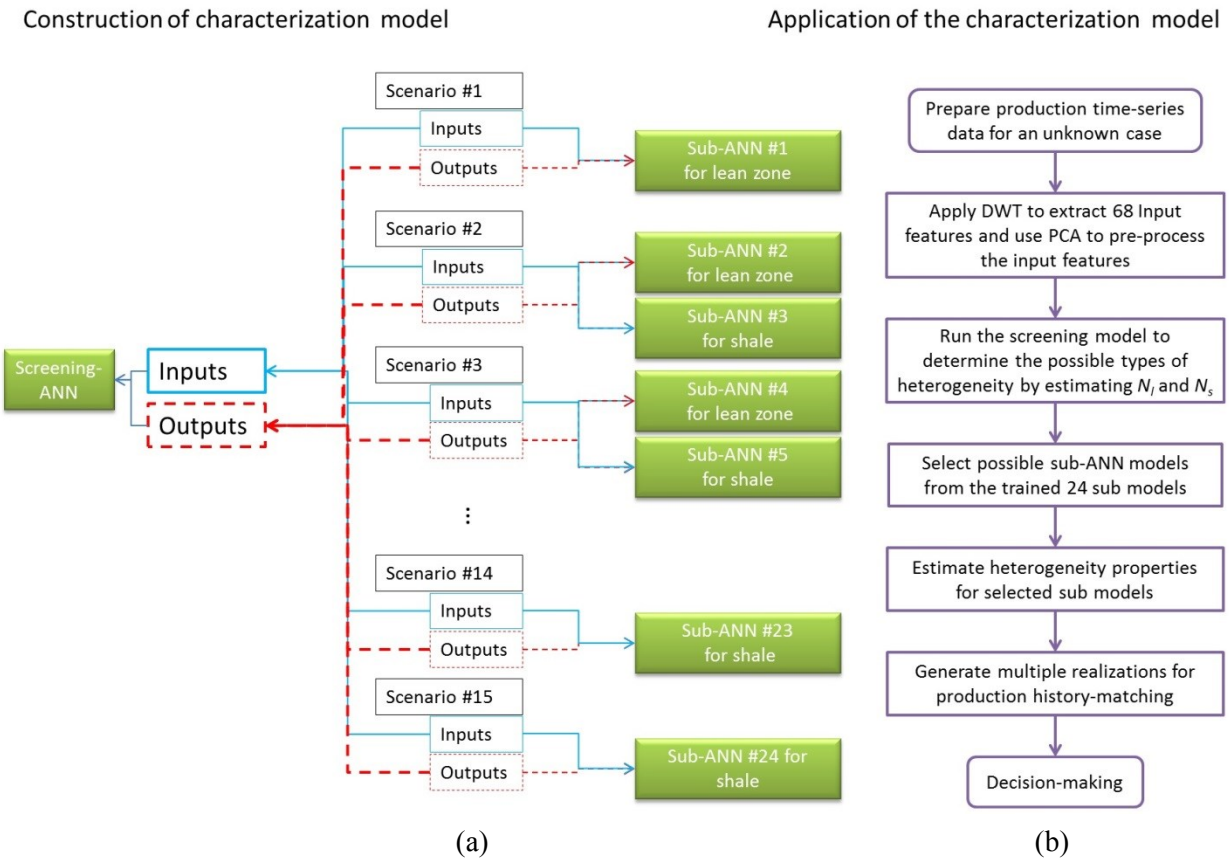


Fig. 6-9 Illustration of the construction (left) and application (right) of the proposed two-level data-driven model for heterogeneity characterization.

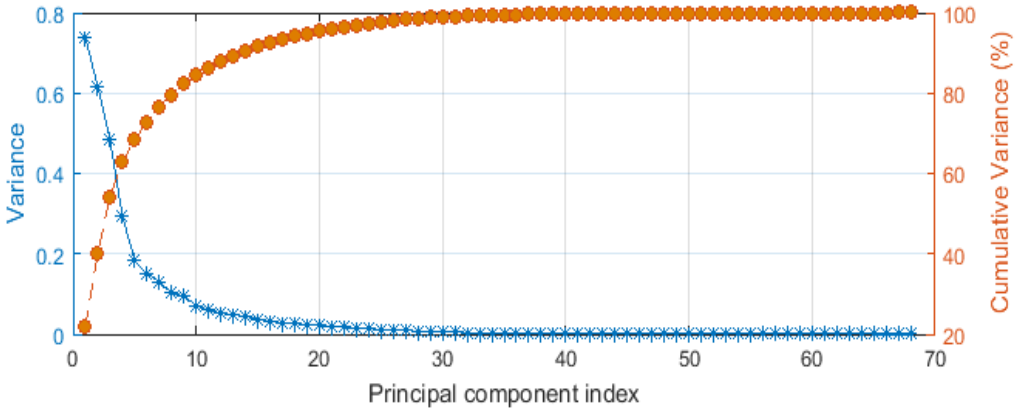


Fig. 6-10 Principal component analysis: variance plot.

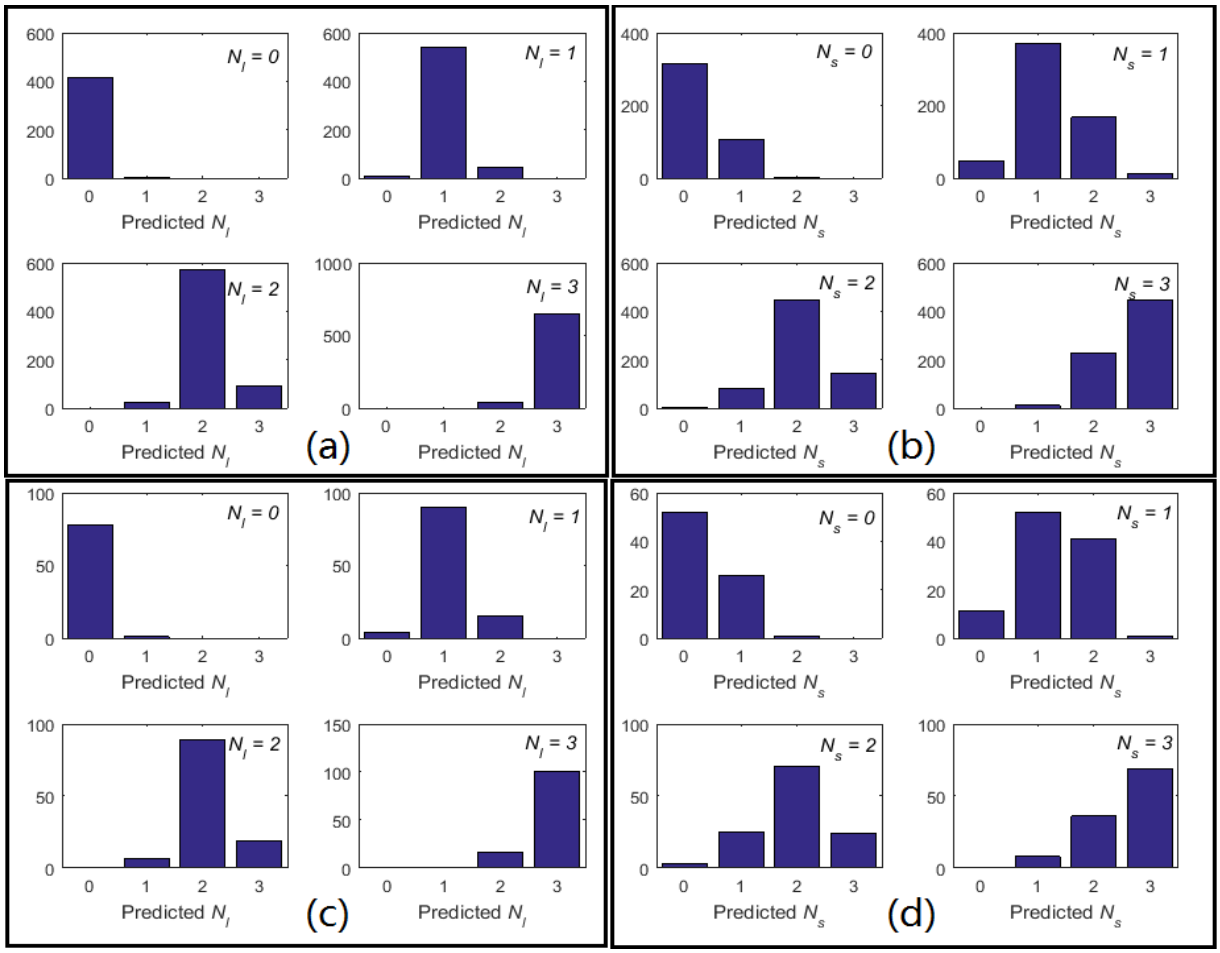


Fig. 6-11 Distributions of the predicted N_l and N_l using the screening-ANN model and the corresponding targets: top – training cases; bottom row – testing cases.

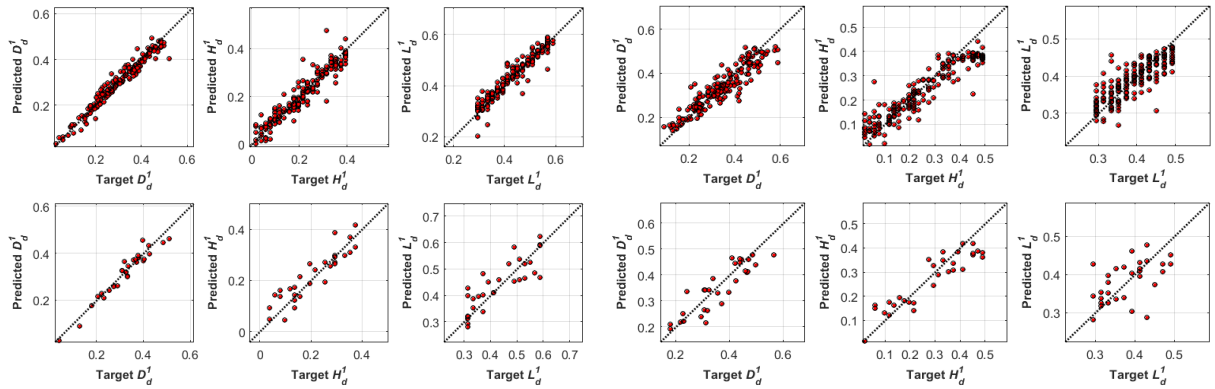


Fig. 6-12 ANN estimation performance for scenario #2 ($N_l = 1$, $N_s = 1$): (a) – lean zone predictions; (b) – shale barrier predictions. For each subplot: top row – training cases; bottom row – testing cases.

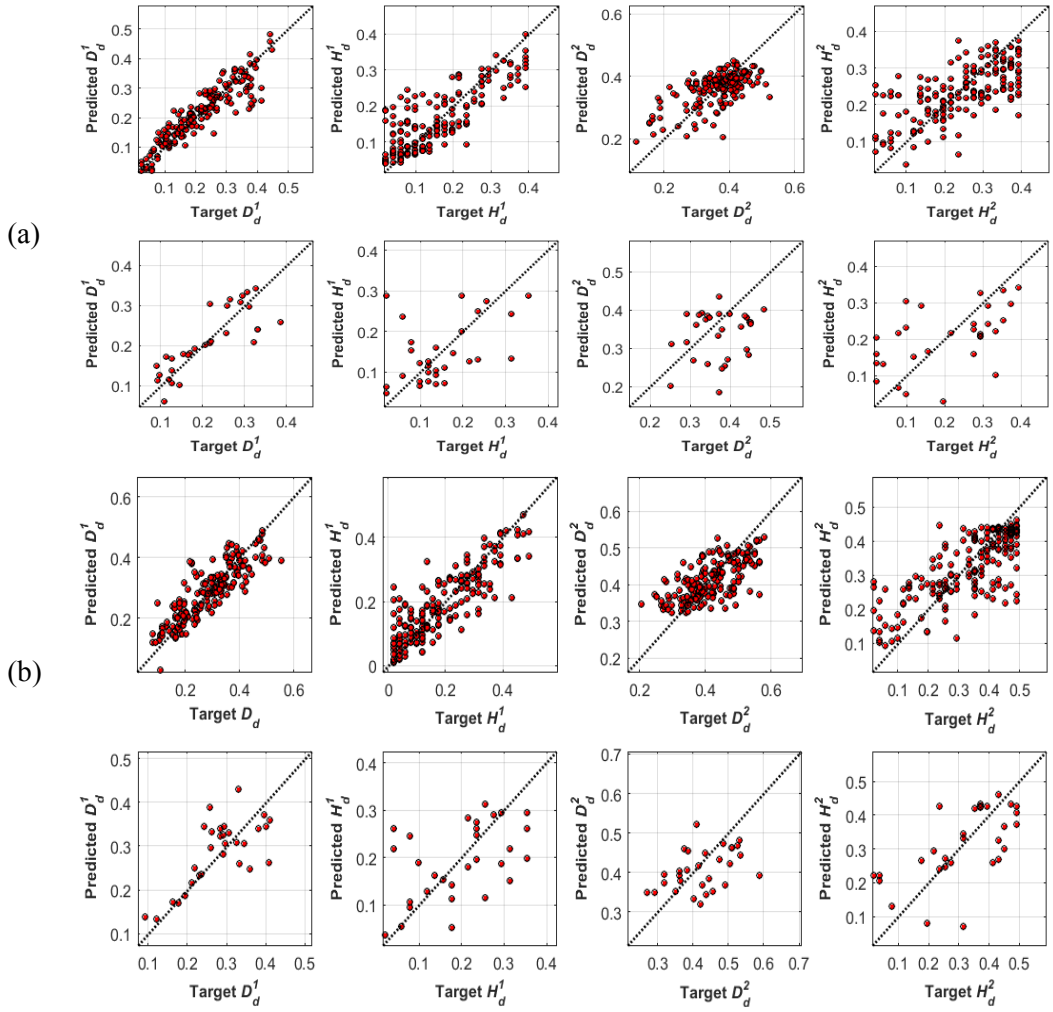


Fig. 6-13 ANN estimation performance for scenario #7 ($N_l = 2$, $N_s = 2$): (a) – lean zone

predictions; (b) – shale barrier predictions. For each subplot: top row – training cases; bottom row – testing cases.

