Improving Production Strategies in Unconventional Oil and Gas Reservoirs Through Machine Learning

by

Derek M. Vikara

B.S. Environmental Science, Allegheny College, 2003

M.S. Environmental Engineering University of Connecticut, 2005

Submitted to the Graduate Faculty of the

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Derek M. Vikara

It was defended on

March 24, 2021

and approved by

Dr. Carla Ng, PhD, Assistant Professor, Civil and Environmental Engineering

Dr. Radisav Vidic, PhD, Professor, Civil and Environmental Engineering

Dr. William Harbert, PhD, Professor, Geology and Environmental Science

Dissertation Director: Dr. Vikas Khanna, PhD, Associate Professor, Civil and Environmental Engineering

Improving Production Strategies in Unconventional Oil and Gas Reservoirs Through Machine Learning

Derek M. Vikara, PhD

University of Pittsburgh, 2021

This research involves the application of supervised, unsupervised, and deep learning ML modeling approaches using empirically-derived well completion, production, and geologic datasets from prominent unconventional O&G plays in the U.S. The anticipated outcome of this work is to provide substantial contribution to the knowledge base pertinent to O&G field development and reservoir management approaches (transferable to other subsurface applications) founded in data-driven strategies. ML-based models built through this work complete a multitude of tasks, including: 1) Evaluating potential well production response to various hydraulic fracturing completion designs using a gradient boosting ML algorithm; 2) hierarchical ranking of well design and geologic reservoir quality parameters and their associated interactions on production response by assessing parametric importance and partial dependence; 3) deriving well design strategies that maximize production given well placement through optimization; 4) development of time series-based predictive forecasting capability using long-short term memory neural networks that can generalize temporal or sequence-based tendencies in water and associated gas production trends; and, 5) to enable rapid identification of stratigraphic units within a basin using multiclass classification given total vertical depth and spatial positioning.

The findings from this work show that ML provides fast, accurate, and cost-effective analytical approaches to a variety of O&G-related functions. These strategies can be used to analyze disparate datasets in innovative ways, provide utility in generating new insights, and may

be used in ways to identify improvements over industry benchmarks. They offer robust approaches that can supplement existing reservoir management best-practices and improve the return on investment from field data acquisition.

Table of Contents

List of Tables

List of Figures

List of Equations

Nomenclature

Abbreviations:

% R_o = vitrinite reflectance

ANN = artificial neural network

API = unit of radioactivity used for gamma ray well logs

bbl = barrel

bbls = barrels

Bcf = billion cubic feet

Bcf/day = billion cubic feet per day

CI = confidence interval

CO_2 = carbon dioxide

EIA = U.S. Energy Information Administration

EEMD = ensemble empirical mode decomposition

EUR = estimated ultimate recovery

ft = foot or feet

g/cm^3 = grams per cubic centimeter

GB = gradient boosting

GBR = gradient boosted regression

GBRT = gradient boosted regression trees

GIS = geographic information systems

HH = Henry Hub Spot Price

IP = initial production

IQR = interquartile range

KW = Kruskal-Wallis

Lad = least absolute deviation

lbs = pounds

LHC = Latin hypercube

LSTM = long-short term memory

M = thousand

Marcellus = Marcellus Shale

Mbbls = thousand barrels

mD = millidarcies

ML = machine learning

MLP = multilayer perceptron

MM = million

MMBtu = million British thermal units

MMcfd = million cubic feet per day

MMcfge = million cubic feet of gas equivalent

MPa = megapascal

MSE = mean squared error

NPHI = neutron porosity

O&G = oil and gas

Ohm-m = Ohm – meter

Pearson r = Pearson correlation

psi = pounds per square inch

R^2 = Coefficient of determination

rbf = radial basis function

RF = random forest

RFECV = recursive feature elimination with cross-validation

RHOB = bulk density

RMSE = root mean square error

RNN = recurrent neural network

scf = standard cubic feet

SMOTE = Synthetic Minority Oversampling Technique

SSE = sum of squared errors

stdev = standard deviation

SVC = support vector machine classification

tcf = trillion cubic feet

TOC = total organic carbon

U.S. = United States

USD = United States Dollar

WTI = West Texas Crude Spot Price

Mathematical Symbols and Variables

α = Significance level

β = Expansion coefficient

γ = Step length or gamma parameter for support vector machine classification

ϕ = average Marcellus Shale porosity (fraction [based on neutron porosity])

ϕ_a = pore space occupied by adsorbed gas (fraction)

ϕ_m = matrix porosity (decimal)

ϕ_{frac} = fracture porosity (decimal)

ϕ_{stdev} = Marcellus Shale porosity standard deviation (decimal)

ρ_b = bulk density of shale (g/cm^3)

υ_a = specific volume of gas absorbed per unit mass of shale (standard ft^3/ton)

σ = sigmoid activation function

ν = Shrinkage parameter

μ = mean

$First\ 12_w$ = First 12-months production (MMcfge)

a = Input variable gradient boosting parameters

A = drainage area (acres)

A_1 = additive per foot (bbls / foot)

A_2 = wellbore azimuth trajectory (degrees)

A_3 = acre spacing (acres)

b = neural network bias or b-factor per the Arps model

B_g = formation volume factor (reservoir ft^3/standard ft^3)

C = cluster centroids or cost margin function for support vector machine classification

C_t = LSTM cell state

$d(a_i, c_k)$ = distance between data points and cluster centroids

D_1 = top of Marcellus Shale depth (ft below ground surface)

D_2 = average Marcellus Shale density (g/cm^3)

D_{2stde} = Marcellus Shale density standard deviation (g/cm^3)

D_i = initial decline per the Arps model (fraction/month)

f_t = LSTM forget gate

G = normalized average Marcellus Shale gamma ray (API)

G_{stdev} = Marcellus Shale normalized gamma ray standard deviation (API)

h = Weak learner (in GBRT context) or prediction horizon (in forecasting context)

h_s = shale net pay zone thickness (ft)

h_t = LSTM cell output

H_o = null hypothesis

H_1 = alternate hypothesis

$H(K)$ = Hartigan's Index

i_t = LSTM input gate

k = decision tree or number of folds in cross-validation

K = total number of decision trees or total number of clusters

l = decision tree node

L = loss function

L_p = perforated interval length (ft)

m = month

max_x = maximum value in an x feature set

min_x = minimum value in an x feature set

$n_{samples}$ = number of samples in a dataset

N =length of dataset or pseudo well count

$OGIP$ = original gas in place (scf)

P = proppant per foot (lbs / foot)

p_m = production value for a given month (MMcfge)

q = monthly oil production per Arps model (bbls per month)

q_i = initial oil flow rate per Arps model (bbls per month)

R = x-feature space region for decision terminal nodes

relu = rectified linear unit function

R_{deep} = average Marcellus Shale deep resistivity (Ohm-m [via deep induction logging])

R_{stdev} = Marcellus deep resistivity standard deviation (Ohm-m)

S = separate clusters

$SS_{Regression}$ = regression sum of squares

SS_{Total} = total sum of squares

S_w = saturation of water (fraction)

t = production month (month)

tanh = hyperbolic tangent function

$Top\ 12_w$ = top 12-months production (MMcfge)

U = weight vector for recurrent component of LSTM layer

Var = variance

W = water per foot (bbls / foot) or neural network weight vector

W_k = within-cluster sum of squares

w = well

x = input / predictor variable

$\tilde{x}$ = population median

$x_{normalized}$ = x values normalized between 0 and 1

y = response variable

$\hat{y}_i$ = simulated / predicted response variable

Z = Z-values

Z_t = LSTM candidate values for the cell state

Unit Conversions

The units used throughout this dissertation are commonly used industry standards for the oil and gas sector in the United States. Conversion factors to the international system of units are as follows:

1 barrel = 0.1589 cubic meters

1 pound = 0.4536 kilograms

1 foot = 0.3048 meters

1 cubic foot = 0.0283 cubic meters

1 acre = 4,047 square meters

1 square mile = 2.5899 square kilometers

1.1023 ton = 1 tonne

Acknowledgements

I dedicate this work to those who have supported me in this academic venture. I would like to give my deepest thanks and gratitude to several specific individuals:

- To Dr.'s Carla Ng and William Harbert for serving on my advisory committee. While our interactions have not been overly extensive, the insights and knowledge I've gained from each of you has been anything but. I greatly appreciate the support you both have given me. I can point to several places in this dissertation that were founded on a concept, resource, or idea that I would have otherwise overlooked without your influence.
- To Dr. Radisav Vidic, many thanks for taking a chance on me as a part-time student. I am sure, at a first glance, my situation as a full-time employee elsewhere appeared a bit unorthodox and probably seemed risky. However, personally, I couldn't imagine any another graduate program being as nearly as amenable to my particular circumstance as Pitt's Civil & Environmental Engineering Program has been. I hope I haven't let you down and your investment in me has been worth it.
- To Dr. Vikas Khanna, thank you for being the guiding light through this undertaking. I couldn't imagine having worked with any other advisor. You've made this process extremely enjoyable, as straightforward as it possibly could, and, mostly importantly, exceedingly beneficial (for me at least!). I could not be prouder of the work products we were able to put together along the way. I am sure this is just the start to us collaborating on exciting research.

- To my Dad and late Mom, thank you both for 1) affording me the opportunity to pursue higher-level education and 2) instilling in me a "never quit" mentality and spirit. I needed both to get through this. I couldn't have gotten to this point without you.
- To my daughter Lilly and son Corey, thank you guys for being my inspiration. You have kept me always cognizant of the important things happening in life aside from work, school, and homework as I undertook this process. I know that school has taken a lot of my time, but you both have been so patient along the way. I hope …by me doing this…that maybe, it might someday help to inspire you to not be afraid to try something new you weren't sure you could take on or manage. While I am extremely happy with the body of work in this dissertation, you guys will always be the creations I'm most proud of.…by far!
- To my wife Amy, any accomplishments I might be awarded from all of this are as much yours as they are mine. I literally could not have done it without you. Thank you for enduring the extra burdens as both a wife and as a mother during this process. I know it wasn't easy…but I think we are coming out on top. Also, you are the only one who never suggested that me giving up on this was a viable consideration. You always believed in me, this process, and its significance more than anyone. However, I'm still not buying you a new kitchen! :)

  And as for me, this is just the beginning…

1.0 Introduction

The increasing demand for reliable, affordable, and secure domestic supplies of energy amplify the need for continued research into ways to economically and efficiently access our Nation's vast unconventional natural gas and oil resources. These types of low permeability or "unconventional" reservoirs (described as tight source rocks containing organic rich matter that has reached thermal maturity and is absent of hydrocarbon migration) are geologically complex and heterogeneous on a variety of scales; from basin scale, to reservoir scale, to core scale and to pore scales [1]. The application of horizontal drilling and hydraulic fracturing techniques in oil and gas (O&G) production has revolutionized the energy system of the United States (U.S.) [2] and has been the leading driver in growth in domestic natural gas and oil production. The relatively rapid expansion of U.S. unconventional O&G gas resources over the last several years has resulted in low gas and oil prices not seen for over a decade (Figure 1) [3]. The benefits have been an emergence in new business markets, lower greenhouse gas emissions, and an increase in the security of U.S. energy resources.[1] For instance, the U.S., in particular, is benefitting from some of the lowest prices for natural gas in the world due to the growth in natural gas production, primarily from shale gas. Industries reliant on natural gas have seen overall costs drop, and some have touted low natural gas prices as the main reason for a manufacturing revolution in the U.S. [3, 4]. Some companies have begun to make major investments to take advantage of the low natural gas prices. Examples exist in the petrochemicals industry [2] as well as in the U.S. electric power

[1] This circumstance is separate from recent events (March 2020 through the January 2021 timeframe) related to natural gas (and oil) demand reduction impacted by the COVID-19 epidemic.

sector, which has seen fossil-based energy generation transition from coal-fired to natural gas-based generation [5].

While significant advances in horizontal drilling and hydraulic fracturing technology were made during the "shale revolution," these technologies are still unable to recover large portions of the natural gas and oil in place [6]. Additionally, the world has more recently experienced an abundance of natural gas and oil supply versus demand for some time now as new supplies and suppliers have entered the global market. Absent supply reductions, lower global demand translates to reduced commodity prices in competitive hub pricing and on spot prices [7]; suppressing the remaining gas (or oil) assets that can be economically recovered.

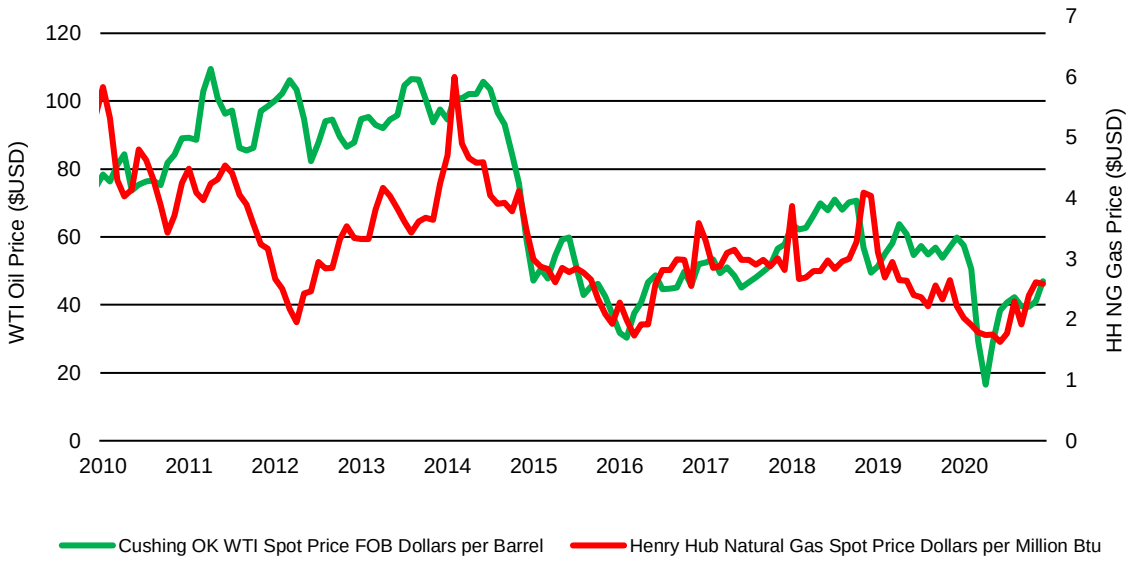

Figure 1. Snapshot of recent and historical O&G market prices in the U.S., limited to West Texas Crude Spot Price (WTI) and Henry Hub Spot Price (HH). Data acquired from U.S. Energy Information Administration (EIA) **[8, 9]**.

Regardless, energy-focused forecasts, including the EIA's Annual Energy Outlook 2020 [10], project continued growth in the production and utilization of natural gas (particularly from

shale gas resources in the eastern U.S. [i.e., Marcellus and Utica Shales]) and crude oil (particularly from the Permian Basin in the southwestern U.S.) through 2050 (Figure 2). The realizations of these EIA forecasts in Figure 2 are contingent upon the occurrence of several technical improvement and economic/market conditions – some of which include:

- Increasing exports of crude oil, petroleum products, and liquefied natural gas from the U.S.;
- Natural gas spot prices at Henry Hub increase from $2.56 to $3.70 per million British thermal units (MMBtu) from 2019 through 2050;
- North Sea Brent crude oil prices increase from $63 to $104 per barrel (bbl) from 2019 through 2050; and
- Technological advancements (in the range of 1 to 6 percent per year) and improvements in industry practices which result in lower production costs (1.5 percent per year) for shale/tight hydrocarbon development in the EIA Reference case and an increase the volume of oil and natural gas recovery per well [11].

Given the suppressed O&G economic climate (December 2020 values: Henry Hub Natural Gas Spot Price - $2.59 per MMBtu; West Texas Intermediate and Brent Crude Oil Prices- $47 per bbl), novel and cost-effective approaches that can supplement existing reservoir and field management strategies and potentially help improve recovery may be needed to keep pace with energy production forecasts like the EIA's Annual Energy Outlook 2020 and assure sustainability in unconventional field development moving forward. The absence of cost or efficiency improvements in unconventional O&G development may pose substantial risks to both national energy security, significant local economic impacts, less than effective use of the national oil and gas resource asset, and can be detrimental to the future investments and effective buildouts of energy infrastructure into the future.

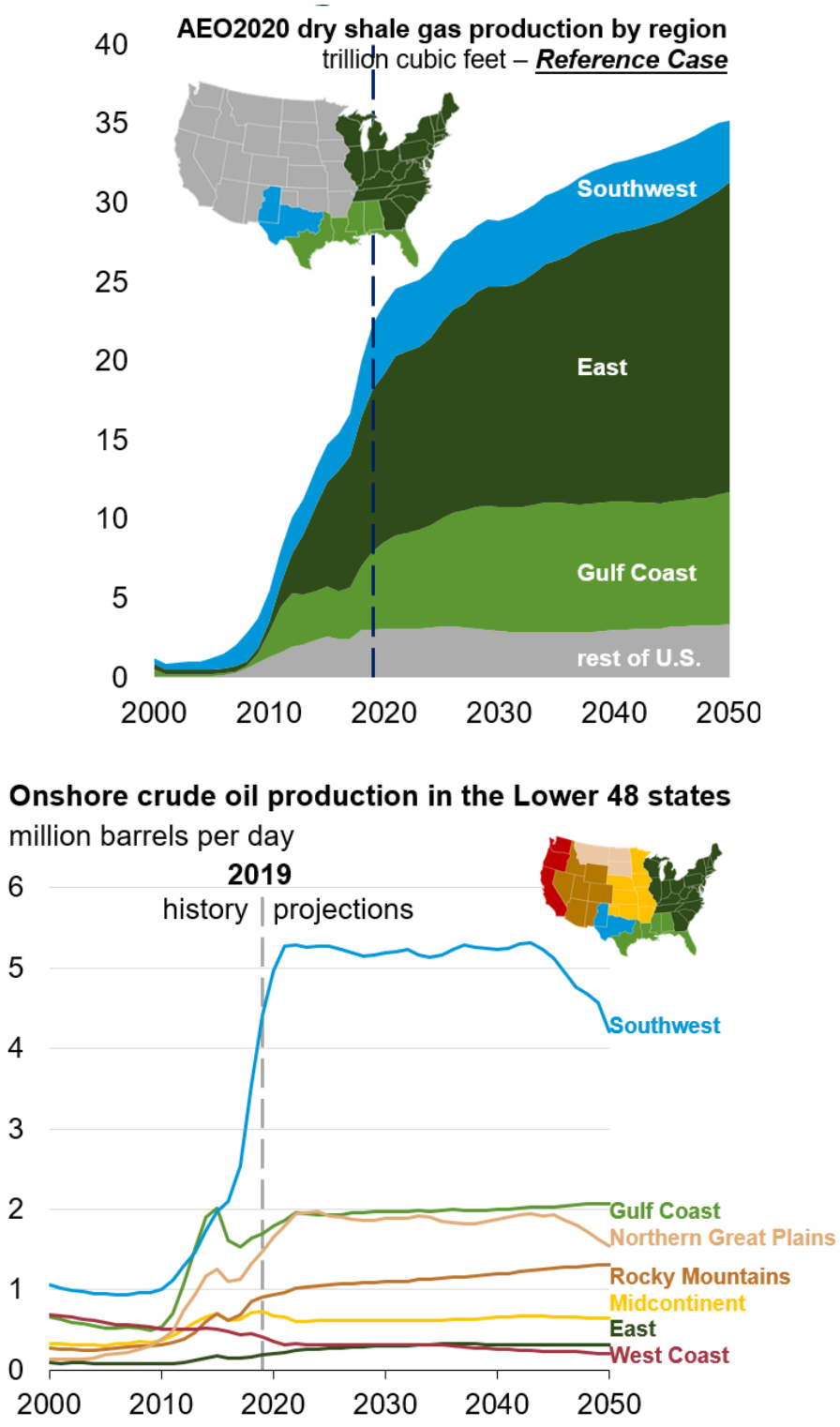

Figure 2. EIA's Annual Energy Outlook 2020 Reference Case showing dry gas production outlooks through 2050 (top) and onshore crude oil prodution outlooks through 2050 (bottom) **[10]**.

Sustained suppression of market prices for O&G can diminish the economically-viable portions of remaining O&G assets. Under these circumstances, solutions would be encouraged that can improve productivity and lower costs and/or increase operational efficiency. It's been mentioned that hydrocarbon production-related performance improvements in unconventional O&G development is expected to occur through tailored well design and completion strategies specific to the geologic conditions for which new wells may be placed [12, 13, 14]. In this pursuit, operators can benefit by considering approaches that can be both cost-effective and insightfully-abundant in order to holistically evaluate the critical factors associated with both well design choices and the geologic conditions (including interactions) of potential drilling areas when considering field developing decisions [15]. Advancements in the computational power that are now widely available coupled with the emergence of digital datasets in the O&G sector creates a unique opportunity for machine learning (ML) to be applied for complex subsurface energy system applications. The intersection of these resources may facilitate the emergence of new strategies that could transform the way subsurface energy systems, including unconventional O&G development, are evaluated. While the applications of ML approaches are anticipated to help towards improving understanding of the phenomena and processes occurring in the subsurface, it also presents a challenge given the large amounts of data that must be collected, transmitted, stored, and processed.