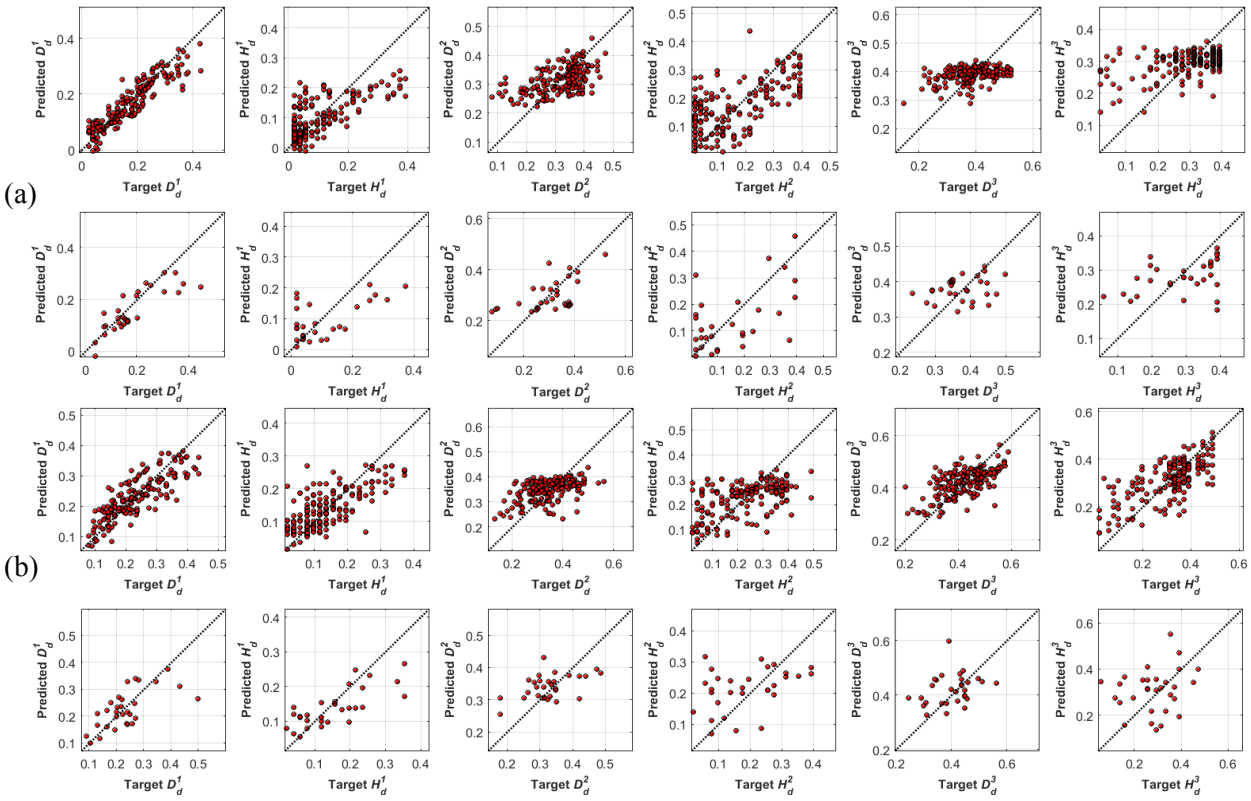


Fig. 6-14 ANN estimation performance for scenario #12 ($N_l = 3$, $N_s = 3$): (a) – lean zone predictions; (b) – shale barrier predictions. For each subplot: top row– training cases; bottom row – testing cases.

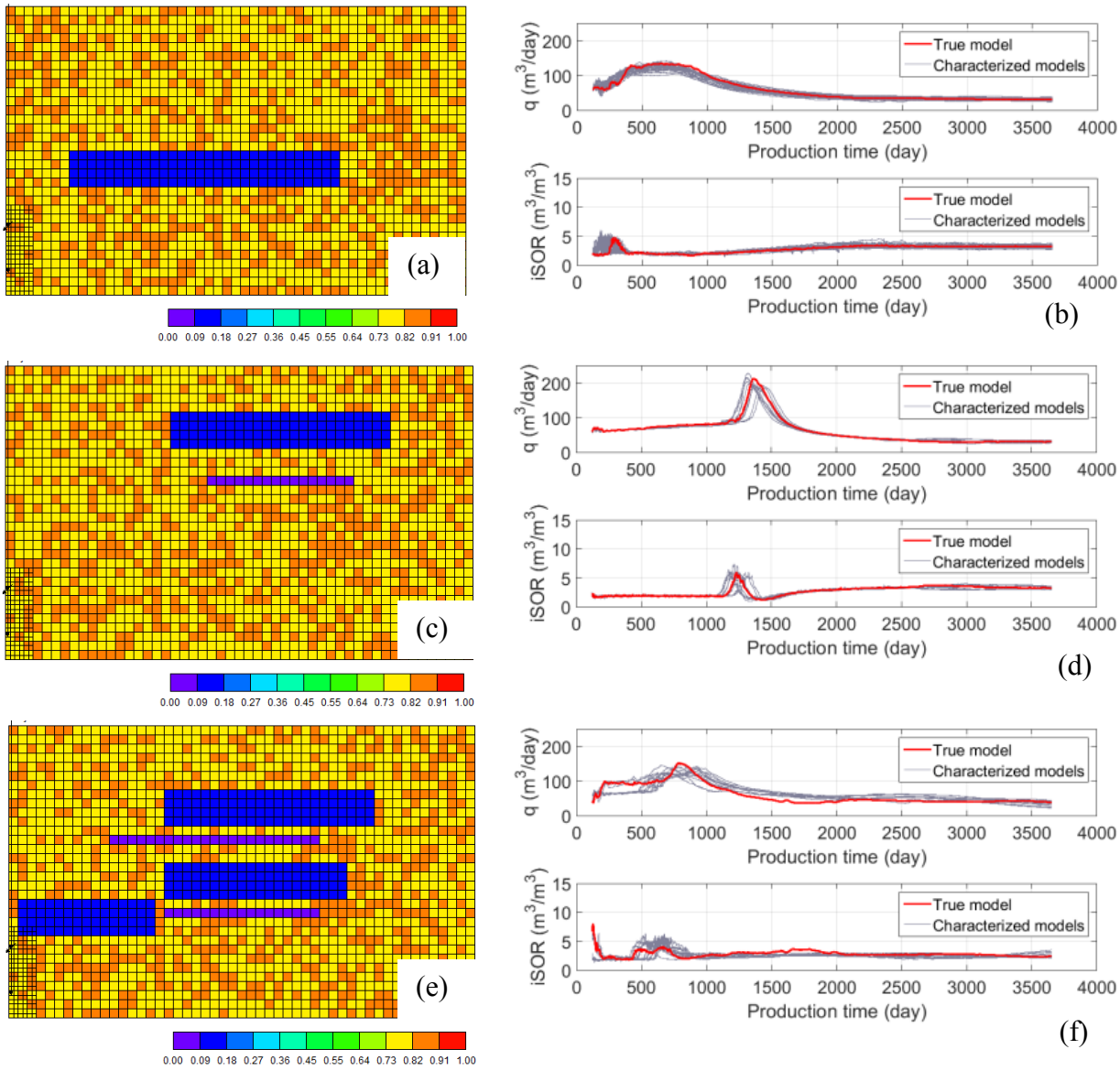


Fig. 6-15 Comparison of production profiles between the true model and realizations of the characterized model for three cases with simply-shaped heterogeneities: a, c, d – heterogeneity configuration of the true case; (b, d, f) – production profiles.

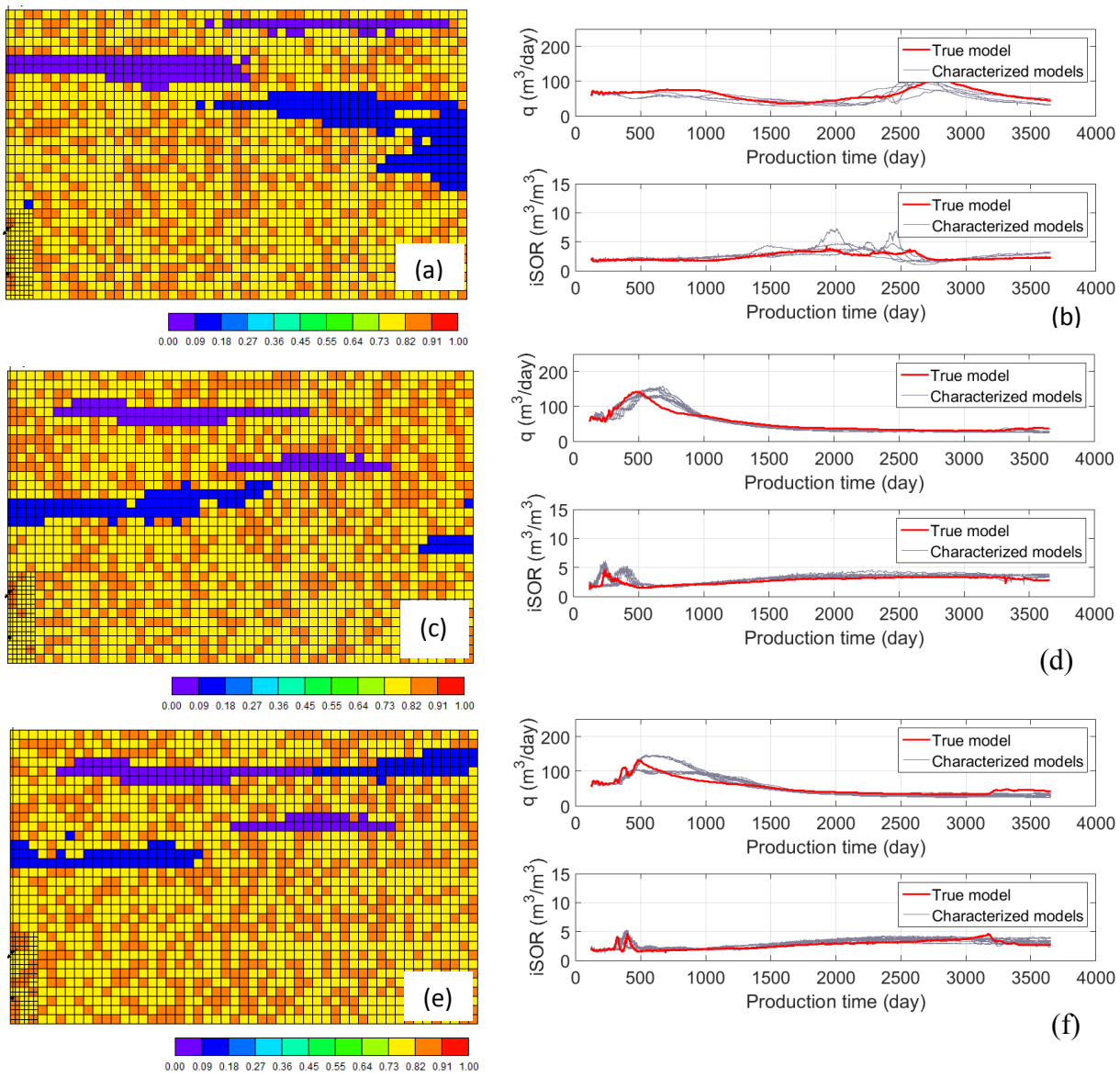


Fig. 6-16 Comparison of production profiles between the true model and realizations of the characterized model for three cases with irregularly-shaped heterogeneities: (a, c, d) – heterogeneity configuration of the true case; (b, d, f) – production profiles.

Chapter 7 Critical Analysis

Chapter Overview

This chapter presents a critical analysis of the proposed methodologies and provides various quantitative comparisons between the developed models and a few other conventional techniques. First, both linear and nonlinear regression models are tested to predict SAGD production using the data presented in chapter 4. The differences in forecast performances with the trained data-driven models are elaborated. Next, a detailed comparison is discussed to illustrate the benefits of the proposed data-driven models for conventional SAGD operation. Third, the potential benefits for integrating data-driven models in practical reservoir management routines are outlined. Finally, the potential limitations of the proposed data-driven models are also presented.

7.1 Comparison with Other Linear and Nonlinear Regression Techniques

In order to further test the capacity of the constructed data-driven models in this thesis, it is would be important to compare the results obtained from the proposed data-driven models to that obtained from other simpler techniques. Therefore, two commonly-used linear and nonlinear regression techniques are also applied to infer the unknown correlation between system input and output variables using the same datasets in Chapters 3-4.

The linear regression model is built using multivariate linear regression algorithm; while the nonlinear regression model is built using the response surface algorithm. In this chapter, the dataset (consisting of 153 data samples) extracted in chapter 4 (as shown in **Table A-1**) is used to build the linear and nonlinear regression models for the comparisons. In order to obtain a comprehensive comparison, the pre-processed dataset via PCA and k-means clustering analysis is used to construct these regression models. Therefore, the input vector contains 6 *PSs* while the output vector includes TISOR and COP, respectively. Same to that in chapter 4, the entire dataset is divided into two groups as training (123 samples) and testing (30 samples) subset, to construct

and test the performance of the models, respectively. The samples in the training and testing subsets are same as that in chapter 4.

7.1.2 Linear Regression Model

The simplest regression function is a linear model, which is capable of building a linear relationship between the predictor (or input) and the response (or output) variables. For a problem with one response and a total number of p predictor variables, the linear regression model can be expressed as:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_p \cdot X_p + e \quad \cdots\cdots (7\text{-}1)$$

where Y is the response or target of the regression problem; X denotes the predictor variable; β represents linear coefficient associated with the predictor X; β_0 refers to a constant coefficient; e is an error term. Least-squares estimation techniques can be used to calculate the corresponding regression coefficients (Montgomery et al., 2012). Similarly, for a problem with λ responses and p predictor variables, the multivariate linear regression model can be formulated as:

$$\begin{aligned} Y^1 &= \beta_0^1 + \beta_1^1 \cdot X_1 + \beta_2^1 \cdot X_2 + \cdots + \beta_p^1 \cdot X_p + e^1 \\ Y^2 &= \beta_0^2 + \beta_1^2 \cdot X_1 + \beta_2^2 \cdot X_2 + \cdots + \beta_p^2 \cdot X_p + e^2 \\ &\vdots \\ Y^\gamma &= \beta_0^\lambda + \beta_1^\lambda \cdot X_1 + \beta_2^\lambda \cdot X_2 + \cdots + \beta_p^\gamma \cdot X_p + e^\gamma \end{aligned} \quad \ldots\ldots\ldots\ldots (7\text{-}2)$$

It should be noted that the same predictor variables are used to construct λ linear models for the λ responses. Many techniques can be employed to calculate the regression coefficients, such as maximum likelihood estimation and covariance-weighted least squares estimation. In this chapter, the "mvregress" function in Matlab™ (MathWorks, 2017) is used to build the linear regression model and compute the corresponding coefficients.

It is easy to know that $p = 6$ and $\lambda = 2$ since the number of predictor variables (i.e., *PSs*) is 6 while the number of response (i.e, COP and TISOR) is 2 in this chapter, respectively. The final obtained coefficients are presented in **Table 7-1** and **Table 7-2** for TISOR and COP, respectively.

The predictions of SAGD production performances using the constructed linear models are shown in **Fig. 7-1**. The overall estimations of TISOR and COP are acceptable as the most points in the cross-plots follow the 45-degree splitting line in both training and testing cases. However, the predictions using the linear regression model are apparently inferior to that using the proposed data-driven models as shown **Fig. 4-11**. It illustrates the nonlinearity exhibited in the extracted inputs and outputs cannot be captured by applying the simple linear regression model. Next, the nonlinear regression models are tested to forecast SAGD production using the same dataset.

7.1.2 Nonlinear Regression Model

In this chapter, the nonlinear regression modes are constructed using the response surface methodology, which consists of many statistical and mathematical techniques for the development, improvement, and optimization of various processes (Myers et al., 2016). Basically, the response surface is a kind of polynomial regression model (Montgomery et al., 2012). The response surface model can be grouped into three categories: the main effect model, the first-order model with interaction, and the second-order model. The main effect model considers only the linear effects from the predictor variables while the first-order model with interaction introduces curvature to the model. The second-order model can be employed for the cases with strong curvature in the true surface (Myers et al., 2016). Due to its good features, such as flexibility and simple calculation of the parameters, the second-order response surface model is widely used in many fields of study. In this chapter, the nonlinear models are built using the second-order response surface methodologies.

Similarly, for a problem with λ responses and p predictor variables, the second-order response surface model is given by:

$$
\begin{aligned}
&Y^1 = \alpha_0^1 + \alpha_1^1 \cdot X_1 + \alpha_2^1 \cdot X_2 + \cdots + \alpha_p^1 \cdot X_p \\
&+ \alpha_{12}^1 \cdot X_1 \cdot X_2 + \alpha_{13}^1 \cdot X_1 \cdot X_3 + \cdots + \alpha_{(p-1)p}^1 \cdot X_{p-1} \cdot X_p \\
&+ \alpha_{11}^1 \cdot X_1^2 + \alpha_{22}^1 \cdot X_2^2 + \cdots + \alpha_{pp}^1 \cdot X_p^2 + \\
&+ e^1 \\
&Y^2 = \alpha_0^2 + \alpha_1^2 \cdot X_1 + \alpha_2^2 \cdot X_2 + \cdots + \alpha_p^2 \cdot X_p \\
&+ \alpha_{12}^2 \cdot X_1 \cdot X_2 + \alpha_{13}^2 \cdot X_1 \cdot X_3 + \cdots + \alpha_{(p-1)p}^2 \cdot X_{p-1} \cdot X_p \\
&+ \alpha_{11}^2 \cdot X_1^2 + \alpha_{22}^2 \cdot X_2^2 + \cdots + \alpha_{pp}^2 \cdot X_p^2 + \\
&+ e^2 \\
&\vdots \\
&Y^\lambda = \alpha_0^\lambda + \alpha_1^\lambda \cdot X_1 + \alpha_2^\lambda \cdot X_2 + \cdots + \alpha_p^\lambda \cdot X_p \\
&+ \alpha_{12}^\lambda \cdot X_1 \cdot X_2 + \alpha_{13}^\lambda \cdot X_1 \cdot X_3 + \cdots + \alpha_{(p-1)p}^\lambda \cdot X_{p-1} \cdot X_p \\
&+ \alpha_{11}^\lambda \cdot X_1^2 + \alpha_{22}^\lambda \cdot X_2^2 + \cdots + \alpha_{pp}^\lambda \cdot X_p^2 + \\
&+ e^\lambda
\end{aligned}
\qquad \text{..... (7-3)}
$$

for each response Y^λ in the above, α in the first row equation represent the linear effect terms; α in the second row are the interaction effect terms; α in the third row represents the quadratic effects terms. Many techniques can be used to solve the coefficients of a response surface function, e.g., least square technique. Please refer to Myers et al. (2016) for the detailed calculations of the coefficients. In this chapter, the interactive response surface modeling toolbox in Matlab™ (MathWorks, 2017) is used to compute the corresponding coefficients for the nonlinear models. Finally, the calculated coefficients are shown in **Table 7-3** and **Table 7-4** for TISOR and COP, respectively. The cross-plots between the predicted TIOSR and COP using the nonlinear regression models and their corresponding targets are shown in **Fig. 7-2**. An improvement in the prediction performance is observed when compared with the linear models, which demonstrates the capacity of the response surface models for inference of the nonlinear and complex relationship among system variables. However, in term of estimation performance, the proposed ANN models shown in chapter 4 outperform the nonlinear models by providing higher R^2 and lower *MSE* for both clusters.

This results presented in this chapter illustrate that it is challenging to apply the conventional linear and nonlinear techniques for capturing the relationship between reservoir/operational parameters and SAGD production performance. In contrast, data-driven

models developed in chapter 4 have demonstrated great potentials to identify and approximate the highly nonlinear, complex, and uncertain relationship among system variables with less prior knowledge of the system. The comparisons in this chapter further demonstrate the capacity of data-driven models for SAGD production prediction. If the number of predictors and responses increases or the relationships between the predictor and response variables becomes more complex, e.g., the inference of reservoir heterogeneities from production time-series data in chapter 5 and 6, the conventional linear and nonlinear techniques would provide inferior results.

7.2 Potential Advantages of Data-Driven Modeling

The potential benefits of the proposed data-driven models, in comparison to conventional workflows of flow simulation and inverse history-matching, for SAGD analysis can be summarized as:

a) Significant reduction in computational time. For instance, the ANN training process for group #2 for multiple shale barriers case in chapter 5 is only 1.048 seconds using the 'Neural Network Toolbox™ (Beale et al., 1992) in MATLAB R2015a; while the corresponding testing process use the trained ANN is only 0.183 seconds. However, a single forward simulation process would require 3 minutes and 6 seconds to obtain the final production profiles using STARS (CMG, 2015). It should be noted that the dataset contains 199 samples. The overall forward simulation for these cases requires 10 hours 16 minutes and 54 seconds, which is 30068 times slower for than the data-driven model. If 3D reservoir models are implemented, the simulation of a single case takes several hours. The detailed history-matching processes would be much more computationally-expensive as they requires to run a large number of models, continue updating the unknown parameters at each step, and solve the complex inverse problems using different algorithms. All the aforementioned computation are performed using a personal computer with an Intel(R) Core (TM) i7-4770 CPU (3.4 GHz) and 12 GB of RAM.

b) Fewer storage requirements for simulation results. The same multiple shale barriers case presented in chapter 5 is compared here the illustration. The overall space needed to store the data for training and testing the ANN model is only 33.4 MB. In contrast, it needs

requires 22288 MB to store simulation results for 199 cases. If 3D reservoir models or more samples are used, it would increase the storage space significantly.

c) No need for the explicit knowledge of the complex system. In order to apply conventional simulation and characterization workflows, the explicit knowledge of the physical system is indispensable, such as the mass, energy, and moment balance equations used in conventional reservoir engineering. Data-driven modeling techniques, as presented in chapters 3 to 6 don't highly depend on such prior information. Actually, data-driven models just analyze the dataset and infer the hidden relationship between system input and output variables.

d) Few assumptions and simplifications regarding the governing equations. In order to construct conventional simulation models, many assumptions and simplifications are involved, such as simple well models and discretization of the PDEs. Data-driven models, by contrast, can avoid these assumptions and simplifications.

e) The reduced dependency on extensive data to build models. Many data are needed to in conventional approaches, such as relative permeability curve, PVT data, and reservoir initial conditions. Construction of reservoir models without such data is impossible. However, it is challenging to obtain all these parameters due to various practical challenges and difficulties. Although such information is missing, data-driven models can still be employed for SAGD production prediction and heterogeneities characterization in this thesis. Most importantly, the performances of data-driven models are promising. If more information becomes available, the trained data-driven models can be easily be updated.

7.3 Integration of Data-Driven models for Practical Reservoir Management Routines

This section provides several suggestions of how the results from data-driven models may impact practical operational reservoir management routines.

a) A set of important variables from field data for SAGD production performance in heterogeneous reservoirs can be identified following the procedures developed in chapters 3 and 4, e.g., porosity, shale index, the effective number of wells, etc. Besides

cumulative oil production, the duration over which the monthly average steam-to-oil ratio exceeds a particular threshold is also defined to comprehensively describe SAGD production performance. These parameters can provide crucial insights to optimize SAGD production strategies.

b) Once several important parameters are computed, the trained model shown in chapters 3 and 4 can be employed to provide a fast and reliable forecast of SAGD production. It does not require extensive data to construct complex reservoir models and solve the governing equations. It is particularly useful to obtain an initial approximate for new wells in the same field or the wells in different fields with similar reservoir parameters.

c) The uncertainty analysis results presented in chapters 3 and 4 comprehensively investigate three types of uncertainties in the application of data-driven modeling techniques in reservoir engineering problems. It offers strategies and guidance to minimize such uncertainties and to improve model accuracy for the future applications of data-driven models in reservoir management workflows.

d) The relationships between SAGD production profiles and reservoir heterogeneities parameters are investigated in chapters 5 and 6. Such correlations would enhance the ability to infer the presence of reservoir heterogeneities (shale barriers and lean zones) from production time-series data. It offers a viable and complementary alternative to the conventional history-matching characterization routines. Similarly, they generate reliable distributions of shale barriers and lean zones in the formation quickly once the production profiles are given. It would greatly benefit reservoir management and decision-making routines by providing an ensemble of heterogeneities distributions. This information can facilitate optimize SAGD operations and change production strategies (e.g., reduce steam injection, drill infilled wells, etc.) to improve recovery factor as well as the steam utilization efficiency

e) The generated multiple heterogeneous models can be regarded as reliable initial guesses for conventional history-matching routines, improving the convergence speed and robustness of the overall process.

7.4 Limitations of the Developed Data-Driven Models

The limitations or disadvantages of the developed data-driven models can be summarized as:

a) Model overfitting and extrapolation: Overfitting and extrapolation are two most common problems for any data-driven modeling techniques. For a given data-driven model, overfitting often leads to the training performance to be much better than that of the testing performance. As shown in **Fig. 3-8**, R^2 in training dataset is equal to 1 while R^2 in testing dataset only 0.14 using the same model. To alleviate overfitting problem, an early-stopping technique is applied in this thesis. Model extrapolation refers to the scenario where values of the input attributes are beyond the ranges encompassed by the training data set. Therefore, it is recommended to check model extrapolation problem before application of the data-driven models.

b) Insufficient data samples: Ideally, a large number of training samples are needed to train the model in order to obtain a set of optimum model parameters. However, assembling a large dataset is not an easy task due to many practical problems. For instance, despite there are many wells in three SAGD projects in chapter 3, only 71 samples are extracted to build the data-driven models. In certain applications, it is possible to construct more synthetic models to span the parameter space and thus get more robust data-driven models in chapters 5 and 6. However, considering the high computational- and storage-cost, a limited number of samples are generated. However, the trained data-driven models using only several hundred cases are quite reliable.

c) Lack of universal data-driven modes: Same to conventional simulation and history-matching workflows that cannot be applied to solve all the problems without any modifications, certain domain knowledge is required to utilize the data-driven models presented in thesis. For instance, the trained SAGD production analysis models in chapter 3 based on three SGAD projects cannot be directly applied to the cases in chapter 4 due to the differences in the number of input variables and differences in their ranges. The data-driven models need to be trained again based on the inputs and output parameters obtained from these fields. Another example is that the lean zone and shale barrier characterization models trained in chapter 6 cannot be employed to infer heterogeneities in other oil sands deposits. That is because the reservoir parameters

used to construct synthetic SAGD models are based on Athabasca oil sands. However, it should be noted that the proposed workflow is generic and can be extended to other projects.

7.5 Reference

Beale, M. H., Hagan, M. T., & Demuth, H. B. (1992). *Neural network toolbox™ user's guide*. The Mathworks Inc.

CMG, 2015. STARS: Users' Guide, advanced processes & thermal reservoir simulator (Version 2015), Calgary, Alberta, Canada: Computer 21 Modeling Group Ltd.

MathWorks (2017). *Statistics and Machine Learning Toolbox User's Guide*. Retrieved December 28, 2017 from https://www.mathworks.com/help/releases/R2017b/pdf_doc/stats/stats.pdf.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*, John Wiley & Sons.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: Process and product optimization using designed experiments*, John Wiley & Sons.

Tables

Table 7-1 Coefficients of the linear regression model for two cluster: TISOR.

Cluster index	β_0^1	β_1^1	β_2^1	β_3^1	β_4^1	β_5^1	β_6^1
Cluster 1	-0.4542	0.1239	-0.4422	0.4951	-0.0246	-0.1855	-0.4669
Cluster 2	-0.5607	0.0845	-0.2928	0.3993	-0.1038	-0.1638	-0.1471

Table 7-2 Coefficients of the linear regression models for two clusters: COP.

Cluster index	β_0^2	β_1^2	β_2^2	β_3^2	β_4^2	β_5^2	β_6^2
Cluster 1	-0.6906	0.2088	-0.3398	0.3377	0.0602	0.5561	0.0072
Cluster 2	-0.5998	0.0148	-0.3817	0.3134	0.1186	0.5097	-0.2170

Table 7-3 Coefficients of the nonlinear regression models for two clusters: TISOR.