Machine learning has gained substantial interest as an innovative approach that can help address challenges facing O&G development. The technology may offer new techniques worthy of consideration by field operators in the pursuit of lowered cost and improved recovery objectives. The unconventional O&G arena has seen research take focus on applying ML-based techniques in various capacities. These include generalizing unconventional well productivity and informing

well completion designs [16, 15], improving well drilling operation practices [17, 18], characterizing lithology and geologic facies from well log data [19, 20, 21], detection of faulting in the subsurface using seismic data [22], operational predictive maintenance planning [23, 24], and detection of potentially high producing pay zones sweet spots via integration of disparate sources of data [25, 26]. The studies listed under each topic (as well as others not referenced here explicitly) provide innovative aspects that serve as a solid foundation for formulating new ML-based research aimed at improving exploration and production efficiencies in unconventional O&G development.

1.1 Background on Machine Learning

Machine learning, a subset of artificial intelligence, has proved advantageous for use across various industries. These technologies help enhance business practices through the utilization of data. They are based upon the use of statistics-based algorithms that enable systems to learn automatically and improve from experience without being directly programmed. Machine learning aims to build model representations of systems of interest by learning from data provided. Once ML-based models are trained by observing data, they can be used to make predictions on unseen data examples that are within the same applicability domain. The ML model development process is a stark contrast from traditional rules-based modeling. A representation of each concept, for comparison, is provided in Figure 3.

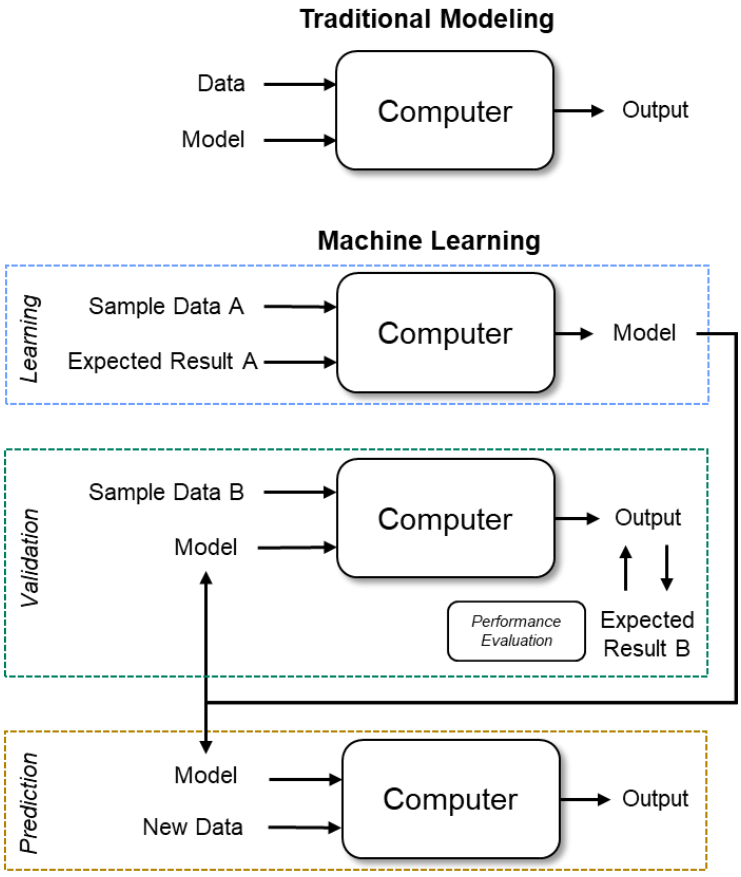

Figure 3. Example dipction of the difference in approachs between traditional predictive modeling (top) and machine learning (bottom). Adapted from Zeiss **[27]**.

Machine learning approaches typically fall under three prominent "learning" categories as follows: supervised learning, unsupervised learning, and reinforcement learning [28]. Each category is aimed towards a particular application depending on the underlying problem evaluated. Figure 4 provides a conceptual overview of the ML concept and the associated learning categories, applications, and problem from which each could be applied. Supervised learning is based on the development of models where training datasets are comprised of classified input-output data pairs that have been assigned under human supervision. Data applicable to supervised learning is often called "labeled" data. These labels apply to known attributes or features within the given dataset

[29]. Specific ML algorithms learn from labeled input or predictor variables (x) to generate a mapping function for labeled output or response variable (y) (i.e., $y = f(x)$). One of the main goals of supervised learning is to develop models that best approximate (i.e., generalize) the mapping function between predictor and response variables so that when new predictor data (x) are applied, a reliable prediction of the response (y) can be achieved [30]. Supervised learning is further subdivided into the applications per Figure 4 that includes 1) regression, where the response variable is typically continuous in nature, or 2) classification, where the response variable is typically categorical in nature. However, for both regression and classification applications, the labeled predictor variables (x) may be either continuous or categorical, or a combination of the two [31].

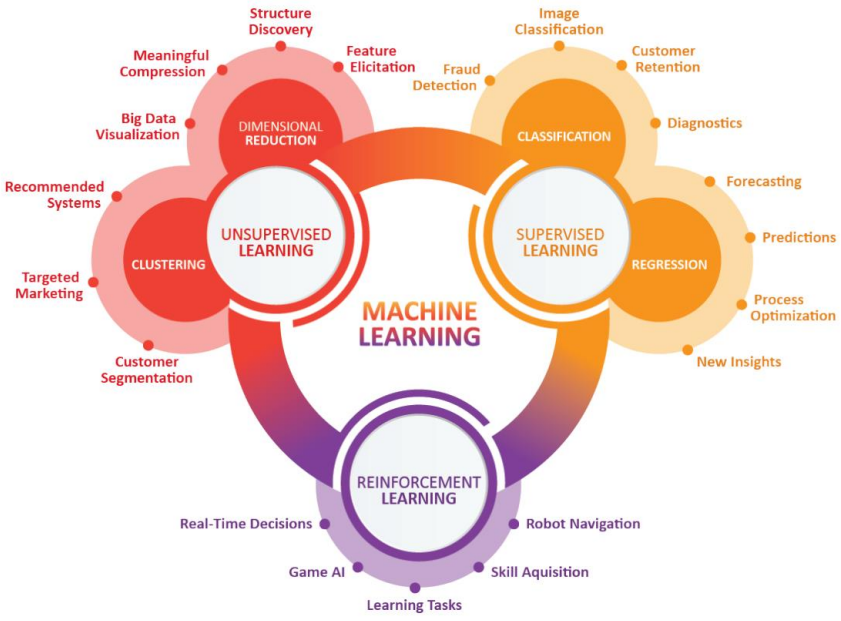

Figure 4. Schematic of machine learning categories and potential application examples. Figure sourced from Bettin et al. 2019 **[32]**.

Unsupervised learning differs from supervised in that it is applied through "unlabeled" data and algorithms are intended to identify patterns heuristically [28]. For instance, there are no predefined targets or responses for the attributes or features being studied (i.e., no explicit response variable (y) is predetermined). The goal in unsupervised learning cases is to enable algorithms to explore datasets and try and identify commonalities between attributes or features, as well as potential structures in the data. Since there is no targeted answer or objective, algorithms are empowered autonomy to discover commonalities and data structures (hence the "unsupervised" namesake) [29, 28]. Unsupervised learning applications can consist of 1) clustering, where the identified commonalities may define natural groupings of features or attributes, as well as 2) dimensionality reduction, where the number of features or attributes (i.e., dimensions) in the dataset under consideration can be reduced or downsized through a set of principal variables [33].

Reinforced learning, the third category of ML outlined in Figure 4, involves learning within a specified environment via interactions and feedback based on a rewards system [34]. Supervised and unsupervised approaches are implemented as part of this dissertation, but reinforced learning strategies are not. Therefore, reinforced learning is not discussed at length here.

The development of ML models typically follows a process of distinct and iterative steps that consists of problem classification (i.e., determining a modeling objective) through data collection, algorithm selection, validation, and model evaluation. A conventional process workflow for a ML development application is presented in Figure 5, which highlights prominent stages and sequence of events [31, 35]. However, despite the relatively linear process as depicted, the progression may, in practice, require iterations depending on realized versus expected results at any given step.

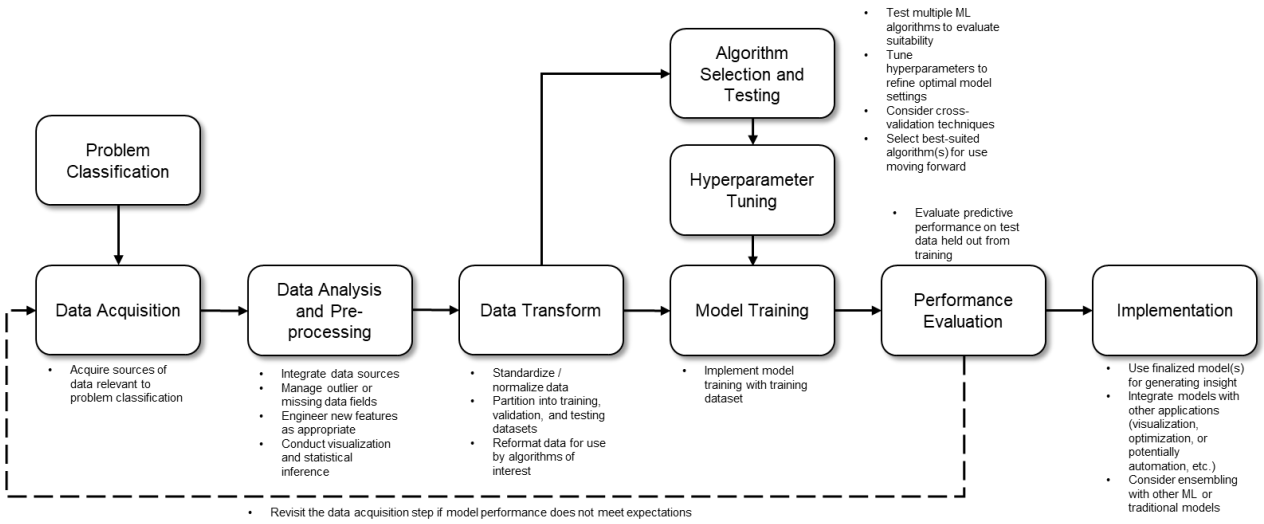

Figure 5. Example of a workflow scematic depicting common stages related to the development of data-driven ML models. Concept based on workflow proposed by Thallam and Dominguez (2019) **[36]**.

In general, ML provides a benefit of enabling analysis of large volumes of data. Therefore, it is gaining popularity and wider-spread use in areas (applications outlined in Figure 4) where digital datasets are becoming available in large quantities, generated rapidly at high velocities, and patterns or relationships of interest may be difficult to represent using more conventional approaches.

1.2 Research Objectives

With the shale revolution, operators are producing unprecedented amounts of oil and gas from unconventional resources; however, only a small percentage of the resource in place is being recovered [37]. Additionally, the increased supply of natural gas and oil has also reduced commodity prices, straining operators who produce those resources to do so profitability. Many

have noted the potential for new, innovative possibilities in the O&G arena targeted at challenges pertaining to improving recovery and operational cost-effectiveness may occur through expanded use of digital empirical datasets and applying ML and data analytics to help improve well performance and operational planning moving forward [38, 39, 40, 31]. These types of data-driven approaches could provide added utility in cases where sufficient understanding of the overarching complexities of unconventional reservoirs is lacking; but the need exists to construct models capable of accurate and reliable replication of complex systems, or when expeditious predictive capability may be needed. They have proven effective in accurately modeling circumstances involving highly complex systems where variable conditions exist – not uncommon to unconventional O&G domains given the inherent relationship complexities between wellbore design, completion and stimulation processes, and prevailing geologic conditions of targeted reservoirs. The application of such approaches in unconventional O&G applications specifically could be aimed toward developing new insights, enabling improved understanding, help toward optimizing well/reservoir interactions [38], and provide a robust and cost-effective approach to supplement existing reservoir management best-practices.

The research conducted in this dissertation applies ML to large unconventional O&G datasets with the aspiration of contributing towards the knowledge base relevant to O&G field development and reservoir management strategies through data-driven approaches.[2] The specific goals of the research conducted through this dissertation include: 1) Developing accurate, multivariate, data-driven models that can generalize well performance in unconventional O&G

[2] The work implemented is not intended to replace detailed reservoir modeling with ML approaches explicitly, but instead consider them as an additional component to reservoir management strategies moving forward.

settings; 2) evaluate the impact of well design attributes, geologic properties, and their associated interactions on productivity in major unconventional O&G plays to assess productivity drivers; 3) implement ML-based models into a framework that can be used to inform future well design that maximizes productivity based on specific placement across a spatially heterogeneous unconventional O&G reservoirs; 4) develop data-driven modeling capability that estimates the volumes of multiple fluid types produced in tandem as part of hydrocarbon development – useful for informing field development strategies based on the volumes and quantities of produced fluids in order to effectively manage, treat, or potentially reuse produced fluids; 5) provide a means to supplement hydrocarbon production outlooks with associated fluid volumes as a function of time; and, 6) outline a methodology for implementing a ML-based subsurface formation identification tool that can be used in delineation tasks. A summary of the specific research objectives and the corresponding chapters of the dissertation in which they are addressed are provided in Table 1.

Table 1. Overview of dissertation research objectives and the chapter in which they are discussed

Research Objective		Chapter Title / Journal Paper	Dissertation Chapter
1	Develop a ML-based predictive model built on a gradient boosted regression tree (GBRT) algorithm capable of accurate generalization of productivity for horizontal wells in the Marcellus Shale using data commonly available over large spatial scales	Vikara, D., Remson, D., and Khanna, V. Gaining Perspective on Unconventional Well Design Choices through Play-level Application of Machine Learning Modeling. Upstream Oil and Gas Technology. 2020. Volume 4, https://doi.org/10.1016/j.upstre.2020.100007	Chapter 2
2	Evaluate the effects of predictors on model production response for the tree-based models developed		
3	Introduce a framework that ensembles a data-driven predictive model capable of accurate estimation of production with a well design optimization approach that maximizes well productivity	Vikara, D., Remson, D., and Khanna, V. Machine Learning-informed Ensemble Framework for Evaluating Shale Gas Production Potential: Case Study in the Marcellus Shale. Journal of Natural Gas Science and Engineering. 2020. Volume 84, https://doi.org/10.1016/j.jngse.2020.103679	Chapter 3
4	Classify the Marcellus region into distinct productivity bins or "grades" from high to low based on threshold cutoffs for simulated productivity from ML models		
5	Evaluate similarity or disparity of the best well design choices that maximize well productivity and geologic properties of the Marcellus within associated rock quality bins		
6	Develop a reduced order model that couples well design and geologic data into a single analytical method that can estimate production at new Marcellus Shale well sites without the reliance on ML		
7	Develop a deep learning-based data driven modeling framework that enables formulation of joint prediction capability for associated gas and water produced alongside oil in the Permian basin region	Application of a Deep Learning Network for Joint Predication of Associated Fluid Production in Unconventional Hydrocarbon Development	Chapter 4
8	Generate oil, water, and gas production outlooks for various combinations of well designs and placement options that can be used to inform basin-level fluid production forecasting		
9	Explore multiple ML classification-based algorithms and evaluate their effectiveness for identifying specific stratigraphic units (i.e., formations) as a function of total vertical depth and spatial positioning in a well-developed geologic basin	Machine Learning Classification Approach for Formation Delineation at the Basin-scale	Chapter 5
10	Outline a data-driven model development framework that can be adapted for subsurface resource delineation or characterization across multiple domains, including oil and gas, geothermal, carbon dioxide storage, or environmental applications.		

1.3 Significance and Technical Implication of Completed Research

Developing and deploying ML technology in subsurface applications like unconventional O&G exploration and production has the potential to provide accurate, efficient, and cost-effective analytical methods that may transform how future subsurface energy resources are utilized into a much more data-driven science [35]. Insights gained from deploying ML as a novel compliment to reservoir management may enable improved understanding and insights of how engineered subsurface systems perform; thereby subsequently reducing the risk, improving the safety, and increasing the effectiveness of developing said resources. For instance, ML-based models have the potential to reduce time needed for reservoir simulations from days to seconds [32], which would provide a fast and reliable complement to the time-consuming operations performed with typical physics-based reservoir simulation models more commonly used. As a result, many more modeling "what if" scenarios can be implemented and evaluated using ML (when the availability of viable datasets exist) versus traditional simulation approaches – a concept depicted commendably by Abubakar in Figure 6. Therefore, there is a significant opportunity to gain more understanding of optimal field development approaches that result in the most prudent utilization of subsurface energy resources through new data-driven strategies. Additionally, ML approaches can be applied at various scales (i.e., basin-scale to the well level), thereby offering utility to evaluate various types of subsurface challenges.

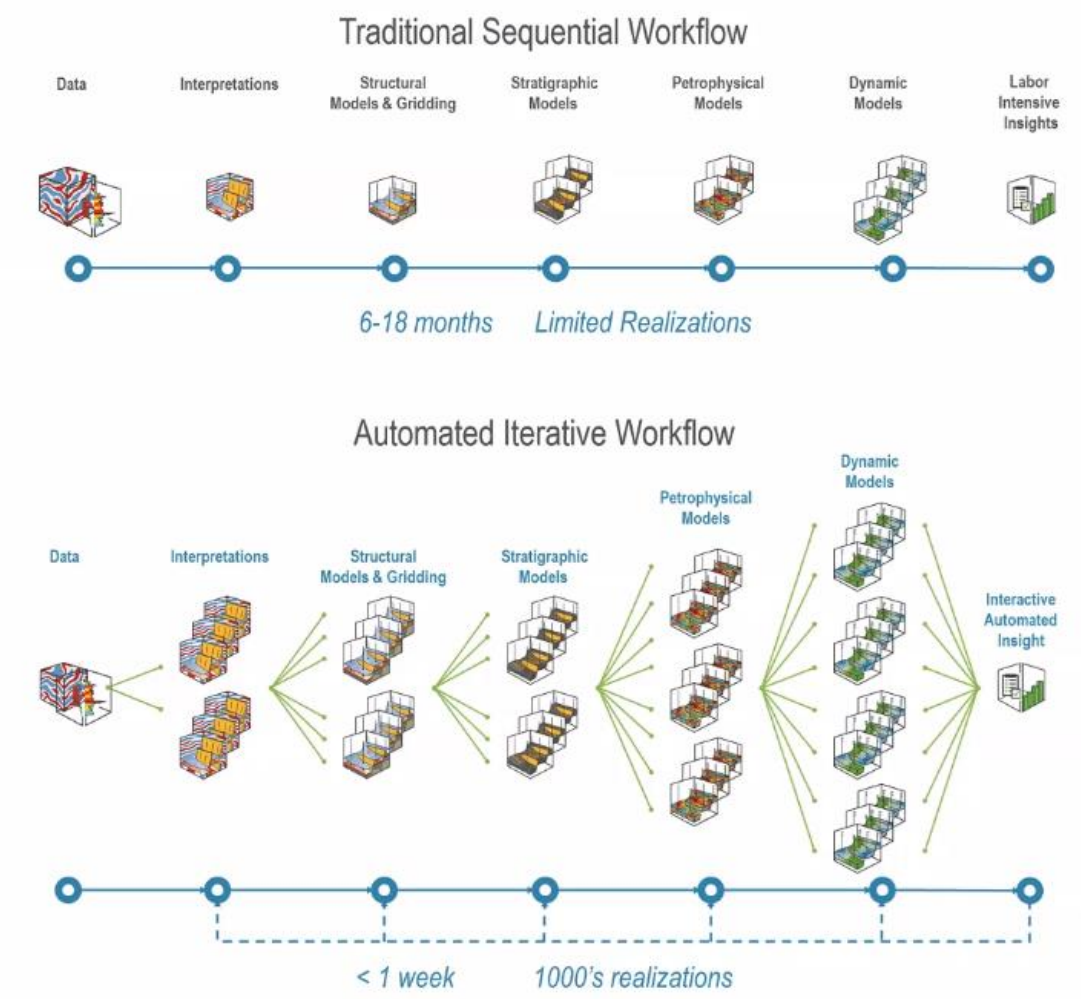

Figure 6. Conceptual of traditional versus automated (i.e., machine learning-based) reservoir evaluation workflow. Sourced from Abubakar **[41]**.

Beside the capability to perform data-driven modeling, analysis of subsurface systems using ML also provides capability, tools, and approaches that may facilitate the generation of new insight and knowledge pertaining to: 1) Acquiring and managing data in large volumes, of different varieties, and generated at high velocities; and, 2) the use of statistical techniques to thoroughly analyze the data and detect hidden patterns and associated relationships in large, complex, multivariate datasets [42, 43, 31]. The combination of these benefits may facilitate data-driven

insights for both understanding and optimizing the performance of engineered subsurface systems [43]. A critical next would involve industry adoption of the notable outputs generated from such approaches, tools, and/or new functionalities so they may be implemented into practice and generate performance improvements in the field. For instance, outputs from ML-based analyses could take multiple forms, including, but not limited to, supplemental data for existing physics-based models, used explicitly to inform field or operational development decisions by operations, or stored or fed into other decision-support and/or situational awareness systems (sharing aspects of transfer learning). In regards the latter point, ML applications have enormous potential to integrate with other capabilities, including enhanced visualization (critical for enabling "human-in-the-loop" functionality), optimization of modeled systems, and potentially autonomous monitoring system capability [44].