Cluster index	α_0^1	α_1^1	α_2^1	α_3^1	α_4^1	α_5^1	α_6^1
Cluster 1	-0.7668	0.8825	-0.5712	0.5866	-0.1196	-0.7682	-0.6338
	α_{12}^1	α_{13}^1	α_{14}^1	α_{15}^1	α_{16}^1	α_{23}^1	α_{24}^1
	0.2025	-0.1118	0.0492	0.6977	-0.0323	0.0461	0.1888
	α_{25}^1	α_{26}^1	α_{34}^1	α_{35}^1	α_{36}^1	α_{45}^1	α_{46}^1
	-0.2271	-0.2088	-0.2140	-0.1482	-0.4294	-0.4592	-0.4632
	α_{56}^1	α_{11}^1	α_{22}^1	α_{33}^1	α_{44}^1	α_{55}^1	α_{66}^1
	0.8225	-0.4550	0.0775	0.0027	0.0501	0.2337	0.1468
Cluster 2	α_0^1	α_1^1	α_2^1	α_3^1	α_4^1	α_5^1	α_6^1
	-0.5014	0.3994	-0.4947	0.5493	-0.5332	-0.5447	-0.6809
	α_{12}^1	α_{13}^1	α_{14}^1	α_{15}^1	α_{16}^1	α_{23}^1	α_{24}^1
	-0.2088	0.0436	-0.3890	-0.4340	-0.7480	-0.0163	0.3471
	α_{25}^1	α_{26}^1	α_{34}^1	α_{35}^1	α_{36}^1	α_{45}^1	α_{46}^1
	0.7261	0.0146	-0.1629	-0.6895	-0.0891	0.3833	-0.6391
	α_{56}^1	α_{11}^1	α_{22}^1	α_{33}^1	α_{44}^1	α_{55}^1	α_{66}^1
	0.9730	0.2977	-0.0963	0.0345	-0.0307	0.0916	-0.1271

Table 7-4 Coefficients of the nonlinear regression models for two clusters: COP.

Cluster index	α_0^2	α_1^2	α_2^1	α_3^2	α_4^2	α_5^2	α_6^2
Cluster 1	-0.5772	0.0250	-0.2520	0.3638	0.3583	0.7304	0.2598
	α_{12}^2	α_{13}^2	α_{14}^2	α_{15}^2	α_{16}^2	α_{23}^2	α_{24}^2
	-0.1995	0.0056	-0.3452	-0.2000	0.1903	0.1854	-0.1095
	α_{25}^2	α_{26}^2	α_{34}^2	α_{35}^2	α_{36}^2	α_{45}^2	α_{46}^2
	0.1870	0.3135	0.2590	0.2470	-0.0454	0.6926	0.4366
	α_{56}^2	α_{11}^2	α_{22}^2	α_{33}^2	α_{44}^2	α_{55}^2	α_{66}^2
	-0.1856	0.0893	-0.1597	-0.0917	0.1073	0.3453	-0.2936
Cluster 2	α_0^2	α_1^2	α_2^1	α_3^2	α_4^2	α_5^2	α_6^2
	-0.6728	-0.4779	-0.2369	0.0476	0.3863	0.8850	0.7415
	α_{12}^2	α_{13}^2	α_{14}^2	α_{15}^2	α_{16}^2	α_{23}^2	α_{24}^2
	0.1657	-0.3311	0.0720	0.0049	0.9778	0.2634	-0.3733
	α_{25}^2	α_{26}^2	α_{34}^2	α_{35}^2	α_{36}^2	α_{45}^2	α_{46}^2
	-0.6025	0.2164	0.2186	0.6650	-0.3826	0.1284	1.9722
	α_{56}^2	α_{11}^2	α_{22}^2	α_{33}^2	α_{44}^2	α_{55}^2	α_{66}^2
	-1.0340	-0.3867	-0.1777	-0.1367	-0.0633	0.5115	0.7575

Figures

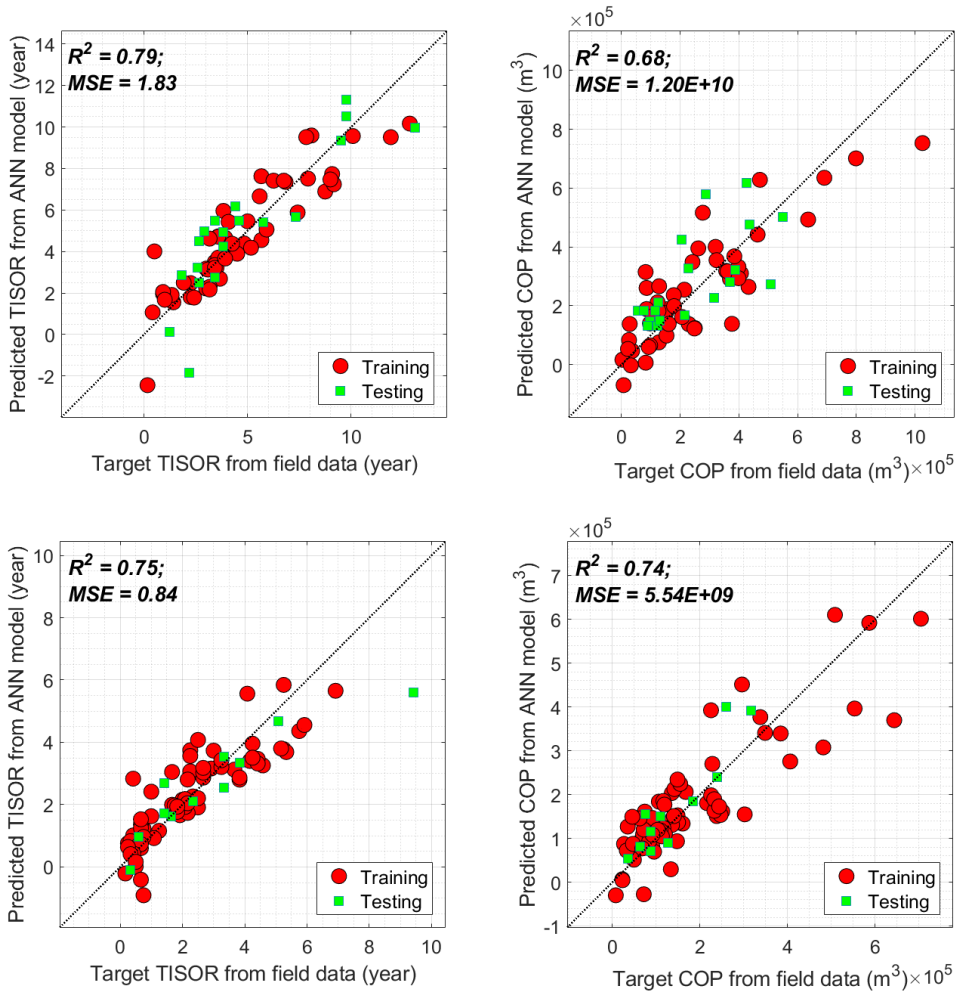


Fig. 7-1 Cross-plots of the predicted TISOR and COP obtained from the linear regression algorithm and target TISOR and COP from field data following k-mean clustering analysis: top – cluster 1; bottom – cluster 2.

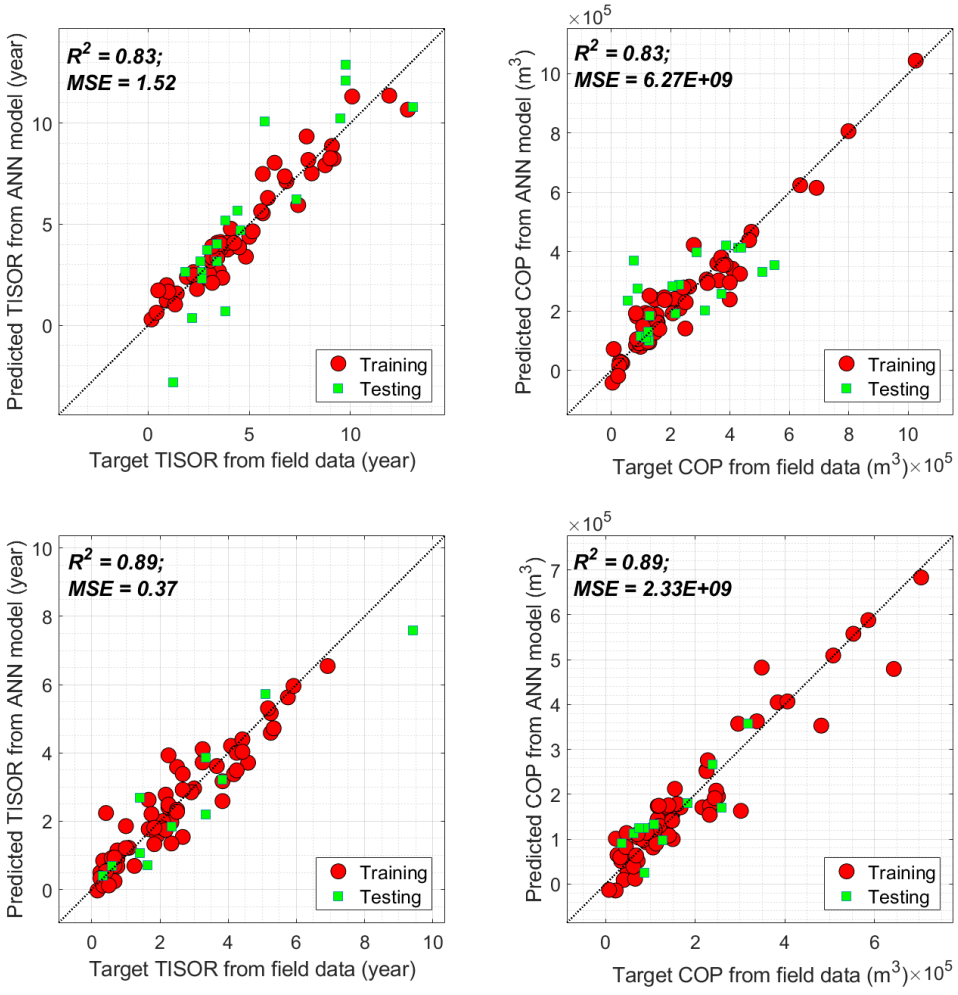


Fig. 7-2 Cross-plots of the predicted TISOR and COP obtained from the nonlinear regression algorithm and target TISOR and COP from field data following k-mean clustering analysis: top – cluster 1; bottom – cluster 2.

Chapter 8 Concluding Remarks

Chapter Overview

In this chapter, the conclusions of the thesis research, the summaries of the main contributions and novelties, and the brief discussions of the future work are presented.

8.1 Conclusions

A comprehensive dataset encompassing over 10 SAGD fields is compiled from public sources, which include well pad information, production performances, and reservoir properties. A practical SAGD field data analysis routine is developed, which is capable of efficiently analyzing a large amount of field data and compiling data samples for construction of data-driven models. Important variables related to reservoir heterogeneities and production constraints are identified through logging interpretation and production analysis.

A SAGD production analysis workflow is developed and employed to predict SAGD production performance using data-driven modeling techniques. ANN is utilized as the main technique to construct data-driven models by identifying the non-linear relationships between various input (e.g., reservoir and operating parameters) and output variables (SAGD performance). Performances of ANN models are shown to be both reliable and satisfactory, as evidenced by the high values of R^2 and low values of *MSE* between predictions and targets in the forward model.

Data-mining techniques, e.g., principal component analysis and clustering analysis are applied to improve the forecast capacity, efficiency, and robustness of the data-driven models. Improvements in forecast performances are observed. Results from this thesis demonstrate that ANN can be employed successfully to predict SAGD recovery performance from log-derived and operational variables alone, thus to correlate reservoir heterogeneities and production characteristics.

Uncertainty analysis is carried out to determine influences of the uncertainties originating from model parameter and data on the final ANN predictions. The analysis reveals that uncertainty due to limited data records is dominating. These results motivate future efforts in expanding the available dataset for uncertainty management.

In order to deal with the challenges associated with SAGD field data, synthetic models are successfully constructed. The influence of shale barrier on production performance is first investigated. A practical implementation of data-driven modeling techniques for characterizing shale heterogeneities is proposed by correlating the heterogeneity parameters with the extracted patterns from production data. A comprehensive set of important observable input parameters that are related to geological properties of shale barriers can be parameterized from production time-series data using piecewise linear approximation, cubic spline interpolation, and discrete wavelet transform. A set of novel output parameterization schemes are implemented to represent shale barrier configurations with reduced dimensional vectors. Once again, ANN is applied to calibrate a relationship between the retrieved production pattern parameters and the corresponding shale heterogeneities.

Characterization results from the ANN models are promising, as the predicted shale parameters match their targets. In comparison to the shale length, higher prediction fidelity is obtained with the location parameters. Model accuracy also decreases with the larger number of shale barriers. Generally speaking, slightly superior characterization performance can be observed for shale barriers that are located closer to the well pair. These calibrated ANN models are integrated into a model-selection workflow to infer shale distribution from actual production history. The outcome of this workflow is an ensemble of reservoir models that are consistent with the production history.

The methodology is extended to characterize more complex heterogeneities in SAGD reservoirs, where both shale barriers and lean zones exist. A total number of 2800 heterogeneous SAGD cases are generated based on field data. A novel two-level data-driven modeling characterization workflow is proposed to characterize such reservoir heterogeneities by correlating production time-series data and heterogeneity parameters using ANN: the first level model (the screening-ANN) is used to predict the number of lean zone and shale barrier regions; while the second level models (the sub-ANN) are used to estimate detailed geological characteristics of each heterogeneity. Results of the screening-ANN model are demonstrated

reliable as the majority of cases are classified into corresponding target scenarios. The estimation performances of the sub-ANN models are also satisfactory as demonstrated with high R^2 and low *MSE*, especially for the scenarios in which the total number of heterogeneities is small. Once the models are trained, they provide a quick and reliable characterization of reservoir heterogeneities.

The proposed characterization methodology is applied in heterogeneities characterization through two case studies whose true heterogeneities are unknown. Good characterization results are obtained as an ensemble of possible heterogeneous realizations that honor the actual production time-series data are generated, which demonstrates the great potential of the workflow for inferring distributions of heterogeneities only from production profiles.

This thesis presents a practical workflow that integrates production time-series data analysis and data-driven modeling techniques for complex SAGD reservoir heterogeneities characterization, a subject matter that is insufficiently explored in both the academics and industries. This methodology presented in this thesis does not aim to replace conventional reservoir characterization routines; instead, it intends to provide an efficient and complementary workflow for heterogeneities characterization from a large amount of SAGD field data.

8.2 Contributions and Novelties

High computational and capital cost, especially for complex and large reservoir models, hamper the practical applications of traditional reservoir simulation and characterization workflows into real-time reservoir management. The data-driven modeling techniques provide fast, efficient, and low-cost alternatives for estimating complex heterogeneities in reservoirs by selecting an ensemble of possible reservoirs models, which would honor the production performances. The contributions of this thesis are summarized as:

a) A large field dataset from over 2000 wells is compiled and analyzed. Analysis of SAGD field data in this size from the McMurray bitumen deposits has not been studied before. A set of important reservoir and operational parameters are identified and formulated from the field data analysis workflow. A forward data-driven model is successfully constructed to predict SAGD production performance. The application of data-driven models involving actual field data from this formation is novel. We first demonstrated that

SAGD performance, e.g., cumulative oil production, can be reliably predicted using data-driven modeling techniques instead of conventional numerical simulations.

b) Influences of the uncertainties originating from the model parameter, data (uncertainty in input attributes due to imprecise analysis criteria and limited number of records in the dataset) on the final ANN predictions are systematically quantified based on Monte Carlo and bootstrapping methods. Since uncertainty assessment is an area that is less explored in the intelligent and expert systems literature, particularly when reservoir-engineering data is involved, this thesis provides an important insight into strategies to improve the accuracy and performance of data-driven models.

c) Shale barriers and lean zones, two types of reservoir heterogeneities that have impacts on SAGD production, have been systematically investigated through detailed sensitivity analysis, such as the locations, geometries, size and numbers of heterogeneities. Although synthetic models are used in this thesis, they are capable of representing the typical SAGD operations in Athabasca oil sands. That is because the parameters that are used to construct the synthetic models are extracted from actual field data.

d) New parameterization methods are proposed to define inputs and outputs. First, based on the geological properties of shale barriers, lean zones, and the mechanism of the steam chamber growth, three parameters are formulated to represent each heterogeneity. Compared to the conventional history-matching workflows, where the unknown parameters are assigned to each grid block, this parametrization of heterogeneities in this thesis would significantly simplify the modeling complexity. Second, a group of time-series data analysis techniques are employed to parameterize input features from production profiles to reduce their dimensionality, e.g., DWT.

e) A practical reservoir heterogeneities characterization workflow is proposed via data-driven models. Unlike conventional reservoir characterization methods, the proposed workflow just analyze the observed production time-series data and apply data-driven modeling techniques to characterize heterogeneities in SAGD reservoirs. A new two-level data-driven model is proposed to correlate the nonlinear, complex, and uncertain relationship between input features and heterogeneity characteristics for more complex reservoir setting. This is a novel application of data-driven modeling techniques in oil and gas industry. Compared to conventional characterization workflows, which alter the

unknown parameters at each time step through different algorithms, the proposed method instantaneously estimates the shale and lean zone distribution and no iterations are required.

The novelties and advantages of implementation of data-driven modeling techniques in the analysis workflows presented in this thesis are listed here:

a) Applying data-driven techniques to characterize shale and lean zone heterogeneity in SAGD reservoir has a prominent advantage is that it does not require any previous assumptions and simplifications about the complex multiphase flow governing equations of the flowing system The methodology proposed in this thesis require only the features extracted from production time-series data to estimate the presence and distribution of heterogeneities. Consequently, implementation of the trained data-driven models may significantly simplify the tedious conventional reservoir characterization workflow, and improve the efficiency and productivity of real-time reservoir management and decision-making.
b) Data-driven techniques can be used to simulate the highly non-linear, complex, highly-dimensional, and uncertain SAGD flowing systems by working behind the scenes to find correlations between shale parameters and production responses. That is because the knowledge and information learned through training process have been automatically learned and stored in the models. Therefore, it can be easily re-used and accessed for new cases without training the models again, which reduces the computational costs.
c) The proposed workflows would not require a large amount of reservoir, fluid, and operational data; instead, they only consider the features extracted from production time-series data as input variables and build the internal correlation between production response with shale heterogeneities in the SAGD reservoirs.
d) The proposed characterization workflows have the capacity to be updated, once the new information becomes available. It does not require building and running the complex geological reservoir models, which is a time-consuming and tedious procedure; instead, it just needs to incorporate the new information or variables into the training process to estimate unknown reservoir properties, which is usually very efficient.
e) The proposed workflow would provide complementary heterogeneity estimation results to the conventional history-matching routines. From the production time-series data only,

it gives an ensemble of possible reservoir descriptions that honor the production history. For a more detailed estimation, one may incorporate the results from the proposed framework as the initial realizations rather than random initializations for conventional history-matching workflow. As a result, it would speed up the entire workflow for existing reservoir management.

8.3 Recommendations for Future Work

a) In this thesis, data-driven models are constructed by ANN and provide reliable production forecasts and heterogeneities characterizations. Considering the growing interest and popularity of big data and deep learning techniques in other fields of study, it would be very interesting to incorporate these types of techniques to analyze a large amount of field data and build data-driven models for characterization and prediction.
b) The current characterization workflows are constructed based on 2D SAGD models, where heterogeneities vary in the cross-well pair plane and have the same extension to the length of the well pair. As the distributions, geometries, and sizes of heterogeneities in a reservoir often vary in 3 dimensions, extending the current workflows to characterize heterogeneities in 3D reservoirs would be more practical.
c) Although the many effective parametrization techniques have been customized to formulate input features from production time-series data and output parameters from the reservoir in this thesis for the characterization workflow, it is interesting to investigate other options of parameterization to further improve the efficiency and robustness of data-driven models.
d) Since the proposed characterization methods have been demonstrated by the results from the data-driven models derived from synthetic data, therefore subsequent efforts would integrate models calibrated from field data directly. As mentioned before, this task would be challenging because many issues associated with actual field data. One option is that working with oil and gas industries to collect sufficient field data to build data-driven models. It would be promising to apply the models built from field data to practically estimate heterogeneity characteristics. In addition, since the characterization workflows are constructed using clean synthetic data, I will consider noisy or uncertain

data to build data-driven models and examine their impacts on characterization performance.

e) Despite the fact that the characterization workflows, presented in this thesis are capable of providing important insights regarding the distribution of reservoir heterogeneities for many cases, future work can also be extended to combine them with the conventional history-matching routines to obtain detailed match. One might just regard the characterized models from the proposed workflows as initialization for conventional history-matching, which would improve the overall speed and efficiency.

Bibliography

Adibifard, M., Tabatabaei-Nejad, S., & Khodapanah, E. (2014). Artificial neural network (ANN) to estimate reservoir parameters in naturally fractured reservoirs using well test data. *Journal of Petroleum Science and Engineering*, 122, 585-594.

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis, 22*(5), 717-727.

Akin, S., & Bagci, S. (2001). A laboratory study of single-well steam-assisted gravity drainage process. *Journal of petroleum science and engineering, 32*(2), 23-33.

Alberta Energy Regulator (2015), *AER Annual Performance Presentation*. Prepared by Nexen Long Lake Project.

Alberta Energy Regulator (2016a), *Alberta energy regulator 2015/16 annual report*. Report prepared by Alberta Energy Regulator, Alberta, Canada.

Alberta Energy Regulator (2016b), *AER Annual Performance Presentation*. Prepared by Suncor Firebag Projects.

Al-Bulushi, N., King, P., Blunt, M., & Kraaijveld, M. (2012). Artificial neural networks workflow and its application in the petroleum industry. *Neural Computing and Applications, 21*(3), 409-421.

Al-Fattah, S. M., & Startzman, R. A. (2001). Neural network approach predicts U.S. natural gas production, Paper presented at the *SPE Production and Operations Symposium*, Oklahoma City, Oklahoma, US.

Al-Fattah, S. M., & Startzman, R. A. (2003). Neural network approach predicts U.S. natural gas production. *SPE Production & Facilities, 18*(02), 84-91.

Aly, M. A. E. E. (2007). Data analysis methodology for reservoir management. Paper presented at the *EUROPEC/EAGE Conference and Exhibition,* London, U.K.

Amirian, E., Leung, J. Y., Zanon, S., & Dzurman, P. (2013). Data-driven modeling approach for recovery performance prediction in SAGD operations. Paper presented at the *SPE Heavy Oil Conference-Canada,* Calgary, Alberta, Canada.

Amirian, E., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Integrated cluster analysis and artificial neural network modeling for steam-assisted gravity drainage performance prediction in heterogeneous reservoirs. *Expert Systems with Applications, 42*(2), 723-740.

Ampazis, N., & Perantonis, S. J. (2000). Levenberg-marquardt algorithm with adaptive momentum for the efficient training of feedforward networks. Paper presented at the *IEEE-INNS-ENNS International Joint Conference on Neural Networks,* Como, Italy, 126-131,

An, P., & Moon, W. (1993). Reservoir characterization using feedforward neural networks. *In SEG Technical Program Expanded Abstracts 1993*, 258-262.

Awoleke, O., & Lane, R. (2011). Analysis of data from the barnett shale using conventional statistical and virtual intelligence techniques. *SPE Reservoir Evaluation & Engineering, 14*(05), 544-556.

Ayala H, L. F., & Ertekin, T. (2007). Neuro-simulation analysis of pressure maintenance operations in gas condensate reservoirs. *Journal of Petroleum Science and Engineering, 58*(1), 207-226.

Azim, M. R., Amin, M. S., Haque, S. A., Ambia, M. N., & Shoeb, M. A. (2010). Feature extraction of human sleep EEG signals using wavelet transform and fourier transform. Paper presented at the *2010 2nd International Conference on Signal Processing Systems*, 3 V3-701-V3-705, Dalian, China.

Bagci, A. S. (2006). Experimental and simulation studies of SAGD process in fractured reservoirs. Paper presented as the *SPE/DOE Symposium on Improved Oil Recovery*, Tulsa, Oklahoma.

Bahrololoum, A., Nezamabadi-pour, H., & Saryazdi, S. (2015). A data clustering approach based on universal gravity rule. *Engineering Applications of Artificial Intelligence, 45*, 415-428.

Bahrololoum, A., Nezamabadi-pour, H., and Saryazdi, S. (2015). A data clustering approach based on universal gravity rule. *Engineering Applications of Artificial Intelligence*, *45*, 415-428.

Barrett, J. P. (1974). The coefficient of determination—some limitations. *The American Statistician, 28*(1), 19-20.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (1992). *Neural network toolbox™ user's guide*. The Mathworks Inc.

Behler, J., & Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14), 146401.

Bhattacharya, S., & Nikolaou, M. (2013). Analysis of production history for unconventional gas reservoirs with statistical methods. *SPE Journal, 18*(05), 878-896.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University press.

Bravo, C. E., Saputelli, L., & Rivas, F. (2014). State of the art of artificial intelligence and predictive analytics in the e&p industry: a technology survey. *SPE Journal*, *19*(04).

Butler, R., McNab, G., & Lo, H. (1981). Theoretical studies on the gravity drainage of heavy oil during in‐situ steam heating. *The Canadian Journal of Chemical Engineering, 59*(4), 455-460.

Caers, J., & Hoffman, T. (2006). The probability perturbation method: A new look at bayesian inverse modeling. *Mathematical Geology, 38*(1), 81-100.

Chang, J., Ivory, J., & Tunney, C. (2012). Numerical simulation of steam-assisted gravity drainage with vertical slimholes. *SPE Reservoir Evaluation & Engineering, 15*(06), 662-675.

Chen, B., Wang, X., Yang, S., & McGreavy, C. (1999). Application of wavelets and neural networks to diagnostic system development, 1, feature extraction. *Computers & Chemical Engineering, 23*(7), 899-906.

Chen, Q., Gerritsen, M. G., & Kovscek, A. R. (2008). Effects of reservoir heterogeneities on the steam-assisted gravity-drainage process. *SPE Reservoir Evaluation & Engineering, 11*(05), 921-932.

Chow, L., & Butler, R. (1996). Numerical simulation of the steam-assisted gravity drainage process (SAGD). *Journal of Canadian Petroleum Technology, 35*(06), 55-62.

CMG, 2015. STARS: Users' Guide, advanced processes & thermal reservoir simulator (Version 2015), Calgary, Alberta, Canada: Computer 21 Modeling Group Ltd.

Dadashpour, M., Rwechungura, R. W., & Kleppe, J. (2011). Fast reservoir parameter estimation by using effect of principal components sensitivities and discrete cosine transform. Paper presented at the *SPE Reservoir Simulation Symposium,* The Woodlands, Texas, USA.

Dang, T. Q. C., Chen, Z., Nguyen, T. B. N., Bae, W., & Mai, C. L. (2013). Numerical simulation of SAGD recovery process in presence of shale barriers, thief zones, and fracture system. *Petroleum Science and Technology, 31*(14), 1454-1470.

Demuth, H., Beale, M., & Hagan, M. (2008). *Neural network toolbox™ 6*, *User's Guide.*

Deutsch, C. V., & Journel, A. G. (1998). *GSLIB geostatistical software library and User's guide (2nd ed.)*. New York: Oxford University Press, Inc.

Doan, L., Baird, H., Doan, Q., & Ali, S. (2003). Performance of the SAGD process in the presence of a water sand-a preliminary investigation. *Journal of Canadian Petroleum Technology, 42*(01).

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Egermann, P., Renard, G., & Delamaide, E. (2001). SAGD performance optimization through numerical simulations: Methodology and field case example. Paper presented at the *SPE International Thermal Operations and Heavy Oil Symposium,* Margarita Island, Venezuela.

Fairbridge, J. K., Cey, E., & Gates, I. D. (2012). Impact of intraformational water zones on SAGD performance. *Journal of Petroleum Science and Engineering, 82*, 187-197.

Fatemi, S. M. (2009). Simulation study of steam assisted gravity drainage (SAGD) in fractured systems. *Oil & Gas Science and Technology-Revue De l'IFP, 64*(4), 477-487.

Fedutenko, E., Yang, C., Card, C., & Nghiem, L. X. (2014). Time-dependent neural network based proxy modeling of SAGD process. Paper presented at the *SPE Heavy Oil Conference-Canada,* Calgary, Alberta, Canada.

Feizabadi, S. A., Zhang, X. K., & Yang, P. (2014). An integrated approach to building history-matched geomodels to understand complex long lake oil sands reservoirs, part 2: Simulation. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Canada.