The goal of this proposed work is to evaluate region-specific industry performance data through time and attempt to identify approaches conducive to improving the recovery of hydrocarbons in unconventional reservoirs, as well as evaluate the implications in terms of resulting fluid production given well design and placement decisions. Much of the proposed work focuses on the application of ML to large well datasets that span predominant O&G plays in the U.S.; most notability the Marcellus Shale (Marcellus) in the Appalachian Basin and in the "Wolfberry" payzones (typically considered the Upper Spraberry through Cisco/Cline [Wolfcamp D] reservoirs) of the Permian Basin.

The International Energy Agency (2017) estimates that widespread use of digital technologies like ML could, for instance, increase O&G reserves by about 5% and reduce production costs by 10-20% [45]. The realization of this level of improvement over current best practices would be substantial towards more prudent development of our nation's oil and gas asset

base through 1) improved recovery efficiency of development operations and 2) an expansion in economically viable hydrocarbon resources that could be developed. Findings from this dissertation research provide insight associated with the interaction of specific well designs and spatially-distinctive geology in the Marcellus Shale, the parameters most influential in terms of model impact and production, as well as a novel, data-driven approach for sweet spot identification, grading and ranking productivity across plays, and identifying well designs settings that maximize productivity based on their placement across the Marcellus. Additionally, this research provides a data-driven approach for a more holistic evaluation towards field development in the Permian Basin where multiple producing reservoir options are co-located, and where unique challenges facing the O&G industry exist related to associated gas and water production. Overall, the knowledge, data, and resources gained and provided through this research (Section 1.4 below) should be of interest to those in industry, academia, government, and otherwise interested in leveraging data-driven approaches to better understand and potentially improve the way unconventional O&G resources can be evaluated. Additionally, the data parameters utilized are relatively common across multiple O&G plays and may be readily acquired from public sources – therefore there is transferability for the frameworks discussed in this dissertation across O&G plays. Developing and deploying ML technology in O&G applications has the prospective to provide numerous global efficient and accurate analytical toolsets that can complement existing best-practices—a combination that can potentially revolutionize how wells are sited, designed, and operated moving forward.

1.4 Dissertation Research Products

The body of work this dissertation encompasses takes form in peer-reviewed journal articles, scholarly research (not published at the time this dissertation was completed), and tangible products (including digital datasets, analytical models, and novel analytical strategies) that can be used by others interested in building upon the work created here. Additionally, portions of the research discussed in this dissertation was selected for presentation to elected officials in Pennsylvania as part of the University of Pittsburgh's 2020 Pitt Day in Harrisburg (Appendix A), as well as presented in various capacities to federal and contractor staff, as well as visitors, at the National Energy Technology Laboratory.

Peer-reviewed journal articles:

1. D. Vikara, D. Remson, and V. Khanna, "Gaining Perspective on Unconventional Well Design Choices through Play-level Application of Machine Learning Modeling," *Upstream Oil and Gas Technology,* Volume 4. 2020. (Chapter 2 in this dissertation)
2. D. Vikara, D. Remson, and V. Khanna, " Machine Learning-informed Ensemble Framework for Evaluating Shale Gas Production Potential: Case Study in the Marcellus Shale," *Journal of Natural Gas Science and Engineering,* Volume 84, 2020. (Chapter 3 in this dissertation)

Scholarly research:

1. D. Vikara and V. Khanna, "Application of a Deep Learning Network for Joint Predication of Associated Fluid Production in Unconventional Hydrocarbon Development." (Chapter 4 in this dissertation)

2. D. Vikara and V. Khanna. “Machine Learning Classification Approach for Formation Delineation at the Basin-scale.” (Chapter 5 in this dissertation)

Tangible products developed:

1. Detailed description of ML-based workflows that can be used to evaluate hydrocarbon and/or water production (both as static/cumulative variables, or in time series) applicable to several oil and gas plays in the U.S. and around the globe.
2. A novel response variable (Top 12-months production) translatable to any unconventional oil or gas play that can better capture productivity potential for wells where production may be interrupted or well designs have been modified.
3. A gradient boosted regression tree-based modeling framework and hyperparameter parameter settings capable of generalizing productivity potential in horizontal Marcellus Shale wells.
4. Gradient boosted regression tree -model generated simulation datasets run across the entire Marcellus productive region under various well design combination (Standard and Tailored designs) that estimate the Top 12-month production estimate. The Tailored design data includes well design parameters that result in the highest Top 12-month production estimate depending on placement across the studied area.
5. A compilation of well log data (51 in total) across the Marcellus Shale interval that are publicly available and includes bulk density (RHOB in g/cm^3), gamma ray (API), neutron porosity (NPHI in percent or decimal), and deep resistivity (Ohm-m) – parameters common to the widely used Triple Combo well log tool string.
6. A reduced order predictive model developed using a multiple linear regression approach which can estimate production at new Marcellus Shale well sites without having to employ

the GBRT machine learning model by the user. This model couples well design and geologic data into a single analytical method so that they can be evaluated in tandem. The reduced order model provides a simplistic and efficient option for evaluating potential well performance based on the specific design choices given known geologic conditions. The model should be helpful in informing future well designs given access to relatively common geologic data. Its linear formulation also enables potential well design parameter optimization.

7. Digest of well completion and associated three-stream production outlook attributes in compilation for the Midland Basin of west Texas. The digest serves as a guiding resource for assessing the potential volumes of produced fluids associated with oil production in the Midland Basin based on well completion design considerations and placement within the basin.
8. Policy Implications based on research findings – Demonstration of the need for continued investment in R&D to develop economically and environmentally prudent ways to access our Nation's vast fossil energy resource base.

1.5 Organization of Dissertation

This dissertation is organized into five chapters followed by three appendices. An overview of each is discussed below:

Chapter 2 is based on application of ML to a large dataset in order to develop models capable of accurate prediction of productivity indicators at the well level that strongly correlate to

estimated ultimate recovery (EUR) (in gas equivalents). The analysis focuses in the Marcellus Shale, a prominent unconventional gas-producing reservoir. The models developed provide for a fast and effective evaluation capability of the impact of various well placement and design choices in the Marcellus Shale. A series of analyses were conducted to 1) test and evaluate model performance and 2) use the developed models to explore the impact and overall effects of predictor parameters on well productivity. The models are used to explore well design optimization strategies that, in hindsight, may have improved in-field well design choices.

Chapter 3 introduces an ensembled framework that couples a data-driven ML predictive model capable of estimating a productivity indicator for unconventional O&G horizontal wells that correlates to EUR with a well design optimization approach that maximizes productivity. The framework is tested and results discussed when applied to the producing extent of the Marcellus Shale. This framework is implemented to generate insights towards identifying the high-priority drilling regions based on productivity potential, as well as informing the tailoring of future well designs to maximize productivity given their placement in the Marcellus and associated controlling geologic conditions.

Chapter 4 presents a combination of supervised and unsupervised ML approaches as part of a framework for the joint prediction of produced water and natural gas volumes associated with oil production from unconventional reservoirs in a time series fashion. The work focuses on the pay zones within the Spraberry and Wolfcamp Formations of the Midland Basin in the U.S.; a region with enormous oil producing potential, but burdened with the management of substantial volumes of water and gas (much of which is flared) produced alongside oil. The ensemble of the supervised and unsupervised elements of this work facilitates a means to forecast oil, water, and natural gas production at the well level as influenced by specific development considerations. Well

level three-stream production volumes can be leveraged to help support the formulation of management and/or remedial strategies based on the volumes of fluids expected from unconventional O&G development operational conditions.

Chapter 5 outlines a framework for generating predictive models using multiple ML classification-based algorithms which can identify the specific stratigraphic units (i.e., formations) as a function of total vertical depth and spatial positioning. The framework is applied in a case study to 13 specific formations of interest (Upper Spraberry through Cisco/Cline [Wolfcamp D] reservoirs) in the Midland Basin, West Texas, United States. The framework is intended to generate classification models that can be applied as resource delineation tools in domains spanning subsurface energy (such as oil and gas or geothermal development) and environmental applications (including geologic carbon dioxide storage or deep well water disposal).

Chapter 6 summarizes the main conclusions derived from this dissertation work and offers suggestions for future or follow-on work.

Supporting information to this dissertation is provided in Appendices A, B, and C. Appendix A provides a poster presentation that aggregates a subset of results generated from this dissertation. Appendix B and Appendix C provide supporting information for Chapters 2 and 3 respectively.

2.0 Graining Perspective on Unconventional Well Design Choices through Play-level Application of Machine Learning Modeling

The following chapter is based on a peer-reviewed journal article published in *Upstream Oil and Gas Technology*, which can be cited as:

Vikara, D., Remson, D., and Khanna, V. Gaining Perspective on Unconventional Well Design Choices through Play-level Application of Machine Learning Modeling. *Upstream Oil and Gas Technology*. 2020. Volume 4, https://doi.org/10.1016/j.upstre.2020.100007

2.1 Chapter Summary

The recent development of unconventional oil and gas (O&G) reservoirs has led to an abundant hydrocarbon supply, both domestically and globally. However, there is a continued push to develop new and innovative approaches to improve exploration and extraction efficiencies and overall well productivity moving forward. Substantial improvements in unconventional O&G development are expected through optimized well completion and stimulation strategies aimed at maximizing well productivity. Optimizing well designs will require tailoring to the distinctive geologic conditions present for any newly placed well. To better evaluate the impact of well design attributes and their associated interactions on productivity in a major unconventional play, multivariate machine learning-based models that use empirical datasets were developed. A gradient boosted regression tree (GBRT) algorithm was applied. GBRT has been narrowly investigated for O&G applications but enables straightforward parametric importance and

influence evaluation, as well as assessment of parameter interaction effects. Models were trained on well design and locational parameters that serve as a proxy for variable geologic conditions to estimate two types of productivity indicator response variables strongly correlated to estimated ultimate recovery (EUR). The dataset utilized consists of over 7,000 well observations that cover the majority of the productive region of the Marcellus Shale. Model performance was evaluated and algorithm parameters tuned by analyzing the goodness-of-fit for simulated results against observed data in a cross-validation approach. Models were found capable of 73–79 percent prediction accuracy on held out testing data of gas equivalent production and can be used to inform future well design and placement decisions for increasing EUR per well and improving overall field-level recovery. Study results indicate that Marcellus well performance improves most with upscaling perforated interval lengths and water and proppant volumes per foot; but relative productivity improvements are spatially dependent across the play. Additionally, optimal combinations of water and proppant on well performance were found to vary depending on well location, emphasizing the utility of data-driven models capable of broad application across a play of interest for informing tailored well design approaches prior to their field deployment.

2.2 Introduction

Unconventional oil and gas (O&G) reservoirs were once economically unattainable resources [46]. However, the development and application of horizontal drilling with multi-stage and multi-cluster hydraulic fracturing techniques has enabled a surge in production from unconventional reservoirs, revolutionizing both the energy system of the United States (U.S.) as well as global energy markets [2, 47, 48]. The combination of horizontal drilling and hydraulic fracturing creates increased contact and flow pathways between reservoirs and horizontal wells, making production possible from these typically low permeability formations [49]. While the overall unconventional hydrocarbon resource in the United States is considerably large, the economically recoverable reserve portion is much smaller [50]. Recent downturns in O&G prices further diminish the economical portion of existing O&G assets, lending a sense of urgency to develop novel and innovative approaches aimed at improving exploration and extraction efficiencies and overall productivity.

Many have argued that the largest improvements in unconventional O&G development will come through optimization of well designs (which includes completion and stimulation strategies) aimed at maximizing well productivity and overall hydrocarbon recovery [12, 13, 14]. Since each O&G production zone is likely to be geologically distinctive from others for any given play, optimum well designs may require tailored approaches to the specific geologic conditions present. Therefore, identification of the most critical factors associated with both well designs and the geologic conditions for potential drilling areas, as well as their interactions, is essential for operators prospecting future well sites with the intent of maximizing gas productivity cost-effectively [15].

Many challenges still exist regarding well design optimization. For instance, further understanding is still needed related to the physical dynamics associated with fluid flow in highly complex fractured systems—a topic that is further complexed by potentially stark contrasts in geophysical conditions from one drilling location to another. This challenge is compounded in that acquisition of adequate levels of geological data at the well level is rare [51]. Reservoir modeling and simulation are the principal tools widely used to inform decision makers about reservoir response to potential hydraulic fracturing designs. However, even these types of models have noted challenges with overall predictive accuracy and can be both time and resource intensive to implement [47, 43, 52].

New innovations in the O&G arena are expected to occur through an increased focus on multi-discipline well design and placement by leveraging the use of digital empirical datasets and applying machine learning (ML) and data analytics to help improve well performance moving forward [38, 39, 40, 31]. The recent expansion in unconventional O&G development has also simultaneously sparked a substantial increase in the amount and types of data generated that would be available for such analyses [53]. Machine learning is a field of artificial intelligence that utilizes statistical algorithms to enable computer systems to progressively improve performance associated with a specific task from data, without relying on rules-based programming of the underlying causal relationships. These advanced techniques are particularly effective in environments where large amounts of data are available, and highly complex, variable conditions are prominent. Recent developments associated with sophisticated ML techniques and data management have expanded rapidly in many commercial sectors [54], providing an array of methods that can be targeted for use in O&G applications. These types of approaches could provide added utility in cases where sufficient understanding of the overarching complexities of unconventional reservoirs is lacking;

but the need exists to construct models capable of accurate and reliable replication of complex systems, or when expeditious predictive capability may be needed. The application of such approaches in unconventional O&G applications specifically could be aimed toward developing new insights, enabling improved understanding, and help toward optimizing well/reservoir interactions [38].

Several recent studies have demonstrated the use of ML and data analytics in subsurface energy and O&G applications. These studies have explored topics pertaining to predicting hydrocarbon production in unconventional reservoirs [55, 56, 15, 57, 58], lithofacies identification and characterization through data inversion [21, 19, 20], hydrocarbon production forecasting [59, 60, 61], and integration with advanced monitoring systems [62, 22]. Specifically, the studies that have developed data-driven approaches for predicting hydrocarbon production provide an initial foundation as well as implementation frameworks for using ML approaches to inform well design optimization. These types of studies involve the development of multivariate models using empirical data associated with design and completion-related well characteristics and use a productivity indicator (typically the cumulative production of hydrocarbons from wells over the first six or twelve months) as the response variable. The handling and incorporation of geologic parameters from these previous studies was varied, and included incorporation of explicit properties from geologic interpretation [56, 57, 47], assumed homogeneity across subsets of wells [63], utilized spatial coordinates as a proxy to evaluate variability in geologic conditions [64, 59, 15, 55], or ignored geologic conditions all together [65, 66].

The focus of this study involves the application of ML to a large well dataset that spans a prominent unconventional reservoir in the United States: the Marcellus Shale. The goal is to develop a predictive model capable of accurate estimation of a productivity indicator at the well

level that strongly correlates to estimated ultimate recovery (EUR) [67] using data commonly available. Both well design and spatial coordinate geologic proxy data parameters are utilized as predictors. This work is anticipated to supplement the advances from prior studies through several means, all focused towards model development intended to help inform future well designs to maximize productivity based on placement within the Marcellus Shale. The study includes the development and evaluation of a new productivity indicator that potentially better captures the production potential of a given well design/reservoir characteristics combination to enable models to make predictions with greater accuracy. As mentioned, one of the more overarching concerns in the modeling and simulation associated with any reservoir management strategy is in the reliability and accuracy of models utilized [43]. In unconventional reservoirs, this can be a greater challenge given the inherent complexity of the systems involved [64]. Therefore, the use of an improved productivity indicator should facilitate accurate improvements for data-driven modeling moving forward. A gradient boosted regression tree (GBRT) ML algorithm was implemented as part of model development. GBRT is in the boosting family of algorithms and believed to be an improved approach to other decision tree-based algorithms (like random forest) because of the way the algorithm sequentially addresses prediction shortcomings [68]. GBRT is also advantageous because it enables straightforward parametric importance and influence evaluation and assessment of parameter interaction effects. Boosting algorithms have recently become widely utilized in many data-science fields due to noted improvements realized in model accuracy. However, they have been narrowly applied in O&G applications. The work of LaFollette and coworkers is one example where boosted tree models were developed to rank the importance of predictor parameters in Middle Bakken Formation of the Willison Basin [55]. In another example,

Wang and Chen (2019) developed a predictive model using AdaBoost to evaluate oil production well performance in the Montney Formation in western Canada [15].

This study is an extension of the National Energy Technology Laboratory's existing research aimed towards gaining data-driven insights for better understanding and optimizing the well performance in the Marcellus Shale [56]. Nine algorithms of various complexities (including linear regression, neural networks, and support vector machines among others) were evaluated in the foundational study focusing on the western portion of the Marcellus Shale. The use of GBRT-based models as part of this study will facilitate improved comprehension of the hierarchy of both well design and geologic reservoir quality parameters and their associated interactions by enabling evaluation of their relative importance—valuable information that could empower operators to determine the best well placement and completion designs that potentially maximize the EUR per well and improve overall field-level recovery in the highly prominent Marcellus Shale play. The authors are not suggesting to replace detailed reservoir modeling with ML approaches explicitly, but instead consider them as an additional component to reservoir management strategies moving forward.

2.3 Data and Methods

As the focus of this study, GBRT-based ML models are developed for estimating natural gas equivalent hydrocarbon production from horizontal wells across the Marcellus Shale. These models use a combination of well design and geologic proxy data parameters as inputs. Literature has demonstrated that methodologies employed for the development of ML-based models in unconventional O&G applications are highly variable and often fit-for-purpose; yet, all possess

some commonality in model development centered on best practices. This is largely due to the unique circumstances influencing the availability of data within a given study region, as well as from the specific application of developed models. With that said, the general approach used for this study was inspired from the model development procedures recommended by Esmaili and Mohaghegh (2016) and Mohaghegh, Gaskari, and Maysami (2017), caveated slightly for circumstances unique to this study (described in the upcoming sections). The framework followed for this study is presented in Figure 7.

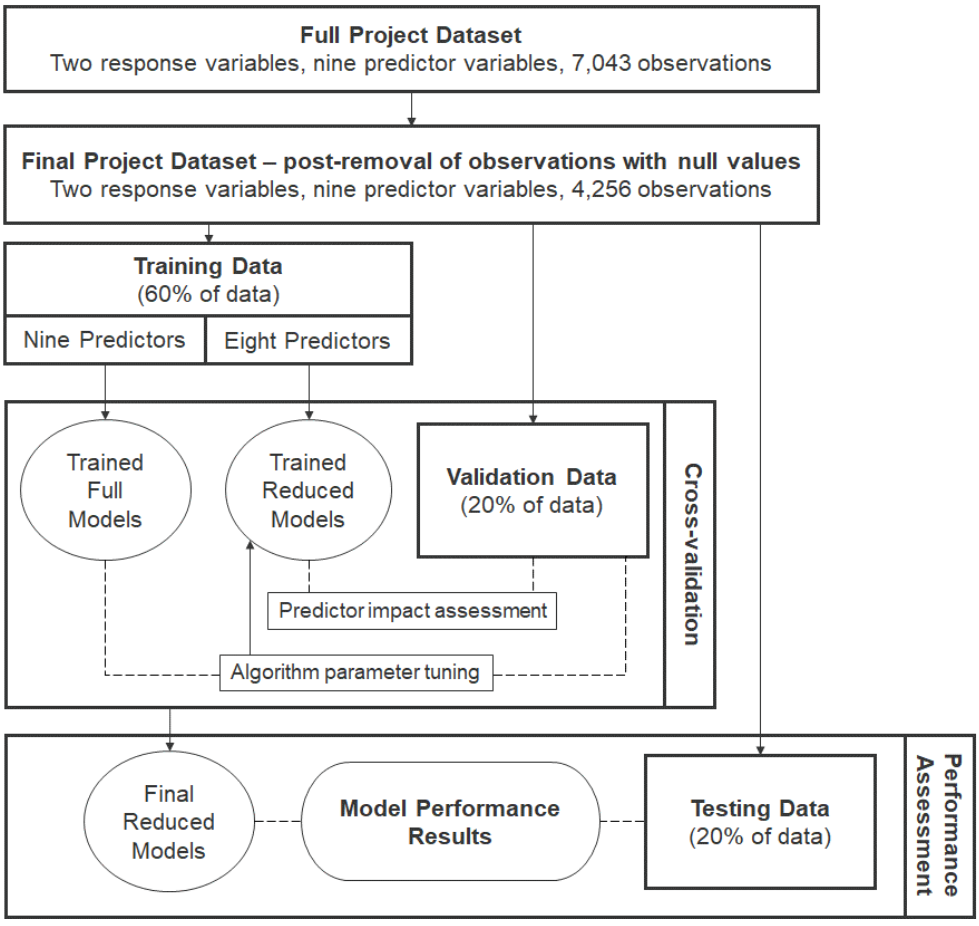

Figure 7. Framework for developing GBRT-based data-driven predictive models for Marcellus Shale wells.

2.3.1 Study Area and Data Sources

The study area focused on a large portion of the Marcellus Shale play in the Appalachian Basin of the United States. The study well data utilized was obtained from the O&G data vendor DrillingInfo [69]. Horizontal wells with first production dates between January 1, 2010 and December 31, 2018 were acquired and utilized—totaling 7,043 well observations. The full dataset underwent a removal of entries that contained one or more missing data values for predictor and/or response parameters of interest (Table 2). As a result, a total of nine predictor parameters and two response parameters were obtained for each of the resulting 4,256 wells in the study dataset.

Table 2. Data parameters available for each well evaluated.