Foroud, T., Seifi, A., and AminShahidi, B. (2014). Assisted history matching using artificial neural network based global optimization method–applications to Brugge field and a fractured Iranian reservoir. *Journal of Petroleum Science and Engineering, 123*, 46-61.

Francis, L. (2001). The basics of neural networks demystified. *Contingencies (11/12 2001), 1*(1), 56-61.

Ge, J., Xia, Y., & Nadungodage, C. (2010). UNN: A neural network for uncertain data classification. *Advances in Knowledge Discovery and Data Mining* (449-460), Springer.

Gharbi, R. B., & Elsharkawy, A. M. (1999). Neural network model for estimating the PVT properties of middle east crude oils. *SPE Reservoir Evaluation & Engineering, 2*(03), 255-265.

Gu, Y., & Oliver, D. S. (2005). History matching of the PUNQ-S3 reservoir model using the ensemble kalman filter. *SPE Journal, 10*(02), 217-224.

Guan, B. T., Gertner, G. Z., & Parysow, P. (1997). A framework for uncertainty assessment of mechanistic forest growth models: A neural network example. *Ecological Modelling, 98*(1), 47-58.

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the marquardt algorithm. *IEEE transactions on Neural Networks, 5*(6), 989-993.

Hammouda, K., & Karray, F. (2000). A comparative study of data clustering techniques. University of Waterloo, Ontario, Canada.

Han, D., Kwong, T., & Li, S. (2007). Uncertainties in real-time flood forecasting with neural networks. *Hydrological Processes, 21*(2), 223-228.

Hanna, S. R., Chang, J. C., & Fernau, M. E. (1998). Monte carlo estimates of uncertainties in predictions by a photochemical grid model (UAM-IV) due to uncertainties in input variables. *Atmospheric Environment, 32*(21), 3619-3628.

Harding, T. G., Zanon, S., Imran, M., & Kerr, R. K. (2016). In-situ reflux: An improved in-situ recovery method for oil sands. Paper presented at the *SPE Canada Heavy Oil Technical Conference*, Calgary, Alberta, Canada.

Hasani, M., & Emami, F. (2008). Evaluation of feed-forward back propagation and radial basis function neural networks in simultaneous kinetic spectrophotometric determination of nitroaniline isomers. *Talanta, 75*(1), 116-126.

Haykin, S.S. (2008). *Neural networks and learning machines* (3rd ed.), Upper Saddle River, NJ, USA: Pearson.

Heaton, J. (2008). *Introduction to neural networks with jav*, Heaton Research, Inc.

Hiebert, A. D., Morrish, I. C., Card, C., Ha, H., Porter, S., Kumar, A., . . ., & Close, J. C. (2013). Incorporating 4D seismic steam chamber location information into assisted history matching for A SAGD simulation. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Hocking, G., Cavender, T. W., & Person, J. (2011). Single-well SAGD: Overcoming permeable lean zones and barriers. Paper presented at the *Canadian Unconventional Resources Conference*, Calgary, Alberta, Canada.

Holdaway, K. R. (2009). Exploratory data analysis in reservoir characterization projects. Paper presented at the *SPE/EAGE Reservoir Characterization & Simulation Conference,* Abu Dhabi, UAE.

Ito, Y., Hirata, T., & Ichikawa, M. (2001). The growth of the steam chamber during the early period of the UTF phase B and hangingstone phase I projects. *Journal of Canadian Petroleum Technology, 40*(09).

Jahanbakhshi, R., & Keshavarzi, R. (2016). Intelligent classifier approach for prediction and sensitivity analysis of differential pipe sticking: a comparative study. *Journal of Energy Resources Technology, 138*(5) pp. 052904.

Jansen, F., & Kelkar, M. (1996). Exploratory data analysis of production data. Paper presented at the *Permian Basin Oil & Gas Recovery Conference,* 331-342, Midland, Texas, USA.

Jia, X., Cunha, L., & Deutsch, C. (2009). Investigation of a stochastic optimization method for automatic history matching of SAGD processes. *Journal of Canadian Petroleum Technology, 48*(01), 14-18.

Jolliffe, I. (2005). *Principal component analysis*, Wiley Online Library.

Joo, S., Oh, S. E., Sim, T., Kim, H., Choi, C. H., Koo, H., & Mun, J. H. (2014). Prediction of gait speed from plantar pressure using artificial neural networks. *Expert Systems with Applications, 41*(16), 7398-7405.

Kempeneers, P., De Backer, S., Debruyn, W., & Scheunders, P. (2004). Wavelet-based feature extraction for hyperspectral vegetation monitoring. Paper presented at the *Proceedings of SPIE*, Barcelona, Spain, 5238 297-305.

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2004). Segmenting time series: A survey and novel approach. *Data Mining in Time Series Databases*, 57, 1-22.

Khamehchi, E., Kivi, I. R., and Akbari, M. (2014). A novel approach to sand production prediction using artificial intelligence. *Journal of Petroleum Science and Engineering, 123*, 147-154.

Khosravi, A., Nahavandi, S., & Creighton, D. (2013). Quantifying uncertainties of neural network-based electricity price forecasts. *Applied Energy, 112*, 120-129.

Kisi, Ö., (2004). River flow modeling using artificial neural networks. *Journal of Hydrologic Engineering, 9*(1), 60-63.

Kjræulff, U. B., & Madsen, A. L. (2008). *Bayesian networks and influence diagrams: A guide to construction and analysis*, Springer Science+Business Media.

Kocadağlı, O., & Aşıkgil, B. (2014). Nonlinear time series forecasting with bayesian neural networks. *Expert Systems with Applications, 41*(15), 6596-6610.

Kröse, B., Krose, B., van der Smagt, P., & Smagt, P. (1993). *An introduction to neural networks*. The Netherlands: University of Amsterdam.

Lacroix, S., Renard, G., & Lemonnier, P. (2003). Enhanced numerical simulations of ior processes through dynamic sub-gridding. Paper presented at *Canadian International Petroleum Conference*, Calgary, Alberta, Canada.

Law, D. H., Nasr, T. N., & Good, W. K. (2003). Field-scale numerical simulation of SAGD process with top-water thief zone. *Journal of Canadian Petroleum Technology, 42*(08).

Le Ravalec, M., Morlot, C., Marmier, R., & Foulon, D. (2009). Heterogeneity impact on SAGD process performance in mobile heavy oil reservoirs. *Oil & Gas Science and Technology-Revue De l'IFP, 64*(4), 469-476.

Lechner, J. P., & Zangl, G. (2005). Treating uncertainties in reservoir performance prediction with neural networks. Paper presented at the *SPE Europec/EAGE Annual Conference,* Madrid, Spain.

Lee, H., Jin, J., Shin, H. (2015). Efficient prediction of SAGD productions using static factor clustering. *Journal of Energy Resources Technology, 137*(3) pp. 032907.

Lee, S. H., Kharghoria, A., & Datta-Gupta, A. (2002). Electrofacies characterization and permeability predictions in complex reservoirs. *SPE Reservoir Evaluation & Engineering, 5*(03), 237-248.

Li, W., Mamora, D., Li, Y., & Qiu, F. (2011). Numerical investigation of potential injection strategies to reduce shale barrier impacts on SAGD process. *Journal of Canadian Petroleum Technology, 50*(03), 57-64.

Li, X., Chan, C., & Nguyen, H. (2013). Application of the neural decision tree approach for prediction of petroleum production. *Journal of Petroleum Science and Engineering, 104*, 11-16.

Ma, Z., Leung, J. Y., & Zanon, S. (2018). Integration of artificial intelligence and production data analysis for shale heterogeneity characterization in steam-assisted gravity-drainage reservoirs, *Journal of Petroleum Science and Engineering, 163*, 139–155.

Ma, Z., Leung, J. Y., & Zanon, S. (2016). Integration of artificial intelligence and production data analysis for shale heterogeneity characterization in SAGD reservoirs. Paper presented at the *SPE Canada Heavy Oil Technical Conference*, Calgary, Alberta, Canada.

Ma, Z., Leung, J. Y., & Zanon, S. (2017). Practical data mining and artificial neural network modeling for SAGD production analysis. *Journal of Energy Resources Technology, 139*(3), 032909.

Ma, Z., Leung, J. Y., Zanon, S., & Dzurman, P. (2015). Practical implementation of knowledge-based approaches for steam-assisted gravity drainage production analysis. *Expert Systems with Applications, 42*(21), 7326-7343.

Ma, Z., Leung, J. Y., Zanon, S., & Dzurman, P. (2014). Practical implementation of knowledge-based approaches for SAGD production analysis. Paper presented at the *SPE Heavy Oil Conference Canada*, Calgary, Alberta, Canada.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Anonymous California, USA, 1, 281-297.

Manshad, A. K., Rostami, H., & Hosseini, S. M. (2016). Application of artificial neural network-particle swarm optimization algorithm for prediction of gas condensate dew point pressure and comparison with Gaussian processes regression-particle swarm optimization algorithm. *Journal of Energy Resources Technology, 138*(3) pp. 032903.

MathWorks (2017). *Statistics and Machine Learning Toolbox User's Guide*. Retrieved December 28, 2017 from https://www.mathworks.com/help/releases/R2017b/pdf_doc/stats/stats.pdf.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

McKinley, S., & Levine, M. (1998). Cubic spline interpolation. *College of the Redwoods, 45*(1), 1049-1060.

Mezić, I., & Runolfsson, T. (2008). Uncertainty propagation in dynamical systems. *Automatica, 44*(12), 3003-3013.

Mirzabozorg, A., Nghiem, L., Chen, Z., Yang, C., & Hajizadeh, Y. (2012). History matching saturation and temperature fronts with adjustments of petro-physical properties; SAGD case study. Paper presented at the *SPE Kuwait International Petroleum Conference and Exhibition*, Kuwait City, Kuwait.

Mirzabozorg, A., Nghiem, L., Chen, Z., & Yang, C. (2013). Differential evolution for assisted history matching process: SAGD case study. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Mohammadpoor, M., Firouz, Q., Reza, A., & Torabi, F. (2012). Implementing simulation and artificial intelligence tools to optimize the performance of the CO2 sequestration in coalbed methane reservoirs. Paper presented at the *Carbon Management Technology Conference,* Orlando, Florida, USA.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*, John Wiley & Sons.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: Process and product optimization using designed experiments*, John Wiley & Sons.

Nigrin, A. (1993). *Neural networks for pattern recognition*. MIT press.

Nilsson, N. J. (2014). *Principles of artificial intelligence*, Morgan Kaufmann.

Norman, C. D. (2013). Correlation of porosity uncertainty to productive reservoir volume. Paper presented at the *SPE Middle East Oil and Gas Show and Conference,* Manama, Bahrain.

Noureldin, A., El-Shafie, A., & Taha, M. R. (2007). Optimizing neuro-fuzzy modules for data fusion of vehicular navigation systems using temporal cross-validation. *Engineering Applications of Artificial Intelligence, 20*(1), 49-61.

Oliver, D. S., & Chen, Y. (2011). Recent progress on reservoir history matching: A review. *Computational Geosciences, 15*(1), 185-221.

Panwar, A., Trivedi, J. J., & Nejadi, S. (2012). Importance of distributed temperature sensor (DTS) placement for SAGD reservoir characterization and history matching within ensemble kalman filter (EnKF) framework. Paper presented at the *SPE Latin America and Caribbean Petroleum Engineering Conference*, Mexico City, Mexico.

Panwar, A., Trivedi, J. J., and Nejadi, S. (2015). Importance of distributed temperature sensor data for steam assisted gravity drainage reservoir characterization and history matching within ensemble Kalman filter framework. *Journal of Energy Resources Technology, 137*(4) pp. 042902.

Papadopoulos, C. E., & Yeung, H. (2001). Uncertainty estimation and Monte Carlo simulation method. *Flow Measurement and Instrumentation, 12*(4), 291-298.

Parada, C. H., & Ertekin, T. (2012). A new screening tool for improved oil recovery methods using artificial neural networks. Paper presented as the *SPE Western Regional Meeting*, Bakersfield, California, USA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830.

Petrović, D. V., Tanasijević, M., Milić, V., Lilić, N., Stojadinović, S., & Svrkota, I. (2014). Risk assessment model of mining equipment failure based on fuzzy logic. *Expert Systems with Applications, 41*(18), 8157-8164.

Pooladi-Darvish, M., & Mattar, L. (2002). SAGD operations in the presence of overlying gas cap and water layer-effect of shale layers. *Journal of Canadian Petroleum Technology, 41*(06).

Popa, A. S., & Patel, A. N. (2012). Neural networks for production curve pattern recognition applied to cyclic steam optimization in diatomite reservoirs. Paper presented at the *SPE Western Regional Meeting,* Bakersfield, California, USA.

Popa, A. S., Cassidy, S. D., & Mercer, M. (2011). A data mining approach to unlock potential from an old heavy oil field. Paper presented at the *SPE Western North American Region Meeting,* Anchorage, Alaska, USA.

Pyrcz, M. J., & Deutsch, C. V. (2014). Geostatistical reservoir modeling, Oxford University Press.

Queipo, N. V., Goicochea, J. V., & Pintos, S. (2002). Surrogate modeling-based optimization of SAGD processes. *Journal of Petroleum Science and Engineering, 35*(1), 83-93.

Radunovic, D. P. (2009). *Wavelets: From math to practice (1st Ed.)*, Springer Publishing Company, Incorporated.

Ramgulam, A. (2006). Utilization of artificial neural networks in the optimization of history matching (Doctoral dissertation, the Pennsylvania State University).

Refsgaard, J. C., van der Sluijs, Jeroen P, Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process–a framework and guidance. *Environmental Modelling & Software, 22*(11), 1543-1556.

Rioul, O., & Vetterli, M. (1991). Wavelets and signal processing. *IEEE Signal Processing Magazine, 8*(LCAV-ARTICLE-1991-005), 14-38.

Rodríguez, G., Soria, Á., & Campo, M. (2016). Artificial intelligence in service-oriented software design. *Engineering Applications of Artificial Intelligence*, *53*.

Romero, C., & Carter, J. (2001). Using genetic algorithms for reservoir characterization. *Journal of Petroleum Science and Engineering, 31*(2), 113-123.

Sarma, P., Durlofsky, L. J., Aziz, K., & Chen, W. H. (2007). A new approach to automatic history matching using kernel PCA. Paper presented at the *SPE Reservoir Simulation Symposium,* Houston, Texas, USA.

Sasaki, K., Akibayashi, S., & Kosukegawa, H. (1996). Experimental study on initial stage of SAGD process using 2-dimensional scaled model for heavy oil recovery. Paper presented at the *International Conference on Horizontal Well Technology*, Calgary, Alberta, Canada.

Scheevel, J., & Payrazyan, K. (2001). Principal component analysis applied to 3D seismic data for reservoir property estimation. *SPE Reservoir Evaluation & Engineering, 4*(01), 64-72.

Secchi, P., Zio, E., & Di Maio, F. (2008). Quantifying uncertainties in the estimation of safety parameters by using bootstrapped artificial neural networks. *Annals of Nuclear Energy, 35*(12), 2338-2350.

Secchi, P., Zio, E., & Di Maio, F. (2008). Quantifying uncertainties in the estimation of safety parameters by using bootstrapped artificial neural networks. *Annals of Nuclear Energy, 35*(12), 2338-2350.

Shin, H., & Polikar, M. (2006). Experimental investigation of the fast-SAGD process. Paper presented at the *Canadian International Petroleum Conference,* Calgary, Alberta, Canada.

Sereda, J. N., & James, B. R. (2014). A case study in the application of bitumen geochemistry for reservoir characterization in SAGD development. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Sheng, J. (2013). *Enhanced oil recovery field case studies*, Gulf Professional Publishing.

Shin, H., Hwang, T., & Chon, B. (2012). Optimal grid system design for SAGD simulation. Paper presented at the *SPE Heavy Oil Conference Canada*, Calgary, Alberta, Canada.

Shlens, J. (2014). A tutorial on principal component analysis. *ArXiv Preprint arXiv:1404.1100*.

Singh, G., & Panda, R. K. (2011). Daily sediment yield modeling with artificial neural network using 10-fold cross validation method: A small agricultural watershed, kapgari, india. *International Journal of Earth Sciences and Engineering, 4*(6), 443-450.

Siu, A., Nghiem, L., Gittins, S., Nzekwu, B., & Redford, D. (1991). Modelling steam-assisted gravity drainage process in the UTF pilot project. Paper presented at the *SPE Annual Technical Conference and Exhibition,* Dallas, Texas, USA.

Smith, L. I. (2002). A tutorial on principal components analysis. Cornell University, USA, 51(52), 65.

Solomatine, D., & Ostfeld, A. (2008). Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics, 10*(1), 3-22.

Solomatine, D., See, L. M., Abrahart, R. (2009). Data-driven modeling: Concepts, approaches and experiences. *Practical hydroinformatics* (pp. 17-30), Springer.

Srivastav, R., Sudheer, K., & Chaubey, I. (2007). A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. *Water Resources Research, 43*(10).

Stundner, M., & Al-Thuwaini, J. S. (2001). How data-driven modeling methods like neural networks can help to integrate different types of data into reservoir management. Paper presented as the *SPE Middle East Oil Show*, Manama, Bahrain.

Sung, A. (1998). Ranking importance of input parameters of neural networks, *Expert Systems with Applications*, *15*(3), 405–411.

Tan, S. S., & Smeins, F. E. (1996). Predicting grassland community changes with an artificial neural network model. *Ecological Modelling, 84*(1-3), 91-97.

Tang, H., Meddaugh, W. S., & Toomey, N. (2011). Using an artificial-neural-network method to predict carbonate well log facies successfully. *SPE Reservoir Evaluation & Engineering, 14*(01), 35-44.

Tiwari, M. K., & Chatterjee, C. (2010). Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs). *Journal of Hydrology, 382*(1), 20-33.

Toffolo, A. (2009). Fuzzy expert systems for the diagnosis of component and sensor faults in complex energy systems. *Journal of Energy Resources Technology, 131*(4).

Tukey, J. W. (1977). Exploratory data analysis, Addison-Wesley.

Ulker, E., & Sorgun, M. (2016). Comparison of computational intelligence models for cuttings transport in horizontal and deviated wells. *Journal of Petroleum Science and Engineering, 146*, 832-837.

Verga, F., Viberti, D., & Gonfalini, M. (2002). Uncertainty evaluation in well logging: Analytical or numerical approach? Paper presented at the *SPWLA 43rd Annual Logging Symposium,* Oiso, Japan.

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, Jeroen P, van Asselt, M. B., Janssen, P., & Krayer von Krauss, Martin P. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment, 4*(1), 5-17.

Wang, C., & Leung, J. (2015). Characterizing the effects of lean zones and shale distribution in steam-assisted-gravity-drainage recovery performance. *SPE Reservoir Evaluation & Engineering, 18*(03), 329-345.

Wang, Y., and Salehi, S. (2015). Application of Real-Time Field Data to Optimize Drilling Hydraulics using Neural Network Approach. *Journal of Energy Resources Technology, 137*(6) pp. 062903.

Waszczyszyn, Z. (1999*). Neural networks in the analysis and design of structures*, Springer.

Wei, T. (2013). Corrplot: visualization of a correlation matrix, *R Package Version 0.73*.

Williams, M., Keating, J., & Barghouty, M. (1998). The stratigraphic method: A structured approach to history matching complex simulation models. *SPE Reservoir Evaluation & Engineering, 1*(02), 169-176.

Wright, W. (1999). Bayesian approach to neural-network modeling with input uncertainty. *Neural Networks, IEEE Transactions On, 10*(6), 1261-1270.

Xu, J., Chen, Z. J., Cao, J., & Li, R. (2014). Numerical study of the effects of lean zones on SAGD performance in periodically heterogeneous media. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Xu, S., & Chen, L. (2008). A novel approach for determining the optimal number of hidden layer neurons for FNN,Äôs and its application in data mining. Paper presented at the *International Conference on Information Technology and Applications: iCITA,* 683-686.

Yadav, S. (2006). History matching using face-recognition technique based on principal component analysis. Paper presented at the *SPE Annual Technical Conference and Exhibition,* San Antonio, Texas, USA.

Yang, G., & Butler, R. (1992). Effects of reservoir heterogeneities on heavy oil recovery by steam-assisted gravity drainage. *Journal of Canadian Petroleum Technology, 31*(08), 37-43.

Yang, R., Zhang, J., Yang, L., Chen, H., & Tang, S. (2016). Performance and calculation method of steam chamber overcoming high water saturation intervals during SAGD process. Paper presented at the *International Petroleum Technology Conference*, Bangkok, Thailand.

Yeten, B., Durlofsky, L. J., & Aziz, K. (2002). Optimization of nonconventional well type location and trajectory. Paper presented at the *SPE Annual Technical Conference and Exhibition,* San Antonio, Texas, USA.

Yohanes, R. E., Ser, W., & Huang, G. (2012). Discrete wavelet transform coefficients for emotion recognition from EEG signals. Paper presented at the *Engineering in Medicine and Biology Society (EMBC), 2012* Annual International Conference of the IEEE, 2251-2254, San Diego, CA, USA.

Zerafat, M. M., Ayatollahi, S., Mehranbod, N., & Barzegari, D. (2011). *Bayesian network analysis as a tool for efficient EOR screening*. Paper presented as the *SPE Enhanced Oil Recovery Conference*, Kuala Lumpur, Malaysia.

Zhang, F., Reynolds, A. C., & Oliver, D. S. (2003). An initial guess for the Levenberg–Marquardt algorithm for conditioning a stochastic channel to pressure data. *Mathematical Geology, 35*(1), 67-88.

Zhang, X. K., Feizabadi, S. A., & Yang, P. (2014). An integrated approach to building history-matched geomoels to understand complex long lake oil sands reservoirs, part 1: Geomodeling. Paper presented at the *SPE Heavy Oil Conference-Canada*, Calgary, Alberta, Canada.

Zhang, X., Liang, F., Yu, B., & Zong, Z. (2011). Explicitly integrating parameter, input, and structure uncertainties into bayesian neural networks for probabilistic hydrologic forecasting. *Journal of Hydrology, 409*(3), 696-709.

Appendix

Table A-1 Raw data used in chapter 4.

ϕ	Sw	*N/G*	*h*	*SI*	*d*	N_e^{prod}	N_e^{inj}	T_{total}	CSI	TISOR	COP
0.28512	0.46598	0.4193	70	0.26242	10.856	1.297	1.0316	10.819	9.68E+05	5.75	4.27E+05
0.28359	0.39252	0.4761	71.033	0.72141	3.2272	1	1.0396	10.736	9.03E+05	6.8333	3.21E+05
0.29185	0.3495	0.69144	55.575	0.26785	15.151	1.4329	1.0238	10.736	8.39E+05	5.6667	4.72E+05
0.31167	0.26466	0.45837	44.333	0.52328	4.8262	1	1.0186	9.1333	8.19E+05	5.0833	3.17E+05
0.2792	0.27025	0.57861	61.7	0.18129	2.4452	1	1.0098	8.7889	6.77E+05	5.75	2.25E+05
0.26534	0.33712	0.51214	60.533	0.8244	23.639	1	1.0098	8.7889	6.79E+05	4.4167	2.29E+05
0.30739	0.22278	0.78345	56.8	0.11565	2.6388	1.1161	1	5.0722	7.25E+05	2.6667	3.49E+05
0.3064	0.15527	0.59751	48.2	0.70893	2.3852	1.0991	1	5.0722	7.73E+05	2.25	3.84E+05
0.29441	0.30548	0.63065	56.65	0.40312	3.6518	1.0991	1	5.0722	7.71E+05	1.6667	4.81E+05
0.28344	0.24777	0.6648	53.7	0.41352	3.6095	1	1.029	2.875	3.28E+05	1.6667	1.49E+05
0.28569	0.30608	0.44016	36.4	0.9529	12.399	1	1.4967	0.84444	1.64E+05	0.75	49034
0.28125	0.35662	0.63545	55.467	0.22093	33.482	1	1	6.6778	9.99E+05	5	4.07E+05
0.29079	0.43178	0.62169	79.3	0.23938	3.4404	1	1	6.6778	8.71E+05	4.5833	3.71E+05
0.30668	0.16197	0.80127	62.9	0.25033	16.766	1	1.1115	1.5194	3.29E+05	0.66667	1.67E+05
0.32629	0.18806	0.66061	34.55	11.508	20.438	1	1.1115	1.5194	2.16E+05	0.58333	1.27E+05
0.309	0.20899	0.57297	55.5	10.22	4.3684	1	1.0519	1.6056	3.17E+05	1.25	1.36E+05
0.32479	0.14295	0.86667	54	0.062075	21.13	1	1.0519	1.6056	2.74E+05	1.0833	1.15E+05
0.29398	0.15402	0.79749	55.8	0.12723	5.6016	1	1.1115	1.5194	3.94E+05	0.41667	2.16E+05
0.29021	0.21296	0.36528	55.3	6.5338	7.1563	1	1.0519	1.6056	2.89E+05	0.75	1.60E+05
0.29673	0.23097	0.71814	40.8	0.42435	4.4028	1	1.102	1.6889	3.03E+05	0.66667	1.47E+05

0.31409	0.21934	0.78558	52.7	0.054844	11.861	1	1.144	1.775	3.18E+05	0.25	2.51E+05
0.30861	0.17688	0.80752	45.2	0.22211	6.6016	1	1.144	1.775	3.58E+05	0.41667	2.38E+05
0.31781	0.19254	0.44318	44	0.5836	10.038	1	1.091	1.8611	3.93E+05	0.66667	2.47E+05
0.31637	0.19267	0.55896	52.733	0.73675	29.415	1	1.091	1.8611	4.11E+05	0.75	2.25E+05
0.3308	0.13669	1	44.6	0	15.084	1	1	0.41944	93854	0.33333	40404
0.3057	0.17388	0.78912	37.467	0.26116	40.899	1	1.2234	0.75833	1.41E+05	0.5	66026
0.31019	0.14533	0.82461	35.8	0.165	48.338	1	1.1363	1.2639	2.81E+05	0.58333	1.50E+05
0.31409	0.21934	0.78558	52.7	0.051522	42.364	1	1.2234	0.75833	82248	0.66667	34979
0.31781	0.19254	0.44318	44	1.8845	27.942	1	1.1045	1.2236	2.74E+05	0.66667	1.36E+05
0.30555	0.29917	0.29014	51.7	3.3691	34.061	1	1.0987	0.84444	1.37E+05	0.41667	65850
0.2914	0.24135	0.59451	56.35	0.42991	21.439	1	1.1987	0.41944	58845	0.16667	26523
0.29687	0.18902	0.36952	47.9	1.4309	5.4033	1	1.1136	0.75833	1.72E+05	0.25	86482
0.309	0.20899	0.57297	55.5	1.9039	5.7322	1	1.1136	0.75833	1.62E+05	0.33333	69946
0.29398	0.15402	0.79749	55.8	3.074	1.8559	1	1.2007	0.84444	2.22E+05	0.5	1.09E+05
0.305	0.23358	0.77462	59.9	0.1797	24.353	1	1.2007	0.84444	2.26E+05	0.33333	1.09E+05
0.31177	0.29172	0.81164	35.9	50	22.929	1	0.95845	4.1444	2.37E+05	2.1667	90975
0.25486	0.63819	0.41441	55.5	0.4428	18.888	1.1087	1.139	8.5361	1.15E+06	5.9167	4.64E+05
0.35266	0.12956	1	21.6	0	6.1716	0.99172	0.99172	10.4	1.04E+06	6.9167	2.96E+05
0.33346	0.19089	0.76235	28.5	0.2125	17.628	1	1.0945	2.7056	2.92E+05	1.9167	1.16E+05
0.3483	0.14667	0.84921	12.6	0.23869	23.858	1.312	1.0074	11.664	1.25E+06	5.25	5.86E+05
0.20253	0.60096	0.71597	59.5	0.21458	36.437	1.3966	1.022	11.494	1.15E+06	3.8333	7.99E+05
0.33361	0.14042	0.94907	21.6	0.094828	4.502	1.3973	1.0147	11.494	1.13E+06	4.0833	7.04E+05
0.33277	0.12382	0.72692	26	0.391	5.9315	1.5563	1.0086	9.7222	1.08E+06	5.9167	5.08E+05
0.21802	0.49612	0.76569	47.8	0.27731	34.038	1	1	3.6333	3.78E+05	3.4167	1.28E+05
0.17336	0.59277	0.58472	30.1	6	55.81	1	1	3.5472	2.87E+05	3.5	97880