Group	Variable Type	Parameter	Mean	Std. Dev.
Well Design	Predictors	Water per perforated foot (bbls)	32	19.1
		Proppant per perforated foot (lbs)	1,475	866
		Additive per perforated foot (bbls)	1.54	3.81
		Perforated interval length (ft)	5,501	2,088
		Well trajectory azimuth (degrees)*	325	29.3
		Acre spacing (acres)	150	126
		Nearest neighbor spacing (ft)	1,197	944
Geology Proxy		Surface hole latitude (decimal degrees)	40.643055	0.97
		Surface hole longitude (decimal degrees)	-78.721317	1.95
Productivity Indicators	Response	First 12-months production (MMcfge)	1,503	1,030
		Top 12-months production (MMcfge)	1,637	1,084

*Per similar approaches by Shih et al. (2018) and LaFollette et al. (2013), all well azimuth trajectory data was adjusted to fall between 180^{o} and 360^{o} to avoid a bi-modal distribution of well orientations.

These wells are plotted in Figure 8 and demarcate the portion of the Marcellus Shale evaluated based on well and data availability.[3] The resulting study area covers nearly 23,700 square miles. One way for ML and data analytics to make a more immediate impact in unconventional O&G exploration and production operations is though the development and validation of accurate models that both can learn from past and inform future well designs based on their specific placement in heterogenous reservoirs. This requires that models be trained on datasets that include parameters reflective of the relevant technological well design components that would be deployed in the near-term (as models trained on certain datasets may struggle estimating the impact of new parameters where extensive data does not exist). This project dataset is large enough to provide ample data coverage to capture both Marcellus Shale well design changes over time and wells placed across the play in both core and peripheral areas.

The Marcellus Shale is a Middle Devonian-age organic-rich formation that extends from New York State in the north to northeastern Kentucky and Tennessee in the south [70]. It is considered one the most prolific natural gas-producing formations in the world. The play has been a major shale gas producing resource since roughly 2008 and became the largest gas producing field in the United States in 2013 [71]. It is anticipated to continue to be a major gas producer into the future—projected to produce more than 20 billion cubic feet of gas per day through 2040 [72]. Similar to other continuous plays, the notable geologic and technical criteria that define the play boundaries and have been shown to influence hydrocarbon productivity include thermal maturity, total organic carbon content, formation thickness, porosity, permeability, depth, pressure, gas-in-

[3] The geographic information systems layer used for all maps to display state and county boundaries as part of this project is provided from the U.S. Department of Commerce [331].

place, the ability to be fractured (brittle vs. ductile), presence of existing natural fracture networks, in addition to lateral target selection and completion design. However, the lithology of the Marcellus Shale is known to be highly heterogeneous and vary significantly across the Appalachian basin [71, 73]; therefore, the geologic criteria mentioned are highly spatially dependent.

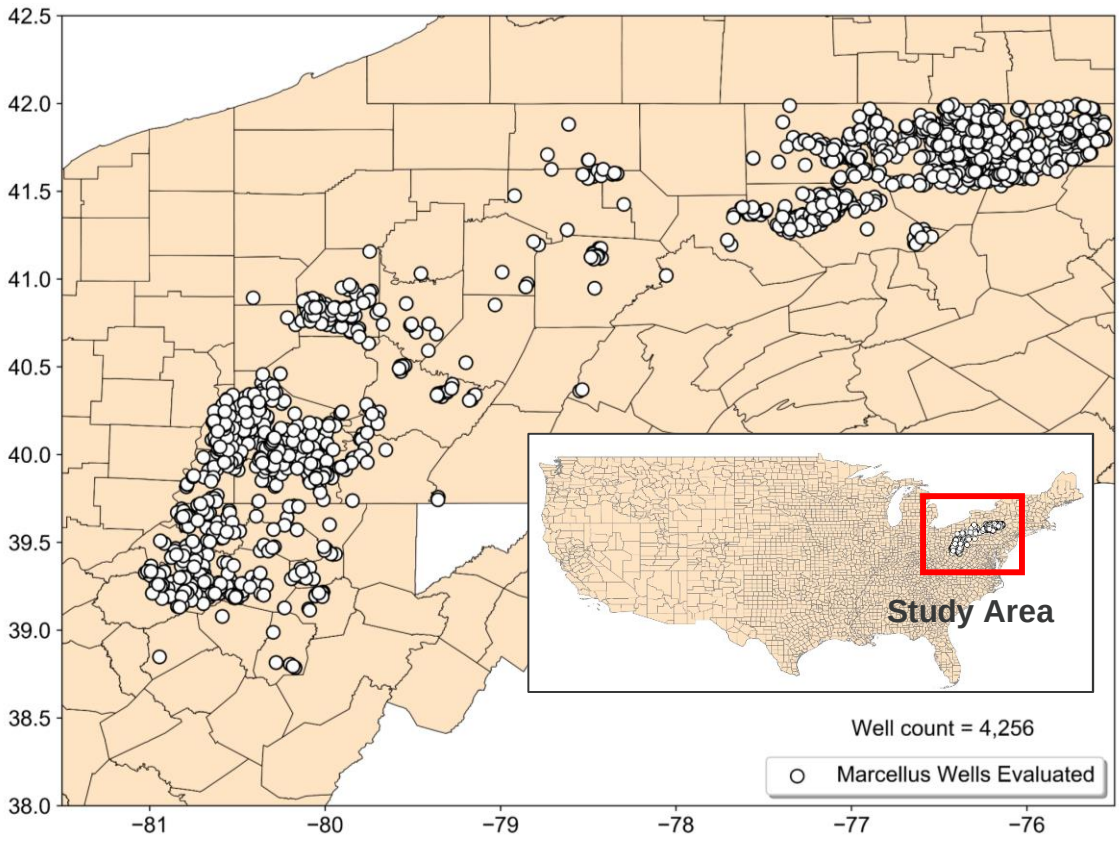

Figure 8. Map of the wells utilized as part of the study. All wells are horizontal wells within the Marcellus Shale with first production dates between the years of 2010 and 2017. A total of 4,256 wells were available for analysis that contain a complete set of data for each parameter of interest (Table 2).

The Marcellus Shale contains two major core areas that have enabled higher relative production capability compared to the rest of the play. These include one in southwestern

Pennsylvania and northern West Virginia (southwestern core) and the other in northeastern Pennsylvania (northeastern core). Both areas are captured in the study dataset. Each core area contains geologic characteristics that make it uniquely distinct. Notable contrasting formational characteristics between these two core areas relate to differences in depth, thickness, pressure gradients, organic content, and thermal maturities among others. The southwestern core is typically higher in total organic content, net (absent limestone intervals) to gross thickness, higher in porosity and permeability, contains a lower pressure gradient on average, and less thermally mature than the northeastern core. Given the thermal maturity conditions, portions of the southwestern core are rich in natural gas liquid content. However, the Marcellus Shale thickens from approximately 100 feet average gross thickness (including interbedded limestone intervals) near the southwestern core to greater than 300 feet average gross thickness towards the northeast core. Increases in the pressure gradient and thermal maturity also occur; the later condition resulting in predominantly dry gas conditions [74, 75, 70]. In addition to these core areas, there is a vast amount of peripheral area largely underexplored that could have promising production potential for future Marcellus Shale development [71].

2.3.2 Predictor and Response Parameters

Model development included incorporation of all relevant and available data across wells in the study region that relate to three broad categories: 1) Well design, 2) spatial coordinates that approximate variability in geologic conditions, and; 3) production response indicators for each well. The data included as predictor variables are associated with the length of the perforated interval contacting the reservoir, the volume of proppant, water, and additive used for hydraulic fracturing on a per foot of perforated interval basis, the azimuth orientation of the well lateral

trajectory, well locational data, and well spacing data to evaluate the potential impact of interference from offset wells. Two productivity indicator response variables were evaluated as part of this study; discussed later in this section. Table 2 lists the selected predictor and response parameters used in this study and basic statistical properties of each. An expanded statistical interpretation is available in Table 22 of Appendix B.

In low permeability unconventional reservoirs, the hydraulic fracturing process involves injecting large volumes of fluid at high pressures into production zones to break the rock down and initiate flow pathways from which hydrocarbons can travel through to the well. The fluid consists of mostly water and proppant, but also includes a small portion of chemical additives. Proppant (which may consist of sand or ceramic material) keeps fractures highly conductive and open long-term. Additives can serve a multitude of purposes aimed at ensuring the wells maintain efficient fluid and proppant delivery as part of the hydraulic fracturing process, as well as hydrocarbon recovery afterwards. Specific additive formulations may vary from well to well, but may include biocides, scale inhibitors, iron stabilizing agents, corrosion inhibitors, friction reducing agents, gelling agents, and cross-linking agents to name a few [48]. Another important factor related to hydraulic fracturing well design is the number and placement of fracturing stages along the lateral. Unfortunately, these data were not readily available, but it is likely correlated with lateral length [59, 76]. Instead, proppant, fluid, and additive were normalized for each well on a per foot of perforated interval of lateral basis.

Other important well design characteristics captured in the dataset relate to the wellbore lateral orientation and well spacing. Well lateral directional alignment (represented by well trajectory azimuth) is influenced strongly by the orientation of in situ stresses in the reservoir. Wells drilled along the minimum horizontal stress tend to generate transverse fractures via

horizontal fracturing, which are considered better suited for improving drainage areas and overall well productivity. When well laterals are oriented properly on azimuth, higher production rates are expected [48]. In terms of well spacing, both acre spacing and nearest neighbor data parameters were utilized. Each parameter can be used to infer the distance between wells evaluated and their anticipated drainage areas. Additionally, these data parameters may provide insight of any potential interference from hydraulic fracturing operations via nearby wells.

Similar to studies performed by LaFollette et al. (2013), Schuetter et al. (2015), Montgomery and O'Sullivan (2017), and Wang and Chen (2019), the variability in geologic conditions across the Marcellus Shale was assessed via proxy using each well's surface locational data (latitude and longitude). Since geological data is typically not available at large scales when evaluating thousands of wells, and interpolation of geologic properties over large areas can introduce uncertainty [77], well locational data provides an approximation approach to evaluate geologic conditions known to vary spatially across the study area. The aforementioned studies have found these input parameters critical for enabling subsequent models to accurately generalize well productivity.

Two response variable productivity indicators were considered as part of this study; both involve the cumulative summation of twelve-monthly empirically-based production values per well in gas equivalent units (million cubic feet of gas equivalent [MMcfge]).[4] The first, referred to as First 12-months of cumulative production, is a summation of each well's total production in its first 12 months of operations (either oil, gas, or an equivalent). This is a widely-used productivity

[4] This step involves combining gas and/or oil production values for each well into a single unit by considering one barrel (bbls) of oil to contain contains six times the British thermal units as one thousand cubic feet of gas.

indicator common to similar types of data-driven unconventional well predictive modeling studies (e.g., Wang & Chen, 2019; Montgomery & O'Sullivan, 2017; Schuetter, Mishra, Zhong, & LaFollette, 2015; LaFollette, Izadi, & Zhong, 2013 among others). In unconventional reservoirs, a well's production rate typically peaks within the first few months and then begins to decline over its productive life. Since a large portion of the hydrocarbons produced from these wells occur early on, this particular productivity indicator is a good proxy for overall well EUR [59]. However, limitations exist with using the First 12-months productivity indicator under many circumstances as it can fail to best represent the production potential of a well given its unique design characteristics and placement in the reservoir. For instance, well production profiles can vary from the typical "peak and decline" trend for a number of reasons, including being shut in or choked back due to less than favorable O&G prices, restrictions in pipeline offtake capacity, and equipment or maintenance issues. On the other hand, recompletion/refracturing efforts or installation of artificial lifting equipment can provide boosts in production. Given the potential impact from these circumstances, the First 12-months productivity indicator may underrepresent the true production potential of wells. Therefore, a second productivity indicator was developed and used as part of this study that may better represent a given well's productivity potential and help towards improving the accuracy and reliability of future data-driven modeling efforts. The second productivity indicator, referred to as the Top 12-months production, is a summation of the 12 highest monthly observed production values for a given well regardless of when they occur during a well's existing productive lifetime. The mathematical expression for First 12-months production is described in Equation 2-1 and for Top 12-months production in Equation 2-2:

$$First\ 12_w = \sum_{m=1}^{12} p_m$$ such that every production month p_m for each well w corresponds to months 1 through 12 Equation 2-1

$$Top\ 12_w = \sum_{1}^{12} p_m$$ such that every production month p_m for each well w corresponds to the highest recorded 12 values of p_m Equation 2-2

where p_m is the monthly production value for a given well (w) in MMcfge, and m is a given production month. Each productivity indicator can only be calculated for wells with at least 12 months of production history. Wells with less than 12 months of observed production data were therefore not included in this analysis. However, the potential benefits of the Top 12-months production is that it can better capture productivity potential for wells where production has been interrupted or well designs have been modified. Figure 9 highlights examples using empirical time-series production data from arbitrarily selected wells in the study dataset where the two productivity indicators are comparable (Figure 9 top) and where they differ (Figure 9 bottom). The reason for the discrepancies observed in the productivity indicators in the bottom portion of Figure 9 is unknown; however, it's clear that the variable time-series production profiles of those wells make it challenging to represent their productivity potential by the first 12 months of production alone. In this study's dataset, nearly 2,500 of the 7,043 wells available (35 percent) have a noticeable difference in the First 12-months production compared to the Top 12-months production of 50 MMcfge or greater; in roughly 100 cases, the discrepancy is greater than a 1,000 MMcfge difference. It should be noted that for relatively newer wells that may have experienced some form of interruption, the 12 potentially highest months of production may not have yet occurred. For these instances, the First 12-months may be similar or equal to the Top 12-months.

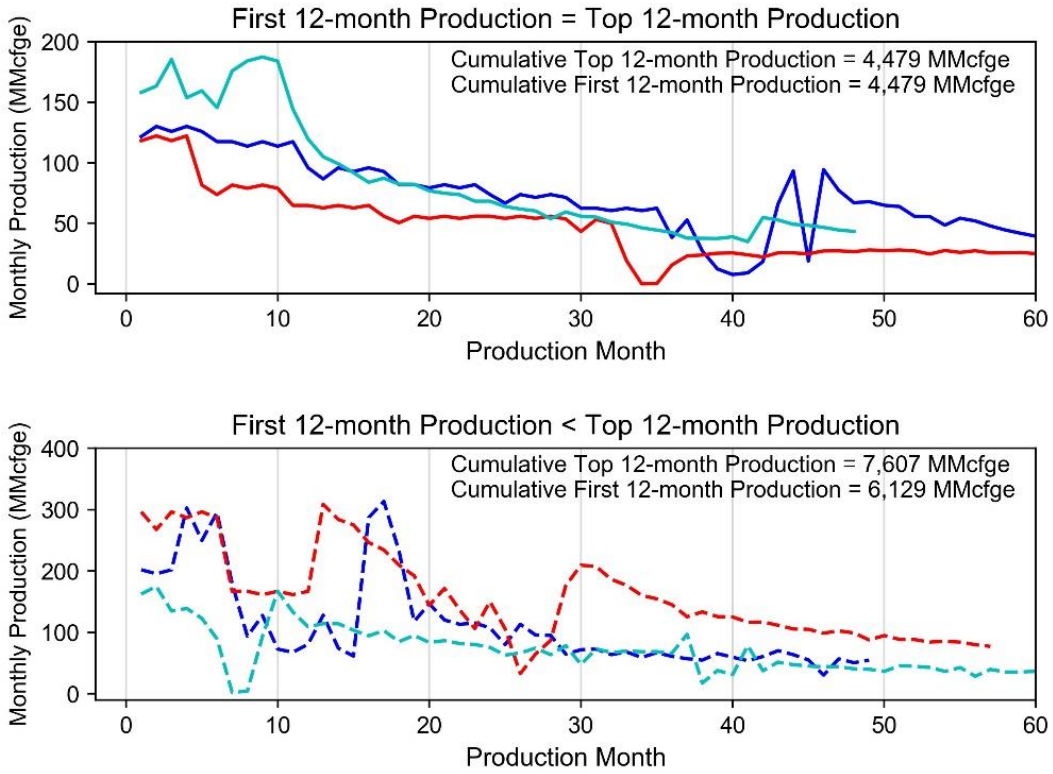

Figure 9. Examples of time-series production profiles for arbitrarily selected wells across the study area. The top chart features three wells (indicated by different colors) where the calculated values for each productivity indicator are the same. The bottom chart features three separate wells where the calculated values for each productivity indicator are different. The cumulated production in each chart is the summation of monthly production that corresponds to each productivity indicator for all wells featured for 12 total months (either First 12 or Top 12).

The Pearson correlation (Pearson r) was used to evaluate the linear relationship between each productivity indicator and EUR estimates calculated by DrillingInfo to verify their utility as a response parameter for assessing well productivity. In Figure 10, scatter plots are presented between each productivity indicator and an estimation of EUR. The plots suggest that both productivity indicators are correlated to EUR, but the Top 12-months productivity indicator is correlated slightly higher with EUR estimates than First 12-months for wells in this study dataset.

Additionally, the histogram in Figure 10 shows the distribution of well counts per associated EUR estimate for perspective. One of the objectives of this study is to develop data-driven models and evaluate and compare model performance for predicting either productivity indicators using the methods described in the upcoming sections.

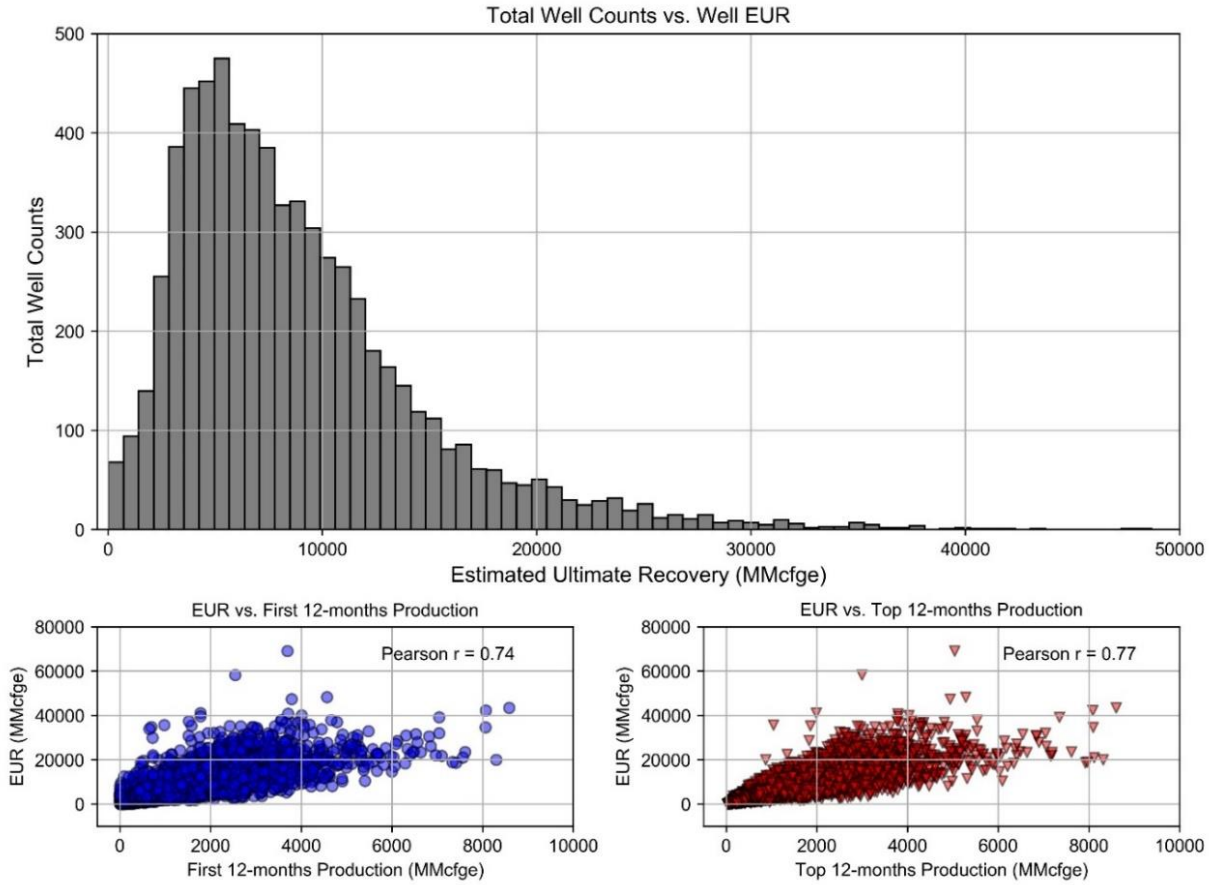

Figure 10. Breakdown of the EUR for the 4,256 wells evaluated in this study. The top chart features a histogram for the distribution of well counts per associated best estimate EUR as determined by DrillingInfo **[78]**. The bottom left chart shows the correlation (via Pearson r) of the widely-used First 12-months production performance indicator to well EUR. The bottom right chart features the correlation (via Pearson r) of the new Top 12-months production performance indicator to well EUR.

2.3.3 Overview of Gradient Boosting for Regression

This section includes a description of the gradient boosted regression (GBR) approach used to develop GBRT-based predictive models for productivity indicators for horizontal wells in the Marcellus Shale. It also provides information about the unique features of GBRT algorithm utilized, as well as important parameters and architectural components manipulated in this study to achieve accurate predictive models. Python version 3 and several packages within the scikit-learn library [79] were used extensively to perform analyses for this study.

Gradient boosting is an ensemble ML technique that can be used for classification and regression problems in which a final predictive model is developed consisting of an ensemble of weak prediction models, typically decision trees [80]. GBRT is inherent to the gradient boosting concept, and indeed uses the regression tree (or decision tree) groups of models [81]. Therefore, GBRT is considered an ensemble method (combination of several ML techniques into one predictive model [82]) that combines the strengths of two types of ML algorithms (boosting and decision trees) in order to improve the overall performance of a single, final model by fitting and combining many smaller models.

Decision trees are sequential models, which logically combine an arrangement of tests that compare a given numeric feature against a threshold value or compare a nominal feature against a set of possible values [83]. The goal of decision trees is to essentially create models that can predict the value of a given target variable or feature (that can contain either discrete or continuous values) based on several input variables [84]. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees [81]. A decision tree is grown through binary splitting of the source dataset into subsets based on an attribute value test, a process that is repeated on subsets of the data in a recursive manner [85]. The splitting points are determined

when prediction error is minimized. The recursive process continues until a stopping criterion is achieved, like a node subset has all the same value of the target variable, or when splitting no longer adds value to the predictions [86]. Decision trees have several noted advantages as they can handle multiple data types and data of widely differing scales, as well as predict complex functions [86, 81] and are widely popular because of their easy interpretability. However, Eilth et al. (2008) and Hastie et al. (2001) have noted several challenges common to decision trees that can affect their overall predictive performance. For instance, decision trees can be less accurate than generalized linear or additive models, and may have difficulty in modeling simple, smooth function types.

Boosting is an approach that builds an additive model in a sequential stage-wise fashion that enables numerical optimization by minimizing a differentiable loss function [80]. Different boosting algorithms vary in how they either quantify the lack of modeling fit from prior stage residuals or select settings for upcoming iterations [81]. In GBRT, each tree is constructed in sequence (i.e., boosting approach) as opposed to in parallel (i.e., bagging approach), where each new tree compensates for shortcomings associated with the previously developed tree [68, 64, 81]. The GBRT concept shares commonality with random forests in the sense that they are an ensemble of decision trees but are developed in sequence as opposed to in parallel. Overall model performance is essentially "boosted" by the addition of new trees fit to the residual errors of the previous model. The resulting final model is a consolidation of the many trees within the ensemble into a linear combination where each tree is essentially considered a term solving for a response variable (y). Gradient boosting through GBRT builds a predictive model in following manner as introduced by Friedman (1999):

$$F(x) = \sum_{k=0}^{K} \beta_k h(x; a_k)$$ Equation 2-3

where $h(x; a_k)$ in Equation 2-3 are the basis functions which are referred as "weak learners" in the context of boosting. GBRT uses decision trees {*0, k ... K* number of trees} of fixed size as the weak learners [79]. The weak learners are a function of input variables (x) with parameters $a = \{a_1, a_2, \ldots\}$. Expansion coefficients (β) and parameters (a) can be fit to training data in a sequential, stage-wise manner. First, an initial weak learner model is determined (i.e., $F_0(x)$), and then others for the specified number of trees $k = \{1, 2, \ldots K\}$ are added in a greedy fashion as follows in Equation 2-4:

$$F_k(x) = F_{k-1}(x) + \beta_k h(x; a_k)$$ Equation 2-4

and

$$(\beta_k, a_k) = \arg\min_{\beta, a} \sum_{i=0}^{K} L(y_i, F_{k-1}(x_i) + \beta h(x_i; a))$$ Equation 2-5

where the newly added tree $h(x_i;a)$ in Equation 2-5 attempts to minimize the loss function L, given the previous ensemble of the model $F_{k-1}(x)$. Gradient boosting attempts to solve this minimization problem for arbitrary loss functions L numerically via steepest descent. The steepest descent direction is the negative gradient of the loss function evaluated at the current model ($F_{k-1}(x)$) which can be calculated for any loss function in each region of a tree:

$$F_k(x) = F_{k-1}(x) + v\gamma_{lk}1(x \in R_{lk})$$ Equation 2-6

where each step length γ_{lm} is chosen using:

$$\gamma_{lk} = \arg\min_{\gamma} \sum_{x_i \in R_{lk}} L(y_i, F_{k-1}(x_i) + \gamma)$$ Equation 2-7

For each iteration *k*, regression trees partition the x-feature space [87] into disjointed, rectangular regions and predict separate constants for each region. The term R_{lk} in Equation 2-6 represents these regions for each corresponding terminal node *l* of the k^{th} tree. The shrinkage parameter (v; where $0 \leq v \leq 1$) in Equation 2-6 determines the learning rate of the procedure, which essentially controls the contribution of each tree within the final model. Smaller learning rate values ($v \leq 0.1$) have shown to result in improved model performance [81, 68]. However, decreasing the learning rate results in an increase in the number of decision trees required, and a greater number of decision trees may require longer computational times to fit the final model.

Performance of the GBRT-developed models depends on the choice of the loss function, learning rate, and other selected parameters. In this study, the least absolute deviation loss function was ultimately used prior to the parameter tuning and optimization step (discussed in Section 2.3.5), along with modifications to the parameter learning rate, boosting stages, the minimum number of sample splits, and the maximum depth of individual regression estimators (i.e., number of nodes in a tree). Overall, the use of the GBRT algorithm is an excellent choice for development of shale gas productivity indicator prediction models because of 1) being able to handle the complexity associated predicting the performance of a shale gas well, and 2) GBRT enables the

straightforward determination of predictor variable relative importance and partial dependence. The latter point provides for relatively expeditious insight into the major contributing production drivers considered and included as part of the model development. Identification of the hierarchy of both well design and geologic reservoir quality parameters and their associated interactions is paramount for determining the best well placement and completion designs that potentially maximize well EURs and help facilitate improved field-level recovery.

2.3.4 Evaluating Results and Model Performance

Model performance was evaluated by analyzing goodness-of-fit for simulated results against held out data (i.e., data not used for training) in several cases. While many performance evaluation approaches are available, the predicted output from developed models versus known observations were evaluated using R^2 and root mean square error (RMSE). The R^2 metric is considered reliable in similar O&G modeling applications [56, 64, 57, 58], and additionally, it is fairly easy to interpret. RMSE provides a complimentary performance metric, one that is directly interpretable in the units of the response variable. Evaluating model performance can provide insights toward each one's susceptibility for generating error in prediction.

The R^2 metric indicates the degree of correlation between simulated and observed values. By definition, R^2 is the regression sum of squares ($SS_{Regression}$) divided by the total sum of squares (SS_{Total}). R^2 values are proportional to the given data evaluated where higher values represent smaller variations between the observed data and predicted values. R^2 values range from zero to one; a value of one would indicate a perfectly-fitted relationship, whereas zero would suggest no correlation exists. R^2 is mathematically represented in Equation 2-8 as follows:

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1}(y_i - \bar{y})^2} \quad \text{Equation 2-8}$$

Where N is the length of the dataset, y_i is the observed value, and $\hat{y}_i$ is the simulated or predicted value. The overbar above variables indicates the mean over the entire portion of the dataset evaluated.

The RMSE metric represents the mean error between simulated values and observed values (i.e., residuals) and represents the variance of errors independent of sample size. Smaller RMSE values are associated with less mean error between simulated and observed results compared to higher RMSE values [87]. RMSE is one of the most commonly used metrics to assess model efficiency and is mathematically represented in Equation 2-9:

$$RMSE = \sqrt{N^{-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad \text{Equation 2-9}$$

For this study, R^2 and RMSE were used to evaluate model performance during several steps: 1) Algorithm parameter optimization and tuning (Section 2.3.5); 2) analyzing model performance with reduced predictor inputs (Section 2.3.6); and 3) evaluating the best performing final model formulation developed for predicting each production performance indicator type (Section 2.4.2).

2.3.5 Model Development Using Cross-Validation

A model cross-validation approach was used to train, validate (often called “calibrate”), and test the developed models as part of this study. The cross-validation approach is a method of estimating the accuracy of classification or regression models in which the input dataset is divided into several distinct segments (at the discretion of the user). Each segment in turn is used to test a model trained to the remaining parts [88, 47]. The goal here is to develop models that can effectively generalize unconventional well performance while avoiding overfitting to the training data. The well data used for this study were broken down into training, validation, and testing datasets at random in a 60/20/20 percentage-based split following the removal of entries that contained one or more missing data values for input parameters of interest (Table 2). The resulting well count breakdown per dataset was 2,553 wells in the training dataset, 852 wells in the validation dataset, and 851 wells in the test dataset. The training dataset provides the GBRT ML algorithm sets of examples used for learning so that algorithm parameters (like expansion coefficients) can be fit. The validation dataset provides a unique set of hold out well completion and production response data examples not used as part of model training that are used to tune and optimize GBRT regularization parameters of the trained model. In this study, the learning rate, number of boosting stages, the minimum number of samples required to split an internal decision tree node, and maximum depth of the individual regression estimators (which limits the number of nodes in a tree) were tuned and optimized using a one-at-a-time grid evaluation consisting of

multiple possible regularization parameter combinations [89] against the validation dataset.[5] For each possible combination of model parameters adjusted as part of tuning and optimization, a new model was fit to the training dataset under the given model regularization parameter combination. Each model was then used to predict either the Top 12-months or First 12-months productivity indicators using the validation dataset. R^2 and RMSE were used to evaluate the performance of each model's prediction. The parameter combination for the best performing models for each productivity indicator type was selected for further analysis based on the highest R^2 score and lowest RMSE combination. Ultimately, the best configuration of GBRT parameters that generalizes well productivity, enables highest performance accuracy, and avoids overfitting are presented in Table 3.

Table 3. Final set of GBRT regularization parameters following cross-validation

Productivity Indicator	Boosting Stages	Min. Sample Splits	Maximum Depth	Learning Rate	Loss Function
First 12-months	3,000	15	14	0.01	least absolute deviation
Top 12-months	3,000	10	14	0.005	least absolute deviation

The testing dataset provides a final collection of unique holdout data examples used to assess the prediction performance of the fully developed model formulations for each productivity indicator that utilize the best performing predictor regularization parameter value combinations.

[5] Only the least absolute deviation (lad) loss function was utilized as part of this study. For the one at a time parameter optimization step, no other type of loss function was used.

2.3.6 Refining the Predictor Parameter Dataset

A predictor parameter evaluation and selection approach was employed to establish a refined predictor parameter dataset that contributes most to the prediction response for models estimating each productivity indicator. For decision tree-based models, the ideal predictor parameter combination would maximize the mean value of R^2 and reduce variance in prediction.

In this study, a parameter removal approach was implemented to evaluate the impact of each parameter on model performance to predict either Top 12-months production or First 12-months production. Additionally, such an analysis enables detection of potentially diminishing returns from adding certain predictor parameters into the model formulation. In general, the training and validation datasets were modified accordingly to reflect the omission of a given parameter attribute. Models were then re-trained 10 different times on the modified predictor dataset and the overall prediction accuracy is determined against the validation dataset. The process is repeated by swapping the previously omitted parameter back into the formulation for another. The GBRT regularization parameter configurations presented in Table 3 were used for this step. Influential parameters removed from the GBRT models would likely reduce model predictive accuracy relative to the impact of that parameter. On the other hand, removal of less influential predictors may only marginally reduce, even possibly improve, model performance through the parameter omission.

The noted change in R^2 and RMSE for the new model formulation in its prediction accuracy against the validation dataset are used to as the overall performance metric for this step. Parameters considered unimportant would be dropped altogether to form new model formulations. In a way, the approach shares similarities with recursive feature elimination, one parameter at a time sensitivity analysis [31], and the R^2-loss evaluation approaches described by Schuetter et al. (2015)

and Mishra and Lin (2017). This type of simplification approach is considered useful under various circumstances, including where potential redundancy may exist between two or more predictor variables, or where developer preferences favor a certain combination of predictor variables [81]. The goal of this step is to establish model formulations that provide for the most accuracy and reliability, and that utilize a reduced but indispensable set of predictor parameters.

2.3.7 Assessing the Effects of Parameters on Production

To evaluate the effects of predictors on the production response for the GBRT-based models developed as part of this study, two approaches were used: 1) relative importance of predictor variables and 2) partial dependence analysis. These approaches can be evaluated collectively to understand the hierarchy of both well design and geologic reservoir quality variables on model response, as well as the unique functional relationships between predictor variables and response.

Predictor parameter relative importance in GBRT-based models is determined from the number of times a given predictor variable is selected for splitting, weighted by the squared improvement to the overall model resulting from each split, then averaged over all trees [64, 90, 68]. Essentially, the more a given parameter is leveraged to influence decisions within decision trees, the higher the importance. Since a relative importance value is calculated for each predictor variable, they can be ranked and compared relative to each other in a hierarchical fashion. For this study, values for importance are normalized relative to the most important predictor variable then scaled by 100 (the most important variable has a value of 100, and all others are less than 100).

Partial dependence analysis enables evaluation of the marginal effect of one or two model predictor variables on the response variable, conducted by averaging the effects of all other

predictor variables and increasing values for the parameter(s) of interest from low to high over several iterations [68, 85, 91]. The mathematical foundation behind partial dependence functions is described by Friedman and Popescu (2005) [92]. Partial dependence can be plotted to visually illustrate the overall complexity in the functional relationships between the response variable and a given predictor variable or variables. Therefore, it is an effective approach for identifying overall causal relationships between predictors and the associated model response [93]. Additionally, it provides for a more robust technique in understanding each predictor parameter effect by evaluating them over a continuous range of values opposed to relying solely on a single numerical score that represents each parameter's impact in its entirety.

2.4 Results and Discussion

The findings from this study, presented in the subsections below, confirm that accurate prediction of Marcellus Shale well productivity indicators can be achieved by using GBRT and data that combine well design and geologic proxy data.

2.4.1 Parameter Selection for GBRT-Based Models

Table 4 depicts the results from the one or two at a time parameter removal evaluation based on the approach discussed in Section 2.3.6. Each row in Table 4 represent aggregated model R^2 and RMSE scores for 12 different cross-validation iterations where GBRT models were re-trained 10 times each for the one or two parameters removed. Results are for prediction against the validation dataset. Predictor parameters are ordered (from top to bottom per Top 12-month

results) based on mean R^2 scores for reduced models associated with their omission. The parameters listed toward the bottom of the table are more impactful for model accuracy than those toward the top in the sense that performance drops more extensively when they are excluded. The Full Model formulations where all nine predictors from Table 2 are included as part of model development and can be used as a baseline reference for model performance changes given the omission of predictors for new, reduced model formulations. Removal of any given parameter for new models trained results in an adjustment to performance relative to the Full Model formulations, either as an improvement or reduction.

In Table 4, its noted that for models predicting each productivity indicator, well azimuth and water per foot both improve performance when omitted relative to the Full Model formulation with all parameters included. This can be attributed to these parameters having 1) a negligible effect on the model response (evaluated in Sections 2.4.4 and 2.4.5) or 2) strong collinearity with another predictor variable (discussed in the following paragraph). The most impactful parameters with the largest relative drops in R^2 and gains in RMSE for predicting either productivity indicator are related to spatial coordinates and perforated interval length.

For two instances, a two at a time parameter removal evaluation was also conducted for predictor variables known to be highly collinear to evaluate the overall impact on model performance by removing both. For instance, both the Shih et al. (2018) and Schuetter et al. (2015) studies identified strong collinearity between proppant and water volumes used in unconventional O&G wells. Therefore, both 1) proppant and water per foot (major components in hydrofracturing well design) as well as 2) surface latitude and longitude (used to evaluate spatial variability in geologic conditions) were removed as pairs to evaluate the corresponding change in model performance vs. omission of any given parameter individually. Findings indicate that the overall

impact is substantially larger for each case (proppant/water or latitude/longitude) compared to removing any single parameter on their own. For instance, when water per foot alone is omitted from the model formulation, the overall R^2 value is shown to increase for Top 12-month and marginally reduces for First 12-months, suggesting the predictor parameter holds little importance. Intuitively, this seems to contradict unconventional O&G well design and stimulation best practices given the acknowledged significance of injecting water under high pressures to facilitate productivity. Removal of proppant per foot, on the other hand, reduces R^2 relative to the Full Model formulation for predicting both productivity indicators, suggesting it is an impactful parameter to some degree. Given the known correlation of water and proppant per foot in unconventional O&G wells (Pearson r of 0.82 for wells in this study dataset), removing one predictor can tend to compensate for the loss by removing the other and possibly imply that neither parameter is of substantial importance. However, when both water per foot and proppant per foot are omitted, their combined impact can be assessed from the resulting R^2 and RMSE change. Overall, the most impactful parameters are the spatial coordinates used to evaluate variable geologic conditions across the study area. This is true when either latitude or longitude are omitted individually or together. With both omitted, the most substantial loss in performance accuracy from the Full Model formulation is observed, emphasizing the importance of spatially dependent geologic reservoir variability when determining future well sites.

Table 4. Results of the one or two at a time parameter removal procedure and associated values of the R^2 and RMSE adjustments against the validation dataset.

Predictor(s) Omitted	Top 12-months Production		First 12-months Production	
	Mean R^2 (stdev)	Mean RMSE (stdev)	Mean R^2 (stdev)	Mean RMSE (stdev)
Azimuth (degrees)	0.792 (0.002)	491 (2.93)	0.738 (0.005)	519 (5.22)
Nearest Neighbor (ft)	0.780 (0.003)	504 (3.35)	0.726 (0.005)	531 (4.87)
Additive per foot (bbls)	0.780 (0.003)	490 (3.22)	0.745 (0.004)	493 (4.31)
Acre Spacing (acres)	0.776 (0.002)	528 (2.84)	0.721 (0.005)	556 (5.08)
Water per foot (bbls)	0.772 (0.003)	514 (3.46)	0.725 (0.003)	532 (3.35)
Full Model (no predictors omitted)	0.771 (0.004)	515 (4.41)	0.726 (0.004)	531 (4.24)
Proppant per foot (lbs)	0.771 (0.004)	516 (4.07)	0.713 (0.004)	539 (3.77)
Proppant (lbs) and Water (bbls) per foot	0.753 (0.004)	535 (4.60)	0.691 (0.005)	560 (4.10)
Perforated Interval Length (ft)	0.726 (0.004)	564 (3.66)	0.656 (0.005)	595 (4.62)
Longitude (degrees)	0.684 (0.004)	605 (3.90)	0.623 (0.006)	622 (4.84)
Latitude (degrees)	0.671 (0.004)	617 (4.07)	0.630 (0.006)	617 (4.86)
Latitude and Longitude (degrees)	0.424 (0.004)	817 (2.73)	0.390 (0.005)	791 (3.40)

Given its placement in the order of parameters in Table 4 and both redundancy and correlation with the acre spacing parameter (Pearson r of 0.86 for wells in this study dataset), the nearest neighbor parameter is omitted as part of the final model formulation for both productivity indicators moving forward. Since removing acre spacing results in a smaller improvement in R^2 and RMSE as indicated in Table 4, nearest neighbor is omitted instead. The final model formulation is therefore the best suited reduced predictor parameter combination that affords for the most accuracy and eliminates redundancy between parameters.

2.4.2 Regression Model Performance

Table 5 compares the prediction performance for the different model formulations predicting either the Top 12-months or First 12-months productivity indicators. The results indicate that developed models are relatively high performing for data-driven O&G predictive models based on literature review. Prediction results are compared again the training, validation, and testing datasets where each formulation was re-trained on the training dataset under 10 separate

iterations prior prediction. The mean R^2, RMSE, and their standard deviation values were used to evaluate performance. Results presented in Table 5 under the validation column confirm that the omission of the nearest neighbor parameter from the full model to final model formulations provided a slight boost in model performance and smaller standard deviation for predicting Top 12-months production, but results in a subtle reduction in model performance and larger standard deviation for the model predicting First 12-months. Based on the parameter impact assessment highlighted in Table 4 above, there appears to be no obvious formulation via predictor parameter removal for a First 12-months production model that can result in a more accurate model compared to predicting Top 12-months (i.e., exceeding a mean $R^2 = 0.780$ against validation dataset, or a mean $R^2 = 0.793$ against the testing dataset).

Table 5. Comparison of the mean and standard deviation (in parenthesis) predictive performance of developed models under different formulations

Productivity Indicator		Training		Validation		Testing	
	Predictor Configurations	Mean R^2	Mean RMSE	Mean R^2	Mean RMSE	Mean R^2	Mean RMSE
Top 12-months Production	Full Model - All Parameters	0.987 (0.002)	126 (9.887)	0.771 (0.004)	515 (4.143)	0.788 (0.006)	499 (7.032)
	Final Model - Nearest Neighbor Dropped	0.987 (0.002)	124 (10.600)	0.780 (0.003)	504 (3.349)	0.793 (0.002)	493 (2.979)
First 12-months Production	Full Model - All Parameters	0.990 (0.001)	100 (6.137)	0.726 (0.004)	531 (4.236)	0.725 (0.004)	536 (4.029)
	Final Model - Nearest Neighbor Dropped	0.991 (0.001)	100 (4.242)	0726 (0.005)	531 (4.871)	0.732 (0.003)	528 (2.872)

The prediction performance for the final model formulations is visually compared with observed data from the testing dataset in Figure 11. The cross plots provide a visual depiction of each model's prediction to actual observed production, which is quantified using R^2. Models that

provide a perfect fit would have an R^2 of one, and all data would fall along the black dotted lines (i.e., 1-to-1 match). Both models predicting either productivity indicator is fairly accurate with regards to observed production values; however, the final model formulation for predicting Top 12-months production holds a performance edge over the model predicting First 12-months. The evaluation of testing dataset observed values to 90% prediction intervals indicate that the majority of observations fall within the prediction interval ranges (67 percent for Top 12-months production and 65 percent for First 12-months production). The majority of the observations that fall outside of the prediction interval ranges occur at either extreme end of the sorted data set (i.e., $\leq$ 800 MMcfge and $\geq$ 3,000 MMcfge) for both productivity indicators.