0.21802	0.49612	0.76569	47.8	1.7879	22.536	1.0905	1	6.5083	9.40E+05	4.0833	4.33E+05
0.29441	0.41085	0.83152	45.45	1.254	20.808	1.0905	1	6.5083	7.24E+05	2.25	4.06E+05
0.20246	0.52594	0.52	37.5	1.7183	38.36	1.1677	1	6.5083	7.42E+05	4	3.57E+05
0.26033	0.455	0.71513	50.9	0.34109	35.563	1.0773	1	6.5083	7.64E+05	3.9167	3.70E+05
0.34153	0.21021	1	14.5	0	25.545	1	1	3.9722	6.50E+05	3.6667	2.33E+05
0.33516	0.35101	0.57986	28.8	1.2338	19.862	1.3151	1	7.775	7.86E+05	4.25	3.38E+05
0.3335	0.14488	0.96382	30.4	0.064365	12.314	1	1	4.1417	4.88E+05	3.8333	1.36E+05
0.33282	0.2679	0.85714	16.8	0.13634	8.3646	1	1	2.45	2.78E+05	1	1.33E+05
0.20524	0.55727	0.39667	42.1	0.31941	52.634	1	1	1.5194	1.23E+05	1.4167	36451
0.33606	0.164	0.79793	19.3	4.2424	10.735	1	1	1.1861	73721	1	22611
0.32572	0.19897	1	24.8	0	42.002	1	1	1.6889	1.09E+05	1.625	35170
0.32451	0.15262	0.77813	32	8.1609	12.1	1	1	5.8333	1.27E+06	2.5	6.43E+05
0.26017	0.291	0.73868	50	0.20575	32.894	1	1	5.8333	1.10E+06	3.4167	5.08E+05
0.30155	0.10222	0.93314	34.4	0.097458	55.888	1	1	5.8333	1.11E+06	3	5.53E+05
0.26764	0.34517	0.58639	73.5	0.63585	6.0056	1	1	5.8333	6.90E+05	2.9167	3.16E+05
0.30396	0.24131	0.8061	45.9	0.12608	6.677	1	1	3.5472	4.95E+05	3.3333	1.84E+05
0.24389	0.34496	0.28857	55.1	0.87518	6.7268	1	1	5.8333	4.99E+05	3.5833	1.52E+05
0.27534	0.35271	0.80653	39.8	0.24881	39.786	1	1	5.8333	5.05E+05	4.8333	1.46E+05
0.29594	0.23188	0.56672	66.7	0.40006	39.532	1	1	4.9861	9.19E+05	3.8333	3.87E+05
0.21457	0.51721	0.46886	85.1	0.61755	14.51	1	1	3.4639	2.72E+05	2.5833	97172
0.28872	0.46628	0.11634	40.4	0.98413	5.8713	1	1	4.8167	6.33E+05	4.5	2.30E+05
0.23807	0.42669	0.40342	93.7	0.7403	21.811	1	1	3.4639	2.60E+05	1.8333	1.26E+05
0.31476	0.18905	0.23984	24.6	4.6633	12.04	1	1	2.2778	2.61E+05	2	95105
0.23659	0.50723	0.64554	50.5	0.16104	7.0164	1.2187	1	6.1722	4.89E+05	3.5833	2.15E+05
0.19324	0.67318	0.54927	47.7	0.58383	31.902	1	1.014	6.1722	2.64E+05	2.6667	88276

0.26536	0.55087	0.61343	78.9	0.33483	39.96	1.0456	1	3.7167	3.71E+05	2.25	1.78E+05
0.25312	0.65494	0.29268	73.8	0.39583	31.494	1	1	3.7167	3.45E+05	3.6667	1.14E+05
0.2247	0.59631	0.61816	96.9	0.69401	13.082	1	1	3.7167	3.09E+05	3.4167	1.14E+05
0.22401	0.68177	0.47559	55.3	2.4631	22.456	1	1.0489	3.4639	6.08E+05	3	2.50E+05
0.26447	0.46888	0.65529	85	0.39585	26.459	1	1.0489	3.4639	3.28E+05	1.9167	1.56E+05
0.24378	0.54798	0.61376	84.8	0.66954	11.951	1	1.1038	3.2917	2.55E+05	3	97236
0.2525	0.69988	0.66434	74.433	0.95389	27.019	1	1.537	2.3639	1.64E+05	2.1667	55919
0.27163	0.71642	0.85277	68.6	0.2157	7.8761	1	1.0745	2.2	2.50E+05	2.25	82754
0.25322	0.68058	0.75081	61.8	0.13001	29.121	1	1.0692	2.45	2.50E+05	2.4167	92958
0.24211	0.73141	0.59772	78.8	1.8599	51.68	1	1.0241	3.4639	3.29E+05	3.1667	1.23E+05
0.24823	0.69548	0.59772	78.8	1.9382	3.062	1	1.0241	3.4639	4.10E+05	3.0833	1.54E+05
0.24193	0.53088	0.50286	87.5	0.46421	53.922	1	1.0241	3.4639	2.49E+05	3.1667	85697
0.23413	0.53554	0.51705	88	0.75929	12.711	1	1.0232	3.7167	3.79E+05	3.4167	1.47E+05
0.24503	0.57359	0.6158	87.9	0.59816	25.727	1	1.0241	3.4639	3.24E+05	3.4167	1.16E+05
0.25189	0.68212	0.58737	68.1	0.2936	32.831	1	1.0241	3.4639	2.90E+05	2.6667	1.23E+05
0.2421	0.60837	0.39465	59.8	8.2231	52.097	1	1.404	0.41944	60114	0.16667	28250
0.2809	0.38187	0.72084	84.1	0.30311	34.47	1.0176	1.0089	9.6361	2.84E+06	8.0833	1.03E+06
0.26602	0.43887	0.45014	70.2	0.32646	30.842	1.0481	1.2361	5.3667	1.61E+06	3.9167	6.36E+05
0.22907	0.50653	0.68311	81.1	0.33094	26.611	1	1.0089	9.6361	2.41E+06	7.8333	6.91E+05
0.28969	0.34413	0.74374	80	0.37095	40.005	1	1.1893	1.35	3.28E+05	1.25	75368
0.30475	0.21616	0.92237	21.9	0.23684	4.6112	1	1	5.7472	4.03E+05	5.25	73910
0.22403	0.48831	0.42118	40.6	1.1194	45.546	1	1	5.7472	4.10E+05	5.1667	86014
0.31826	0.15248	0.865	20	0.31308	3.5822	1	1	5.7472	3.99E+05	5.3333	1.06E+05
0.31612	0.25991	0.85252	27.8	0.24376	30.646	1	1	4.1417	4.37E+05	3.8333	1.40E+05
0.27632	0.34715	0.69643	5.6	0.47887	32.665	1	1	3.1278	2.75E+05	2.9167	71632

0.33558	0.17095	0.87919	14.9	0.34951	8.8973	1.0299	1.0916	2.7889	1.79E+05	2.5	32861
0.21956	0.42343	0.72401	37.62	0.69058	35.471	1	1	14.289	1.36E+06	9.75	5.49E+05
0.17994	0.44764	0.51613	38.867	1.6046	35.569	1	1	14.289	1.77E+06	9.75	2.87E+05
0.18607	0.45435	0.48889	40.5	0.18659	28.953	1	1	9.1306	1.01E+06	7.9167	3.62E+05
0.21819	0.42913	0.468	27.4	0.17109	28.222	1	0.99391	13.694	9.00E+05	11.917	2.62E+05
0.2573	0.29967	0.732	35.567	0.81469	45.628	1	1	13.611	1.08E+06	9.5	4.38E+05
0.22286	0.30883	0.27559	38.1	12.94	47.108	1	1	0.84444	43082	0.41667	3434
0.20135	0.42828	0.72533	37.5	1.8644	7.4864	1	1	11.664	1.24E+06	10.083	3.99E+05
0.2078	0.3915	0.69863	43.8	1.3546	39.156	1	1	10.061	9.53E+05	9.0833	3.24E+05
0.2955	0.31381	0.79012	32.4	0.080175	25.083	1	1	10.061	1.05E+06	9.4167	2.59E+05
0.18607	0.45435	0.48889	40.5	2.7243	51.01	1	1	9.0444	8.60E+05	8.75	2.42E+05
0.13125	0.59083	0.79968	63.4	0.17737	23.131	1	1	8.0306	4.32E+05	5.5833	1.79E+05
0.25528	0.41297	0.38384	21	1.1677	19.263	1	1	8.0306	4.74E+05	7.3333	1.29E+05
0.25011	0.35531	0.71269	37.4	0.61194	24.869	1	1	8.0306	4.92E+05	7.4167	1.26E+05
0.30925	0.19347	0.70968	27.9	1.3736	15.177	1	1	4.9028	4.61E+05	4.5833	1.16E+05
0.28347	0.2963	0.72279	31.867	1.5423	20.602	1	1	5.4944	3.23E+05	4.4167	91137
0.27483	0.29618	0.65493	33	2.3231	22.057	1	1.015	5.7472	6.80E+05	5.1667	1.41E+05
0.28701	0.35656	0.66229	29.825	0.65417	28.537	1	1	2.5333	88611	2.3333	22953
0.29963	0.33868	0.43771	29.7	1.251	29.881	1	1	1.0972	38140	0.66667	7756
0.21667	0.3884	0.42088	29.7	0.29598	22.334	1	1	1.5194	41529	0.91667	7520
0.33063	0.18296	1	27	0	5.3289	1	1	2.7028	2.58E+05	2.3333	83815
0.33584	0.1985	0.91064	23.5	0.16977	22.662	1	1	2.6194	2.06E+05	2.3333	63661
0.32956	0.1886	0.93644	27.25	0.13671	19.955	1	1	2.6194	3.15E+05	2.0833	1.20E+05
0.32872	0.15957	0.94001	24.9	0.13142	24.161	1	1	2.6194	3.32E+05	2.5	1.10E+05
0.33457	0.14464	1	25.8	0	21.524	1	1	2.2778	2.54E+05	1.75	1.13E+05

0.34959	0.1951	1	26.8	0	45.067	1	1	2.2778	2.74E+05	2.1667	1.03E+05
0.32998	0.17148	1	25.3	0	39.372	1	1	2.2778	2.75E+05	2.0833	91405
0.33592	0.13054	0.85068	22.1	0.4059	54.808	1	1	2.2778	2.24E+05	1.8333	71434
0.33948	0.21313	0.76926	31.05	0.57317	44.021	1	1	2.5333	1.88E+05	1.4167	86681
0.33383	0.19319	0.9262	27.1	0.13541	17.489	1	1	2.5333	2.01E+05	2.1667	46893
0.26477	0.42047	0.72628	27.4	0.26157	6.6333	1	1.0078	10.736	4.08E+05	9.1667	1.06E+05
0.21794	0.48277	0.74093	38.6	0.58739	29.295	1	1.0078	10.736	5.64E+05	9	1.28E+05
0.27199	0.35113	0.86048	35.75	4.4713	26.948	1	1.0078	10.736	8.32E+05	5.6667	3.99E+05
0.25835	0.4032	0.75	35.6	0.19376	48.274	1	1.0078	10.736	8.93E+05	6.25	3.85E+05
0.31587	0.21861	0.94982	27.9	0.097971	7.4106	1	1	5.7472	4.10E+05	2.25	2.32E+05
0.32285	0.19441	0.5945	29.1	1.8557	15.104	1	1	5.7472	5.82E+05	3.3333	2.38E+05
0.26829	0.30195	0.5957	27.45	1.8601	41.592	1	1.0617	5.4917	3.24E+05	3.25	62143
0.28448	0.262	0.57518	25.683	2.5825	46.221	1	1.048	5.3222	2.60E+05	2.6667	45420
0.2965	0.41177	0.78747	51.875	0.18727	8.5376	1	1.0054	15.892	6.71E+05	13.083	2.06E+05
0.27901	0.45401	0.77077	71.9	0.26302	14.014	1	1.0054	15.892	8.17E+05	12.833	2.77E+05
0.32029	0.36404	0.44584	45.2	1.0777	9.4452	1	1	12.508	3.79E+05	6.75	82388
0.279	0.43626	0.59004	68.3	0.27439	2.4907	1	1	1.7806	1.05E+05	1.3333	32305
0.35098	0.26155	0.57912	29.7	0.30461	4.5685	1	1	2.2778	2.72E+05	1.8333	68264
0.20284	0.62147	0.57861	61.7	0.21092	46.166	1	1	2.1139	1.57E+05	0.91667	26062
0.24522	0.48429	0.66782	57.8	0.8935	15.004	1	1	5.8333	6.05E+05	3.8333	2.17E+05
0.28508	0.42758	0.677	50	0.73918	15.814	1	1	5.8333	6.49E+05	3.5833	2.07E+05
0.27781	0.35907	0.22785	23.7	2.158	20.944	1	1	5.8333	5.62E+05	4.25	1.61E+05
0.35806	0.18265	0.86567	20.1	0.58824	19.904	1	1	5.8333	4.90E+05	3.25	1.55E+05
0.31099	0.27996	0.50135	37.1	0.82767	42.522	1	1	5.8333	6.77E+05	4.1667	2.28E+05
0.28963	0.33237	0.71558	50.2	0.25304	19.336	1	1	5.8333	4.93E+05	4.4167	1.49E+05

0.2034	0.63122	0.60645	62	0.20679	45.672	1	1	1.6056	74677	1	22383
0.33486	0.25955	0.81212	16.5	0.22946	43.056	1	1	6.0028	3.72E+05	4.25	1.19E+05
0.28605	0.38512	0.81361	38.2	0.66511	33.061	1	1.0141	5.9194	2.91E+05	3.8333	75069
0.27543	0.38015	0.52008	39.9	0.71776	12.581	1	1	6.0028	4.77E+05	3.1667	2.49E+05
0.35535	0.24283	0.83333	13.2	0.18018	35.652	1	1	4.6472	5.88E+05	2.6667	2.44E+05
0.31194	0.26158	0.33758	15.7	0.91837	35.082	1	1	3.2917	3.20E+05	1.4167	86442
0.16468	0.62925	0.25382	32.7	1.0037	44.874	1	1	4.4778	4.97E+05	0.5	3.77E+05
0.329	0.26059	0.79474	22.7	0.96331	46.869	1	1	4.3917	4.02E+05	0.41667	3.02E+05