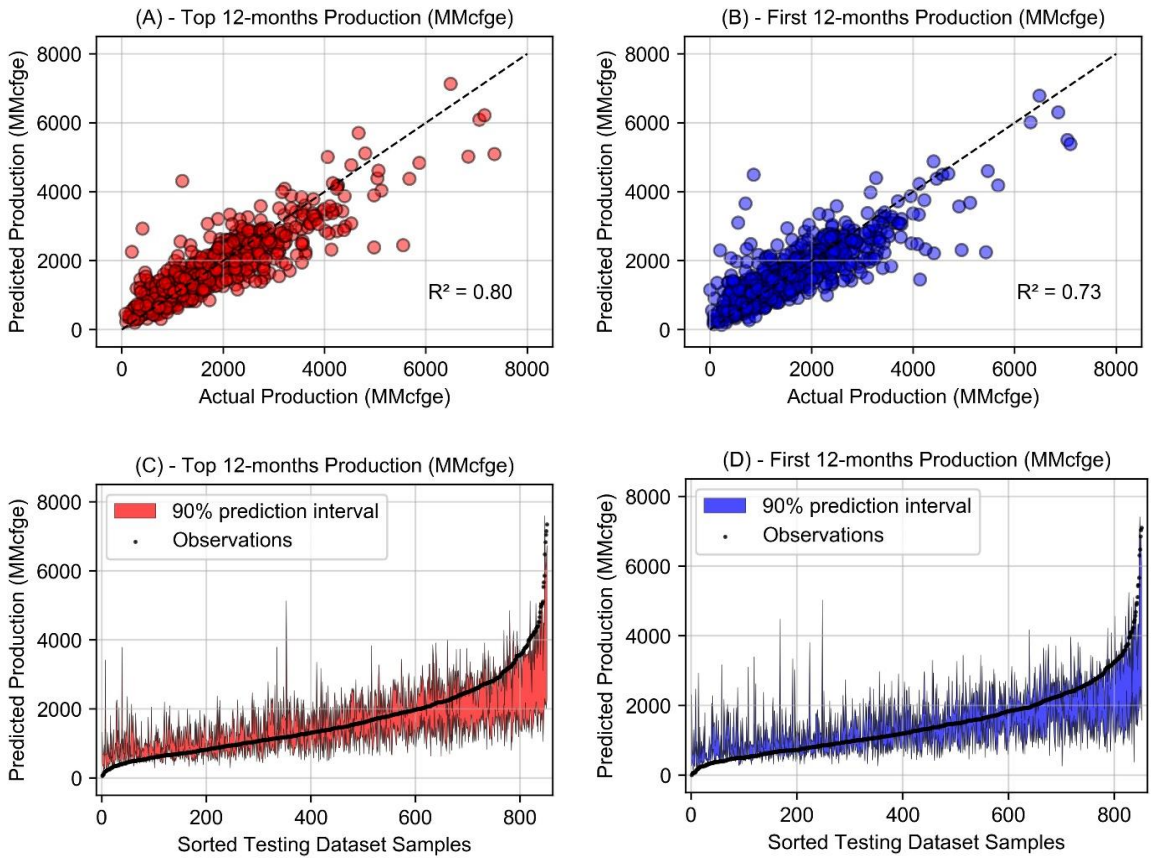

Figure 11. Assessment of model performance for predicting well production. The top plots depict actual (i.e., observed) production values plotted against predicted values for the Top 12-months production responses (A) and First 12-months production responses (B) for each well in the testing dataset using a single run under the final model formulations. The bottom plots show sorted testing data observations and 90% prediction intervals for Top 12-months production (C) and First 12-months production (D).

An analysis of model residuals between final model formulations predicting either productivity indicator against the testing dataset is provided in Figure 53 in Appendix B.

2.4.3 Evaluating Parameter Importance

Parameter relative importance plots are presented in Figure 12. The predictor parameter relative importance is compared for both the Top 12-month and First 12-month response variables under the final model formulation. Variables are ordered from high to low based on their resulting relative importance. Examination indicates two notable findings: 1) There are noted commonalities in four of the top five ranked predictor parameters for both productivity indicators but differences in their relative importance, and 2) the relative importance magnitude of predictor parameters differs to some degree depending on the predicted productivity indicator.

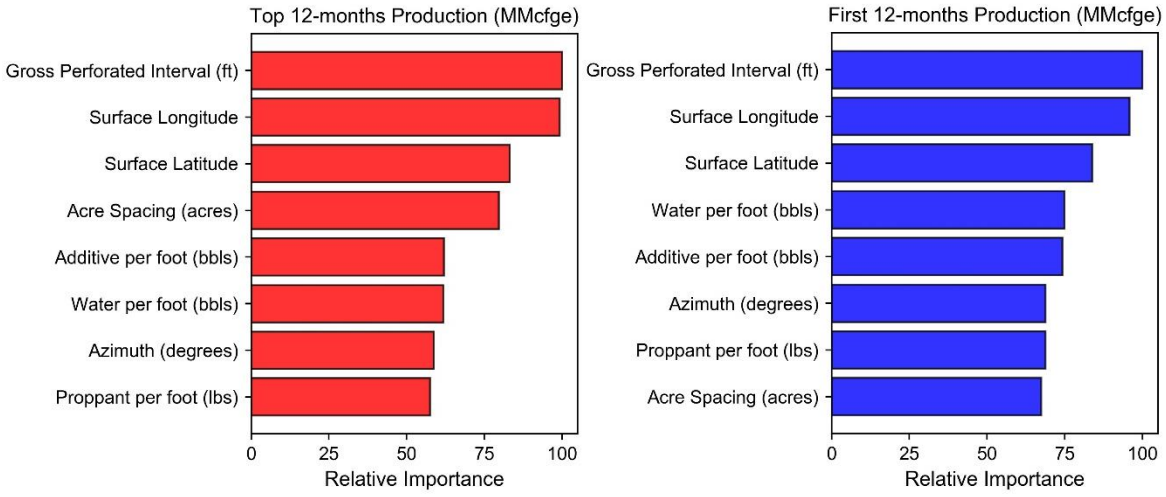

Figure 12. Summary of the relative importance of the predictor variables for the final model formulations for the Top 12-months response (left) and First 12-months response (right).

Gross perforated interval length, the geologic proxy parameters of well surface longitude and latitude, and additive were commonly ranked in the top five for predicting either productivity indicator, with gross perforated interval length as most important in both cases. Similarly, most of these same parameters were found to be important as having substantial detrimental impacts on

model performance when removed from training data (Table 4). In general, these parameters are commonly considered by many to be highly influential in unconventional O&G well design and placement decisions. Therefore, their positioning in the overall rankings in Figure 12 is not surprising. Furthermore, the results suggest consistency with others studies that have also modeled and evaluated/ranked well design and geologic parameters on productivity performance in other unconventional plays [55, 58, 15, 64], thereby helping to validate that models developed as part of this study are effectively identifying parameters commonly believed to be critical in producing from unconventional reservoirs. The remaining parameters (azimuth, water per foot, proppant per foot, additive per foot, and acre spacing) are ranked lower, but they would still be considered important toward estimating production given their overall relative importance magnitude (>50 for Top 12-months; >65 for First 12-months) in relation to the highest-ranking parameter. However, the disparity between the highest versus lowest importance parameters is much larger for the model estimating Top 12-months production compared to the model estimating First 12-month production.

For comparison prior to dropping the nearest neighbor parameter, Figure 54 in Appendix B provides predictor parameter relative importance for both the Top 12-month and First 12-month predictor parameters under the Full Model formulation.

2.4.4 Partial Dependence

Fitting data-driven-based models with large, multivariate datasets can result in the formation of complex parameter interactions that are often highly nonlinear. Therefore, challenges exist in attempting to gain straightforward understanding of input and output relationships, as well as key parameter sensitivities, based on an evaluation of model performance results alone [31].

This is where partial dependence can be handy in helping to identify causal relationships and assess the functional relationship between the response variable and a given input predictor.

The partial dependence for the two production prediction indicators for each of the eight predictor variables that constitute the final model formulation are presented in Figure 13. In this figure (and similarly in Figures 8 and 9), the vertical axes depict the partial dependence for each variable. The average effect for each variable occurs when set at the point(s) on the horizontal axis that results in a partial dependence equal to zero. Any variable setting that results in a negative partial dependence value correlates to a response less than the model average. For instance, a given variable at a value setting that correlates to a negative partial dependence value would generate a model response lower than the average MMcfge production response assuming all other variables are set at their average. Conversely, any variable setting that results in a positive partial dependence value indicates a model response greater than average. In general, the models developed to predict either productivity indicator are highly similar in terms of partial response. The findings from this analysis indicate that for the eight variables evaluated, Marcellus Shale well performance improves most substantially with increasing perforated interval length and proppant and water per foot volumes. These findings are not necessarily surprising as these parameters have been identified by many as highly influential well productivity design factors [59, 57, 14, 63, 58]. Interestingly, the partial response for proppant remains relatively flat until approximately 850 lbs/ foot; at which it then increases linearly with added proppant intensity. Additionally, well performance improves in regions where latitudes trend less than ~40.3 (toward southwestern Pennsylvania and northern West Virginia) and longitudes trend greater than -77 (towards northeast Pennsylvania). Partial response is influenced strongly by the availability of training observations. Figure 13 highlights how model response varies for each parameter based on the availability or lack of training data.

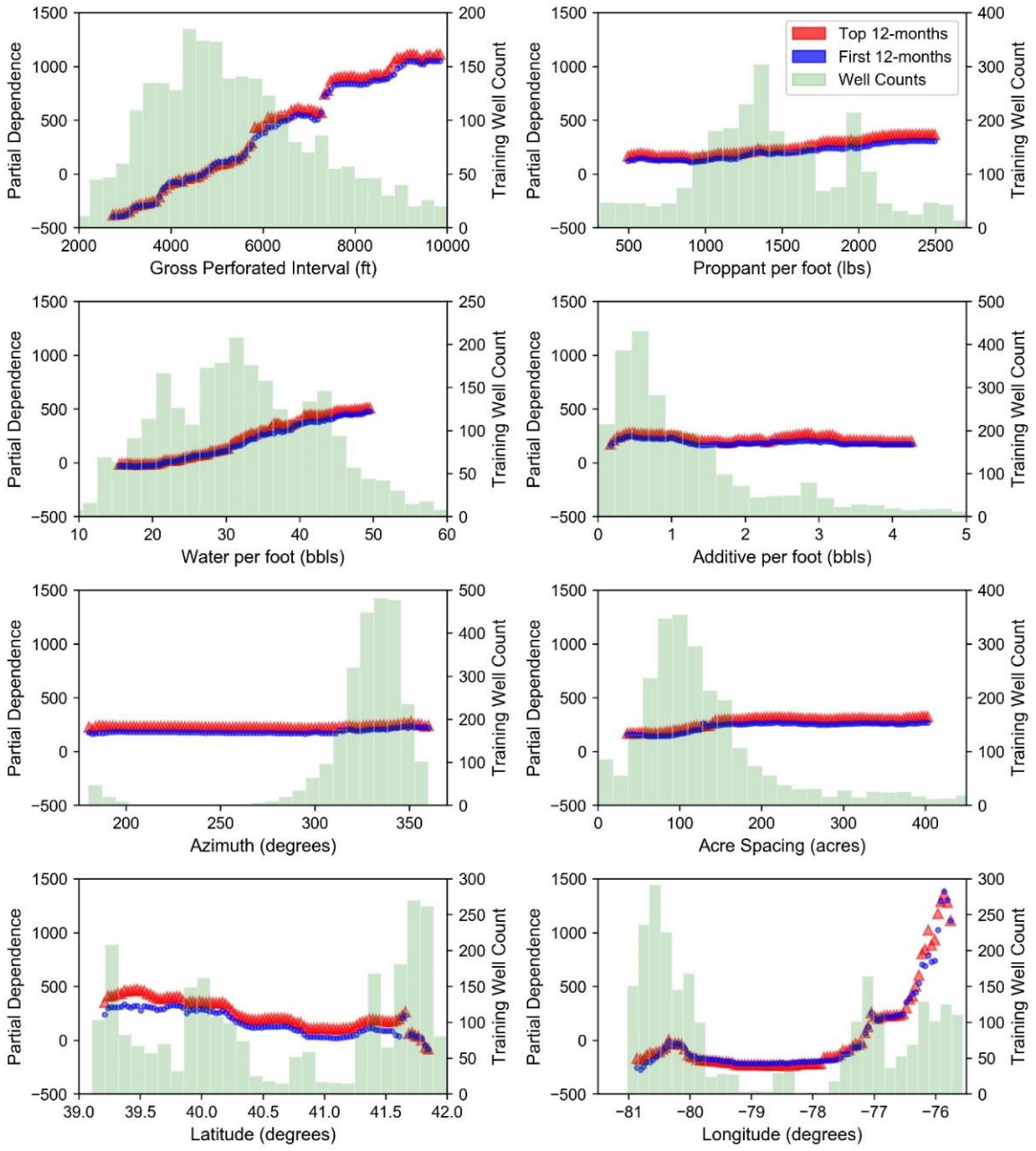

Figure 13. Partial dependence plots for the eight predictor variables as part of the final model formulations. The red lines pertain to the Top 12-months productivity indicator response, and the blue lines pertain to the First 12-months productivity indicator partial response. Histograms (green) emphasize the prominence of training data available at given values along the x-axes. Ranges on the x-axes evaluated over the scales between each parameter's 5th and 95th percentile based on observations in the dataset (azimuth ranges between 0 to 100th percentile).

Additive and acre spacing have noticeable, but more subtle effects than the previously described parameters. Acre spacing, in particular, is shown to have a positive effect on well production as the spacing value for wells increases. This supports the premise that more closely-spaced wells can be subjected to reduced production due to proximal well interference, or possibly from intentionally reduced proppant and water injection design regimes aimed at preventing well interference via smaller stimulated reservoir volume. Well azimuth trajectory seems to have limited effect on production. However, it is possible that operators have identified close to optimal wellbore trajectory for given parts of the Marcellus Shale play that would maximize hydraulic fractures and associated production. Therefore, if true for the wells within this study's dataset, contrast in partial dependence for various levels of azimuth would not be expected.

2.4.5 Evaluating Key Parameter Interactions

Evaluating single partial dependence evaluation (Figure 13) alone may imply that each parameter is independent if not evaluated in context, and that no correlations exist with other features. However, two parameters can be evaluated in concert using partial dependence to assess their interactive effects. The predictor parameters with more noticeable effects highlighted in Figure 13 were evaluated in select combinations to evaluate their interaction effects with joint partial dependence plots in Figure 14 and Figure 15 below. These plots were constructed where all parameters are marginalized by being held constant at their means except for the two evaluated for interaction. Only the Top 12-months productivity indicator is evaluated here, but Figure 55 and Figure 56 in Appendix B feature similar plots for First 12-months.

The interaction between well perforated interval length and latitude and longitude in Figure 14 demonstrates that improving production is not just dependent on longer laterals, but

also on their placement spatially within the Marcellus Shale. In this case, the interactions support the highest productivity typically occurs in areas common to the Marcellus Shale core regions with longer perforated interval lengths. Additionally, Figure 8 (B) highlights that shorter well laterals in core areas achieve common partial responses as longer laterals in peripheral areas. In general, there is less disparity in the partial dependence response under the perforated interval interaction with latitude compared to with longitude. It is suspected that the aggregated effect of moving north to south in the play is suppressed due to a blending of geologic reservoir qualities (i.e., both high and low reservoir qualities occur at any given latitude) comparted to moving west to east (where both core areas are more apparent in the plots).

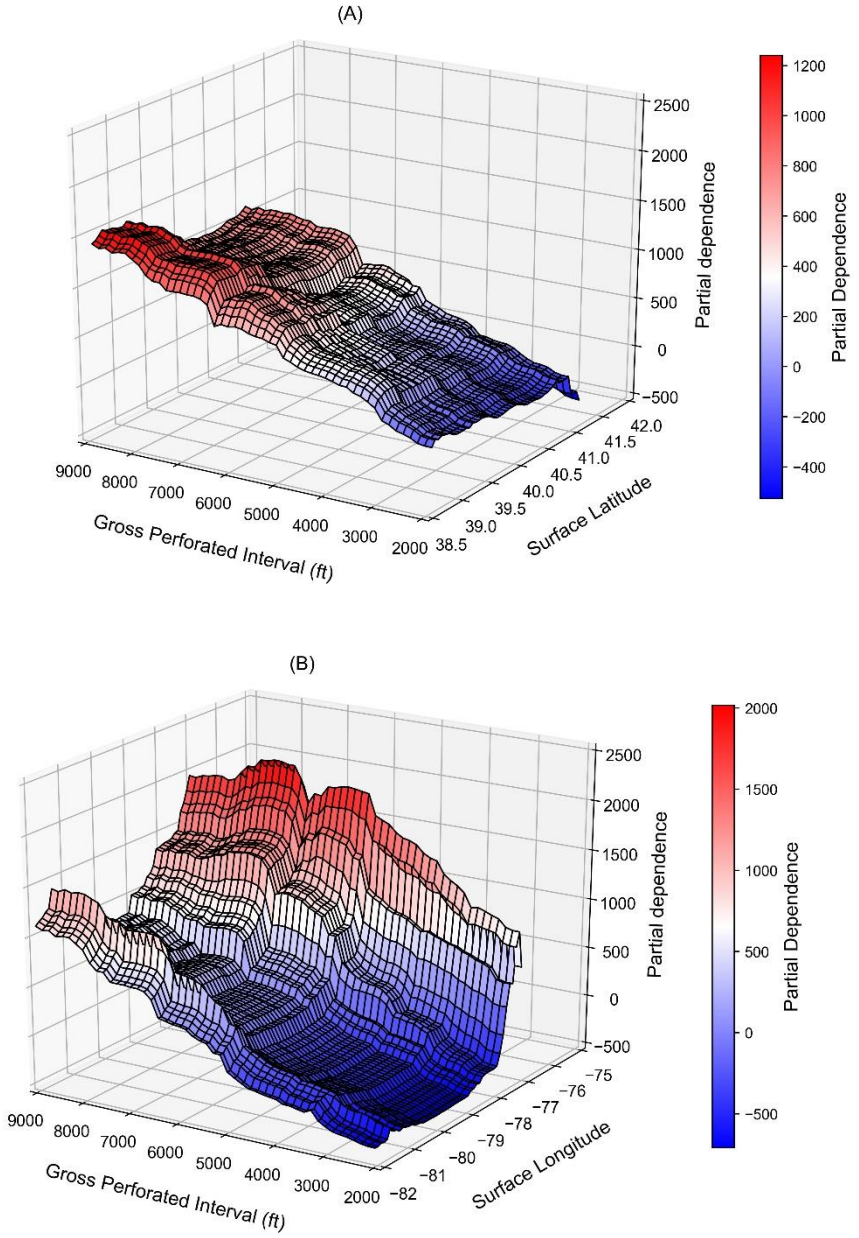

Figure 14. Three dimensional plots of partial dependence for predicting the Top 12-months productivity indicator using the final model formulation. The top figure (A) evaluates the interaction of perforated interval length and surface latitude. The bottom figure (B) evaluates the interaction of perforated interval length and surface longitude.

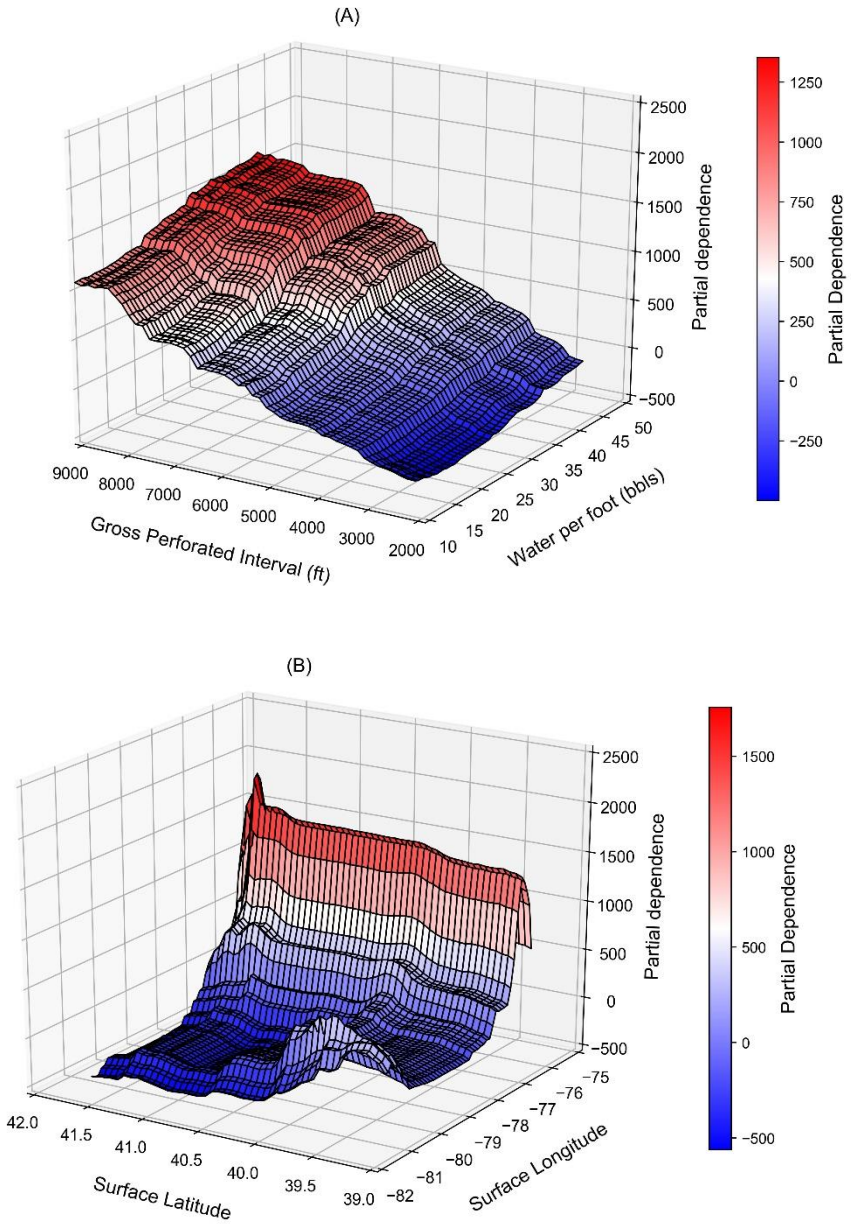

Figure 15. Three dimensional plots of partial dependence for predicting the Top 12-months productivity indicator using the final model formulation. The top figure (A) evaluates the interaction of perforated interval length and water injected per foot. The bottom figure (B) evaluates the interaction of latitude and longitude.

Another notable interaction between perforated interval length and water per foot (Figure 15) indicates that higher productivity has been achieved by applying increased volumes of water during hydraulic fracturing or during any refracturing processes. Given that proppant and additive per foot are correlated to water per foot to some degree (Pearson r of 0.82 and 0.19 respectively), they would correspondingly scale up with increased volumes of water injected.

Figure 15 also evaluates the latitude and longitude interaction. The emergence of the two core regions of the Marcellus Shale can be seen by the prominence of local maxima when the effects from well design are marginalized. The southwestern core is represented by a local maximum between 39.5^{o} to 40^{o} in latitude, and between -80.5^{o} and -79.8^{o} in longitude. The northeastern core dominates the entire eastern portion of Marcellus Shale development and represented by local maxima at longitude greater than -77^{o}. The Marcellus Shale is considered to be generally under-pressured to the southwest and normal-pressured to potentially over-pressured to the northeast, with a transitional area in between. Higher well productivities are expected from the normal to over-pressured areas relative to the transitional area [70, 94]. This dynamic seems to be largely reflected in the model based on the surface response per Figure 15B, which demonstrates the model's utility in capturing the effects from the highly variable geologic conditions across the play. However, the existence of distinctive pressure regimes and highly variable lithology across the Marcellus Shale suggests that tailored approaches to well stimulation and completion would be needed to maximize the production potential depending on where future wells are placed.

2.4.6 Application of Models Through a Case Study

There is substantial utility in evaluating the effects of predictor parameter via partial dependence as highlighted in Sections 2.4.4 and 2.4.5. However, partial dependence is limited to

two-dimensional (i.e., two parameters) representations [91]; but regression models, particularly for complex subsurface applications, likely contain more variables of interest. One of the important functions of regression models, like the ones developed through this study, is in their predictive capability. Given the existence of highly variable geologic conditions across the Marcellus Shale and the noted effects of water and proppant volumes on productivity (Figure 13 and Figure 15 above), the developed GBRT models can be employed to essentially optimize the volumes of proppant and water as part of the well design process. This concept was applied in a case study described below which intends to provide an initial look at using the capability of ML-based models to 1) assess the potential shortcomings in current well designs implemented and 2) help inform the tailoring of future well designs for maximizing the production potential depending on their placement in the Marcellus Shale.

For this example, six existing wells in the study dataset were selected at random across the study area for evaluation. Table 6 provides a summary of the characteristics associated with each and supplemented with approximate geologic data common to each well location acquired from literature [71, 95, 96, 56]. Each well evaluated has different productivity characteristics that may be attributed to spatial placement in the Marcellus Shale, distinctive well design characteristics, and potentially from unique operational circumstances (i.e., reflected in differences in Top 12-months versus First 12-months). Simulations were performed for predicting both productivity indicators (only Top 12-months featured here; First 12-months results can be found in Figure 57) by holding constant all parameters, but enabling water and proppant per perforated foot to vary independently within operationally feasible ranges of +/- 1.5 standard deviations from their mean values across the dataset (originally presented in Table 2).

Table 6. Characteristics from the six wells used as part of the case study.

Parameter	Well A	Well B	Well C	Well D	Well E	Well F
Water per perforated foot (bbls)	36.9	20.1	34.7	33.7	45.7	32.5
Proppant per perforated foot (lbs)	1,684	1,929	1,603	1,347	1,999	863
Additive per perforated foot (bbls)	0.192	1.615	1.248	0.428	2.54	2.96
Perforated interval length (ft)	4,560	1,107	3,005	7,797	3,102	5,186
Well trajectory azimuth (degrees)	337	193	303	344	338	319
Acre spacing (acres)	101	81	66	124	57	345
First production year (year)	2015	2012	2015	2016	2014	2017
Surface hole latitude (decimal degrees)	39.936853	41.677194	40.305922	39.556682	41.675722	39.43821
Surface hole longitude (decimal degrees)	-80.156403	-75.835547	-80.234289	-80.579967	-76.032447	-80.772416
True vertical depth (ft)	8,086	7,456	6,569	7,679	7,391	6,649
Thermal Maturity (% R_o)	1.50	3.30	1.50	1.50	2.75	1.55
Thickness (ft)	65	300	50	55	300	50
Normalized Gamma Ray (API)	570	350	450	453	420	330
First 12-months production (MMcfge)	1,687	2,096	818	1,834	2,310	1,496
Top 12-months production (MMcfge)	1,702	2,434	840	1,899	2,523	1,526

The results from simulations are presented as contour maps in Figure 16, where the colorations represent different levels of predicted Top 12-months productivity in MMcfge as a function of existing well designs and variable proppant and water per foot volumes. Warmer colors indicate higher predicted productivity and colder colors indicate lower predicted productivity. The black dots in Figure 16 represent actual proppant and water per foot design volumes used for each well in the field (which correspond to the associated Top 12-months production value).

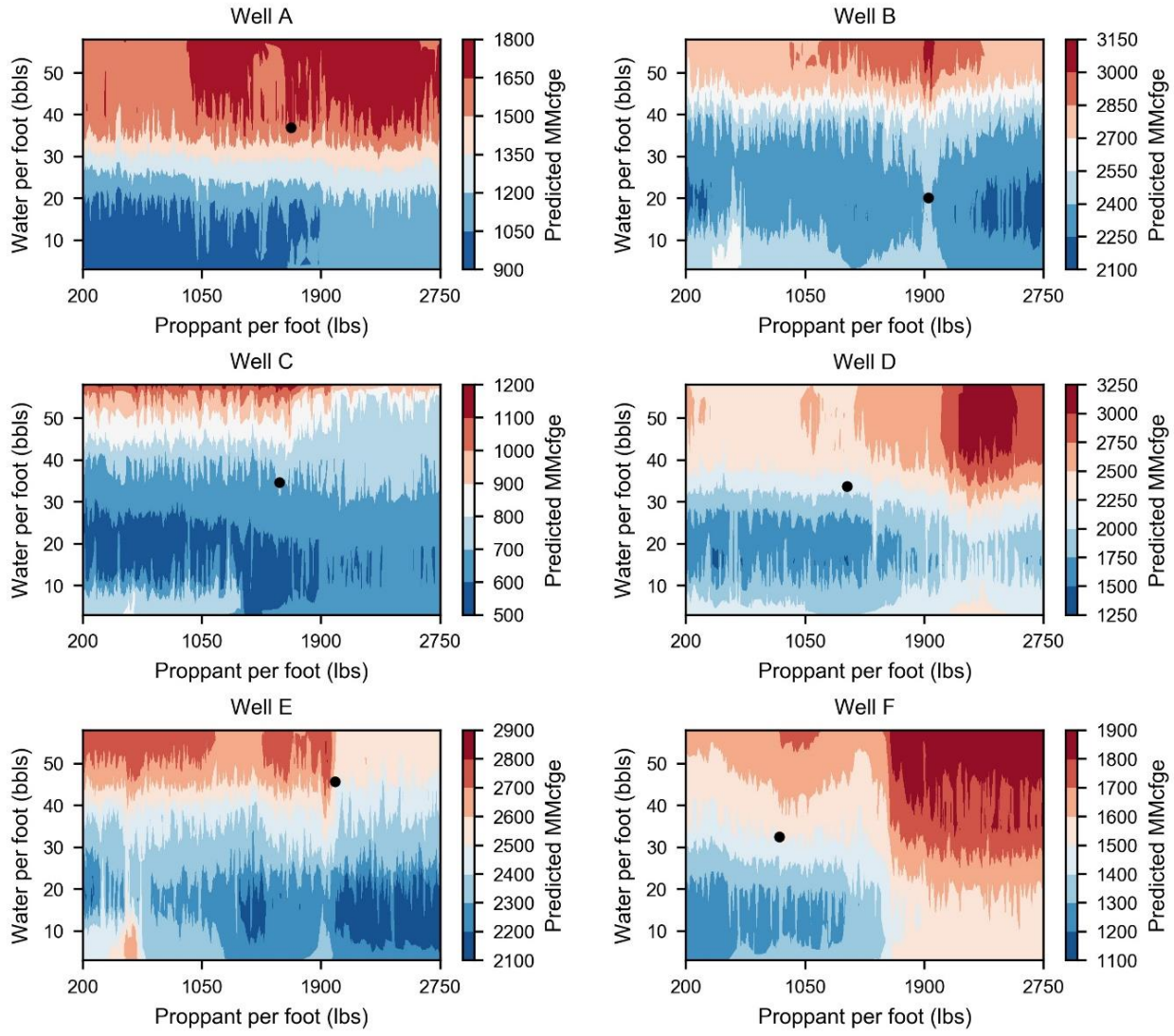

Figure 16. Contour diagrams for estimated Top 12-months production for each well evaluated in the case study with varying water and proppant per foot input values. The black dots represent the implemented field designs for each corresponding well.

Simulation results concur with parameter partial dependence analysis in that water per foot seems to have the largest influence on improving well productivity compared to proppant. However, the optimal combinations of water and proppant (within the bounds of water and proppant levels simulated) for maximizing productivity vary differently across wells evaluated

based on their location. While latitude and longitude data were used as part of the GBRT model input set, the geologic features in Table 5 can be referenced in tandem to provide some insight to why the resulting heat map responses to different proppant and water combinations vary for each well. Additionally, each well appears to fall short of the optimal water and proppant combination that would have ultimately maximize productivity at their respective well locations. For instance, Well A (southwestern core) is near optimum but could improve production by adding more water per foot. Well B (northeast core) could improve productivity by substantially increasing the water per foot to near 50 bbls/foot of perforation without any change to the proppant volume. Well C (liquids rich portion near southwestern core) could benefit by increasing water volume and may see no productivity improvement from any increase in proppant. Well D (liquids rich portion of southwestern core) falls short of the noticeable optimal design point near 45 gallons per foot and 2,200 lbs proppant per foot. Well E (northeast core) is close to an optimal point, but interestingly, the model suggests a subtle reduction in proppant concentration may improve productivity. Finally, Well F (liquids rich portion near southwestern core) could benefit strongly from increasing proppant per foot greater than 1,800 lbs/foot and water to 40 bbls/foot or greater. Further analyzes would ultimately be needed to fully evaluate the economic, logistical, or environmental considerations for deploying wells closer to their optimal design settings (related to proppant and water volumes).

Productivity response is highly variable across the wells evaluated (given that common water and proppant volumes were applied to each well), which suggests that spatial variability in geology is highly influential on production despite the specific choices in well design. However, for each well, unique adjustments to the design choices implemented in the field may potentially improve overall productivity given each well's specific location based on modeling results.

Therefore, moving forward, the bottom-up design considerations for each new Marcellus Shale well will be critical in setting the production trajectory of the play's remaining development potential. Any advances for how wells are designed (from modeling tools like the ones presented in this study, technology improvements, or otherwise) based on the specific geologic conditions for which they are placed can potentially alter future production forecasts, recovery factors, and overall resource estimates across the entire play. While this case study is a relatively straightforward example application (and does not include manipulation of other design parameters like additive, azimuth, or acre spacing), this type of data-driven model has the capability to quickly and effectively evaluate the impact of given well design choices on productivity and potentially help inform future reservoir management strategies.

2.5 Conclusions

This study applied ML to a large dataset that spans across a prominent unconventional reservoir and developed models capable of accurate prediction of productivity indicators at the well level that strongly correlate to EUR. The models developed provide a capability beneficial to reservoir management in the Marcellus Shale that enables fast and effective evaluation of the impact of various well placement and design choices. Prudent and efficient extraction of the Marcellus Shale's remaining resources though the most effective reservoir management strategies are vital to its contribution as a sustained, long-term hydrocarbon asset.

Dependable evaluation of tailored well designs and their associated interactions within the reservoir requires models that capture the inherent complexity associated with unconventional systems with the highest accuracy possible. The models developed here were found to perform

accurately (based on R^2 and RMSE scores against hold out datasets) for predicting the productivity indicators evaluated over a large, play-wide study scale. The noted performance accuracy is attributed to two main factors: 1) the use of GBRT that handles the complexity associated with unconventional O&G systems quite well, and its inherent sequential construction of weak learners that compensates for shortcomings of those previously developed; and 2) the Top 12-months productivity indicator provides added utility for estimating the production potential for a given well based on its design and placement in the reservoir.

A series of analyses were conducted using the developed models to explore the impact and overall effects of predictor parameters on well productivity. Findings demonstrated the importance of the geologic proxy parameters of well surface longitude and latitude on well productivity, as well as gross perforated interval length and water and proppant per foot. Relative improvements in well productivity are tied to upscaling of water and proppant volumes per foot as part of hydraulic fracturing. However, the magnitude of productivity improvements is spatially dependent across the play as influenced by geologic heterogeneity. While upscaling water and proppant are shown to improve productivity, optimal combinations of water and proppant volumes were found to vary depending on well placement within the play as indicated when models were applied under a case study. Additive per foot and acre spacing were shown to have noticeable, but more understated effects on productivity compared to gross perforated interval, water and proppant per foot, and geology. The effect of azimuth was found to be marginal at best despite being widely considered a critical well design consideration in unconventional reservoirs.

While the methodology and models developed as part of this study are specific to the Marcellus Shale, the basic framework can be utilized and applied to other O&G reservoirs relatively quickly. Additionally, the data parameters utilized are relatively common across plays

and may be readily acquired from public sources. Developing and deploying ML technology in O&G applications has the prospective to provide efficient and accurate analytical toolsets that can complement existing best-practices—a combination that can potentially revolutionize how wells are sited, designed, and operated moving forward.

3.0 Machine Learning-Informed Ensemble Framework for Evaluating Shale Gas Production Potential: Case Study in the Marcellus Shale

The following chapter is based on a peer-reviewed journal article published in *Journal of Natural Gas Science and Engineering*, which can be cited as:

Vikara, D., Remson, D., and Khanna, V. Machine Learning-informed Ensemble Framework for Evaluating Shale Gas Production Potential: Case Study in the Marcellus Shale. *Journal of Natural Gas Science and Engineering*. 2020. Volume 84, https://doi.org/10.1016/j.jngse.2020.103679

3.1 Chapter Summary

Artificial intelligence and machine learning (ML) are being applied to many oil and gas (O&G) applications and seen as novel techniques that may facilitate efficiency gains in exploration and production operations. Significant improvements in that regard are likely to occur when ML can be applied to evaluate O&G challenges with inherent synergies that may have otherwise not been evaluated concurrently. This study introduces an ensembled framework that couples a data-driven ML predictive model capable estimating a productivity indicator for unconventional O&G horizontal wells that correlates to estimated ultimate recovery (EUR) with a well design optimization approach that maximizes productivity. The framework is then applied to spatially rank productivity potential from low to high across the Marcellus Shale. The ML model developed used a gradient boosted regression tree (GBRT) algorithm and is capable of 82 percent prediction accuracy on holdout data. The distribution of geological properties as well as the resulting

optimized well design and completion attributes specific to regions commonly ranked in productivity potential are evaluated statistically to comprehend controlling factors on shale well production, and to identify if commonality or disparity exists in the prominent features. The highest productivity ranked region is isolated in the Marcellus Shale's northeastern core region and its periphery. Statistical analyses indicate that regions higher in productivity ranking show a significant difference for certain (but not all) geologic features favorable to gas production potential relative to lower productivity regions; most notably net thickness and porosity. Optimized well design parameter settings vary relative to their placement across the study area and subsequent productivity ranking region. Overall, the ML-based framework discussed in this chapter attempts to analyze shale controlling factors concurrently, to deliver a systematic evaluation result for production potential that accounts for and quantifies controlling features associated with geologic properties and well design attributes.

3.2 Introduction

Horizontal drilling and hydraulic fracturing are transformational technologies that have enabled the widespread development of unconventional oil and gas (O&G) reservoirs that were otherwise uneconomical to develop. Expansive development of unconventional O&G resources has facilitated substantial growth in hydrocarbon production enabling an energy abundance – the result of which has led to a revolution in the energy landscape of the United States (U.S.) as well as in energy markets across the globe [2, 47, 48, 97]. However, more recently, the increased supply (particularly for natural gas) in the U.S. has been coupled with decreasing demand growth resulting in suppressed oil and gas prices. Additionally, in certain cases, a lack of pipeline infrastructure prohibits hydrocarbon supplies to potentially emerging markets; a circumstance which prevents a possible increase in overall hydrocarbon demand.

Suppressed market prices for O&G can diminish the economically-viable portions of existing O&G assets unless 1) technological advancements that can improve productivity can keep pace with declining market prices or 2) mechanisms that lower cost and/or increase operational efficiency are developed and implemented. Such a predicament may compel O&G operators to optimize unit costs ($/unit of energy) by balancing higher-cost completion designs with enhanced performance, as well as through improved operational efficiencies [98]. Artificial intelligence and machine learning (ML) have emerged as promising approaches that may reshape the exploration and production landscape for the O&G arena and provide new techniques for operators to consider in the pursuit of lowered costs and improved recovery objectives [38, 39, 40, 99, 31]. The International Energy Agency estimates that widespread use of digital technologies like ML could, for instance, increase O&G reserves by about five percent and reduce production costs by 10 to 20 percent [45]. Furthermore, the increase in the types and volumes of digital data formats becoming

available via unconventional O&G development [32, 53] makes applying ML for use cases in O&G and other subsurface applications highly intriguing.

The challenge in developing unconventional fields lies in the inherent high-dimensional decision-making problem, especially when substantial uncertainty may exist with subsurface conditions and prevailing economic factors [100]. An integrated ML-based decision-making framework that can accommodate two or more field development dimension considerations would seem to provide efficiency improvements in multiple facets. A multi-faceted but correlated example O&G operators routinely manage include 1) cost-effectively locating drilling locations with high production potential (i.e., sweet spots) in a regionally extensive and highly heterogenous unconventional play and 2) engineering well designs tailored to geologic conditions present at new well sites that maximize production potential. In general, sweet spots are target locations or regions within a play or a reservoir that represent known high productivity or have the potential for high productivity. These locations are often delineated to enable wellbore placement in the most productive areas of producing reservoirs. Sweet spots in unconventional reservoirs are typically defined by several controlling geologic factors (described in detail in Section 3.3). Traditionally, geoscientists and engineers have employed the likes of core analysis, well log data, and seismic attribute data to attempt to identify regions with high productivity potential [101, 102, 103]. A challenge is that existing approaches for sweet spot identification often require disparate datasets that are expensive to acquire and are specific to an isolated portion of a potentially expansive hydrocarbon play. Given the current O&G economic climate (April 2020 values: Henry Hub Natural Gas Spot Price - $1.74 USD per MMBtu; West Texas Intermediate Crude Oil Price - $16 per bbl), continued investment in expensive data acquisition pursuits to inform reservoir management approaches seems disadvantageous. The disparate types of data required for

traditional analyses also present a challenge to integrate when they occur on different scales – both spatially and temporally [104]. In addition, these forms of reservoir evaluation provide a comparison of relative productivity potential within the area evaluated based on geologic conditions only. They can inform the well completion design process but do not translate directly to well productivity. Furthermore, they cannot be scaled across an expansive play given data requirements; and interpolation approaches can be fraught with uncertainty when applied over large spatial domains [77].

Another on-going field development challenge has been determining the controlling factors on shale well productivity potential given geologic conditions inherent to new drilling sites and well completion design choices. Physics-based multiphase flow models are widely used tools for gaining an understanding of the subsurface response to engineered permutations from operations like geologic carbon dioxide storage, liquid waste disposal, as well as hydraulic fracturing processes. However, predictions from multiphase flow modeling are known to contain inherent uncertainty given the challenges associated in sufficiently representing the heterogeneity in subsurface media. Furthermore, challenges may exist in procuring geologic data in sufficient volumes at the lowest level possible (for instance, at the well level) to build representative reservoir models [51]. The more typically used physics-based fluid flow models for reservoir simulation can also be both time and computationally rigorous to carry out [47, 52, 32, 105]. However, given certain reservoir geological properties at potential drilling sites, major production efficiency gains may come through determining the best tailored completion strategy that maximizes the production potential at that location. Machine learning-based models may offer a faster and potentially more reliable complement to commonly used reservoir modeling when evaluating well designs that can be tailored to site conditions and more amenable to productivity gains.

Several studies have examined ML-based approaches for a variety of different O&G applications. Examples of topics investigated include modeling hydrocarbon production in unconventional O&G plays [56, 15, 58, 16], well drilling operations [17, 18], lithology and geologic facies classification [21, 19, 20, 106], subsurface fault detection [22], operational predictive maintenance planning [23, 24], and detection of potentially high producing pay zones sweet spots [25, 26]. The studies under each topic provide innovative aspects which may facilitate exploration and production efficiencies and/or more cost-effective approaches, enabling sustainable development of hydrocarbon resources moving forward. However, significant improvements in that respect may likely occur when two or more of the concepts are integrated into a more robust decision informing framework; one that connects the cause and effect of applications included and when applied, may offer improvements over common benchmarks or industry best practices.

Machine learning has recently been applied as an innovative approach addressing the challenges described above, most notably 1) delineating and appraising the productive quality in unconventional reservoirs and 2) generalizing unconventional well productivity and informing well completion designs. In terms of appraising reservoirs, Tahmasebi et al. (2017) developed a hybrid ML technique that integrates neural networks and fuzzy logic; an approach that can effectively predict total organic carbon (TOC) and reservoir fracturability given a suite of well log data and mineral compositions from the x-ray diffraction analysis [25]. Their approach helps in estimating the probability of targeting sweet spots, as well as identify the necessary well log data needed to inform reservoir evaluation most effectively. Qian et al. (2018) developed a workflow based on fuzzy mathematics and support vector machines that enables characterization of sweet spots in unconventional O&G reservoirs by correlating seismic attributes to petrophysical

characteristics [26]. In mature plays, the use of historic well completion and production data has facilitated analyses and developed alternative approaches for evaluating new well sites over larger spatial domains [107, 108, 59, 109, 110]. This same concept can be applied using ML on existing data that is widely available and low-cost to attain. Regarding well optimization, Wang and Chen (2019) used an adaptive boosting ML algorithm to formulate predictive models for oil production focusing on the Montney Formation in Canada [15]. Their study implemented sensitivity analyses modifying the proppant and water intensity of in-field well designs to determine optimal attribute settings that maximized productivity. Shih et al. (2018) performed a similar optimization exercise using Kernel Ridge regression on hydraulic fracturing additive and water usage in horizontal wells in the western region of the Marcellus Shale [56]. Luo et al. (2018) is another example in which a neural network model was developed to correlate the relationship between the first-year oil production in the Bakken with input parameters consisting of well design attributes and geologic properties. The neural network model was used to perform a sensitivity analysis to investigate the relationships between well completion strategies and geologic conditions on the first-year production response [16].

The application of ML has proved to provide fast, accurate, and cost-effective analytical approaches to many O&G-related problems and can be used to analyze disparate datasets in innovative ways. However, many of the existing ML-based approaches that have proven capable for specific applications (i.e., identifying sweet spots, generalizing well performance, characterizing lithology, etc.) are rarely applied in ways to demonstrate notable improvements over more widely used field development methods. These types of ML topics have been explored mostly in parallel, but larger benefits may exist by coupling approaches at various scales (pad, field, or basin-scale). This concept could facilitate the development of more robust decision-

making frameworks, thereby helping assure longer-term sustainability and overall improvement in unconventional field development moving forward.

This study introduces an ensembled framework that couples a data-driven ML predictive model capable estimating a productivity indicator for unconventional O&G horizontal wells that correlates to estimated ultimate recovery (EUR) [67] with a well design optimization approach that maximizes productivity. The novel framework is applied to the Marcellus Shale unconventional reservoir located in the Appalachian Basin of the northeastern U.S. The Marcellus is a relatively mature play with approximately a decade of production history. As a result, data exists for wells from multiple operators that have employed a variety of well design options across the assortment of geologic conditions of the play. The framework is applied in a fashion to rank productivity potential across the Marcellus. The distribution of geologic properties together with the resulting optimized well design attributes specific to regions common in productivity potential are evaluated statistically to analyze controlling factors on shale well productivity potential. This step will identify if statistically significant commonality or disparity exists in the prominent geologic and/or well design characteristics depending on their geographic placement across the Marcellus. The insights gained should be advantageous in both the 1) identification of high-priority drilling regions and their geographic extent along with 2) informing future well designs tailored to their geographic positioning and spatially-dependent geologic conditions across the Marcellus that can potentially offer improved productivity.

3.3 Overview of Controlling Factors for Gas Production in Shale Reservoirs

Comprehensive evaluation of the controlling factors influencing unconventional gas recovery requires consideration of both geologic and engineering (i.e., well design) aspects [26]. Specifically, well stimulation and completion designs need to be tailored to maximize productivity potential for any given well site based on the controlling geologic conditions present. Some of the major geologic parameters that influence a given unconventional play's boundaries and productivity potential include total organic carbon (TOC), formation thickness, porosity, hydrocarbon and/or water saturation, thermal maturity, depth, pressure, presence of existing fracture networks, and ability of the formation to be hydraulically fractured (i.e., fracturability) [111, 107, 112, 102]. Several of these parameters, which are explored to some degree in the analyses presented later in this study, are briefly discussed below as they relate to shale gas productivity potential. Aside from the effects on hydrocarbon production, these geologic controlling factors can also strongly influence well drilling logistics and safety concerns, but those are not discussed here.

- **Total Organic Carbon -** TOC is widely considered amongst the most crucial parameters for shale reservoir evaluation. Thermal degradation of organic material like kerogen and bitumen creates hydrocarbons – the process is what defines a source rock. Generally, in shales that are thermally mature (> 1 % vitrinite reflectance [R_o]) the higher the TOC, the greater the potential exists for higher gas adsorption capacity [111]. However, measurable TOC tends to decrease with increasing thermal maturity and subsequent hydrocarbon generation [113]. TOC is typically required to be above two percent to have viable production potential [114]. Higher TOC content in unconventional reservoirs tends to increase porosity, permeability, and fracture density,

and is shown to reduce the overall bulk density of the reservoir [114, 101, 115]. One of the best proxy measurements of TOC content in a reservoir is gamma ray count as well as resistivity. A strong correlation exists between reservoir organic content, well log gamma ray intensity, and high resistivity [116]. As an example, five percent TOC content can be detected with gamma-ray counts of 200 API units or greater. Gamma-ray counts in the lower member of the Marcellus formation often exceed 400 API units, which generally indicates higher TOC contents in the basal part of Marcellus [70].

- **Thickness -** The thickness of an organic-rich unconventional reservoir may or may not correlate with overall hydrocarbon storage capacity and productivity potential, because the entire reservoir segment may contain both productive and relatively non-productive rock intervals. Therefore, thickness is typically demarcated in terms of gross and net thickness. Net thickness is the aggregate sum of the reservoir intervals with geologic properties favorable for hydrocarbon production. Gross thickness is the total thickness of the stratigraphically defined reservoir interval which may include non-productive segments with unfavorable geologic conditions for hydrocarbon production interbedded within. The larger the net thickness, the greater the likelihood of overall hydrocarbon storage capacity and productivity potential.
- **Porosity -** Reservoir porosity consists of the pore space that can potentially hold gas (or oil), but may also contain water. The greater the porosity, the greater the gas storage potential. Unconventional reservoirs may contain both free gas and adsorbed gas. Free gas is stored in the effective porosity of the rock matrix (gas saturation = 1 – water saturation) and the absorbed gas adheres to the organic content of the reservoir. Low reservoir density typically correlates with higher porosity, especially because organic-

hosted porosity dominates unconventional shale porosity, and organic material has lower bulk density relative to the mineralogical content. While no standard logging technique for assessing porosity in situ exists, neutron, density, and acoustic velocity (or sonic) logs can be used to infer porosity [117]. The presence of natural fractures can provide additional pore space for hydrocarbon accumulation, as well as a flow path network favorable to productivity. Alternatively, natural fractures can be thief zones for expulsion and migration of hydrocarbons out of the reservoir over geologic time, or be thief zones for hydraulic fracture energy over the completions timeframe.

- **Resistivity –** Resistivity is a material property associated with how strongly a material opposes the flow of an electric current. The materials within most rock matrices are inherently insulators and resist current, but fluids in the pore space could be either resistive or conductive. For instance, oil and gas have higher resistivity than water. Therefore, a favorable productivity interval may contain high resistivity log readings [118] when coupled with higher porosity values. Induction logs are commonly used to assess resistivity in situ by measuring the conductivity of rock formations by using electromagnetics.
- **Depth -** True vertical depth can impact multiple geologic properties. Deeper reservoirs may experience greater compaction effects, resulting in reduced porosity and permeability [114], which could be detrimental to gas productivity. However, hydrostatic pressure increases as a function of depth but may vary location to location (from under- to normal- to over-pressured) dependent upon on the geomechanics of the reservoir and seal, the structural evolution of the region, and the basin's burial history. Generally, over-pressured unconventional reservoirs correlate to higher production

rates [119]. Depth also correlates with thermal maturity within a region with similar basin burial history.

- **Fracturability -** The mineral composition of the reservoir controls the brittleness, a favorable condition that enables fracturability. The higher content of brittle minerals like quartz and calcite generally better enables fractures – either natural or operationally-induced. Fracturing is critical to productivity by establishing hydrocarbon flow pathways from the reservoir to the producing well. The presence of higher content of ductile material like clay, on the other hand, can lead to flow path blockage [120]. Reservoirs high in Young's Modulus (reflection of rock rigidity) and low Poisson's ratio (reflects rock elasticity) typically possess greater fracturability.

The impact of many of these factors on the upper limit of the gas available in situ can be observed mathematically. The volume of original gas in place (*OGIP*) in standard cubic feet (scf) in unconventional reservoir that includes free and adsorbed gas per Richardson and Yu (2018) is outlined in Equation 3-1 [121]:

$$OGIP = 43{,}560\frac{Ah_s\left[\emptyset_m(1-S_w)+\emptyset_{frac}(1-S_w)-\emptyset_a\right]}{B_g}+1{,}359.7Ah_s\rho_b\upsilon_a \qquad \text{Equation 3-1}$$

Where A is the drainage area (in acres), h_s is pay zone thickness (ft), ϕ_m and ϕ_{frac} are matrix and fracture porosity respectively (fraction), S_w is the water saturation (fraction) which may vary between the matrix and fractures, ϕ_a is pore space occupied by adsorbed gas (fraction), ρ_b is the bulk density of shale (in g/cm^3), υ_a is the specific volume of gas absorbed per unit mass of shale (standard ft^3/ton), and B_g is the formation volume factor (reservoir ft^3/standard ft^3). The volume of *OGIP* per Equation 3-1 generally increases with thicker, more porous rock with high gas and lower

water saturations. With increasing pore pressure, the υ_a term may increase as well, resulting in a greater amount of the overall gas volume in place. Equation 3-1, however, does not account for hydrocarbon flow to the wellbore; an aspect influenced by both the presence of natural fractures as well as those induced by the hydraulic fracturing process.

Several studies have classified and ranked prominent controlling geologic attributes in unconventional reservoirs [114, 101]. However, little emphasis has been placed on how the distribution of various controlling geologic parameters at candidate drilling sites can be collectively evaluated and used to inform well designs options that maximize productivity. Additionally, variation that likely exists in controlling geologic characteristics from one site to another given inherent heterogeneity makes differentiating prospective drilling sites from a productivity perspective a challenge. This study implements an evaluation approach that enables ML to leverage existing well design and production data to differentiate and rank productivity potential in the Marcellus from a spatial perspective. The controlling geologic factors are compared within and across regions differentiated as a function of well productivity by analyzing well log data post-hoc to the systematic assessment of productivity using ML. Since the availability of comprehensive, reliable, and represented geologic data associated with the shale gas controlling factors for production at the well-level (where production was observed to complement the well design attribute data) is not feasible to acquire at large scales, these types of data were not integrated into the front end (i.e., ML development component) of the modeling framework.

3.4 Data and Methods

The predictive model utilized for this study was built on a Gradient Boosted Regression Tree (GBRT) ML algorithm. The model estimates natural gas equivalent hydrocarbon production for horizontal wells producing from the Marcellus Shale. Through our previous research, we have found that GBRT performs admirably in generalizing the performance of horizontal wells in the Marcellus [122]. Developed models are then used to simulate well productivity across a bulk of the Marcellus producing region in an attempt to 1) grade and rank regions based on productivity potential from low to high and 2) identify the optimal well design configurations within each region driven by controlling geologic conditions. Statistical analyses are then performed to identify differentiable geologic and/or well design attributes between graded regions. This analytical approach that leverages readily available datasets provides a fast and efficient compliment that can supplement existing prospecting site selection and reservoir management practices. Additionally, the insights gained should be useful in informing tailored well designs given their potential drilling location. Figure 17 highlights the study framework that was implemented. The following sub-sections discuss the framework component in detail.

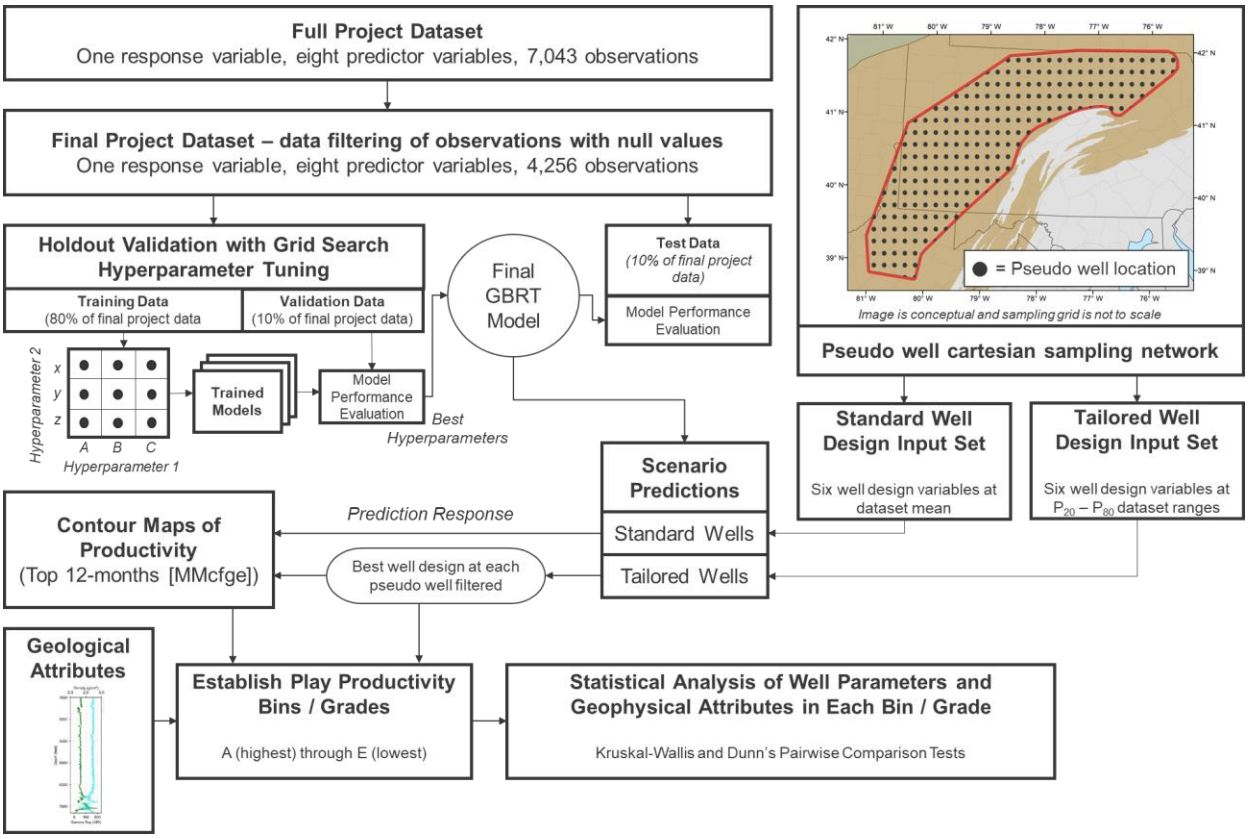

Figure 17. Study framework using GBRT-based machine learning predictive models to grade and rank producing regions in the Marcellus.

3.4.1 Study area overview

Well data used for this study was obtained from the O&G data vendor DrillingInfo [123] and consists of horizontal production wells placed in the Marcellus all with first production dates between the January 1, 2010 to December 31, 2018 timeframe. The resulting dataset consists of 7,043 wells in total extending across Pennsylvania and northern West Virginia. An initial data preprocessing step included data filtering for dataset entries with missing data values for either predictor or response parameters of interest (Table 7); no attempts at data interpolation on missing values were made. The resulting dataset consists of 4,256 wells and is the identical to the dataset

used in our previous study [122] – all of which contain data for every predictor and response parameter evaluated. Figure 18 delineates the extent of the study area and the distribution of well data available. The study area includes the most currently active, appraised, and developed portions of the Marcellus.

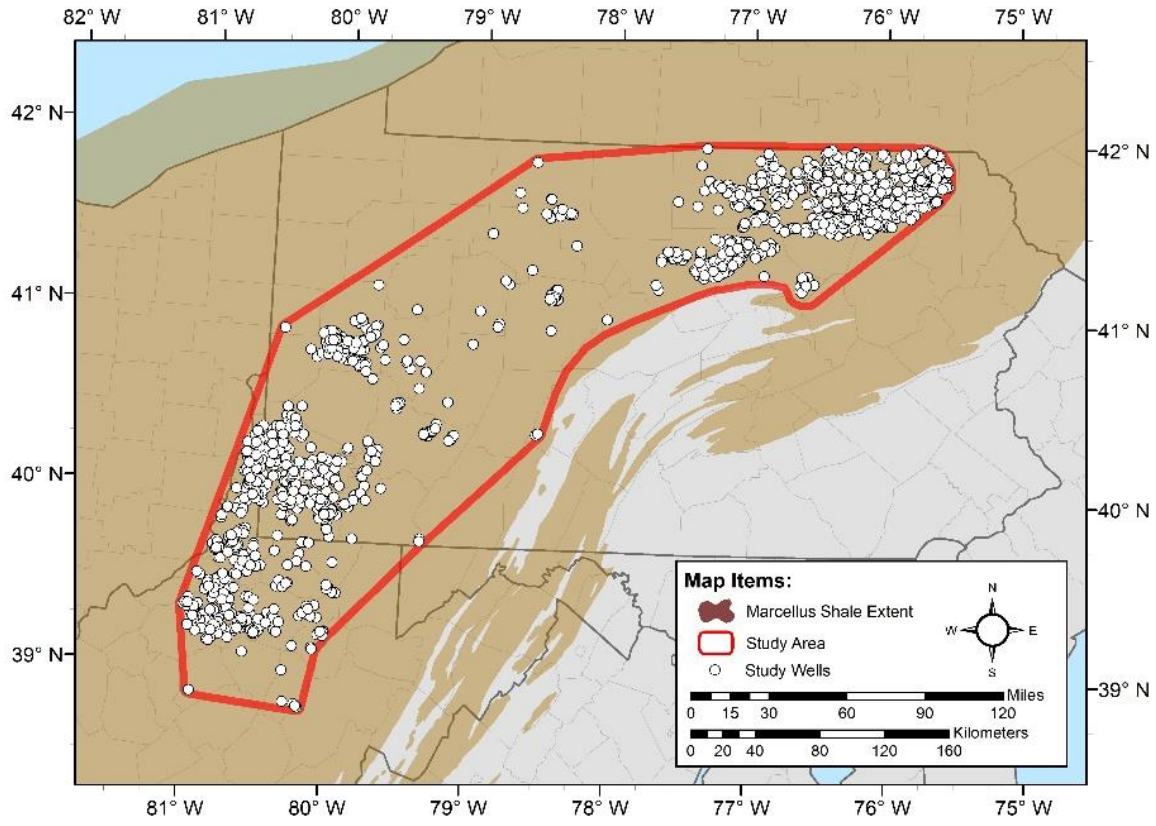

Figure 18. Map outlining the study area of interest and the well set used as part of the study. Well data is used in the machine learning workflow to train, validate, and test predictive models.

The Marcellus Shale is a marine sedimentary carbonaceous rock of Middle Devonian-age located in the Appalachian Basin of the U.S. The formation covers nearly 95,000 square miles and underlays portions of New York, Pennsylvania, Ohio, Maryland, West Virginia, Virginia, Kentucky, and Tennessee, as well as extending into Ontario, Canada, [70, 124, 125]. Since 2013,

the Marcellus has become the largest gas producing play in the U.S. [71] and is expected to be a vital natural gas resource moving forward with estimates of recoverable resource potential ranging from 220 to upwards of 560 trillion cubic feet (tcf) of natural gas [126, 127, 2]. Implementing effective and efficient reservoir management strategies when planning production towards the remaining resources of the Marcellus Shale will be critical for the play to remain a leading hydrocarbon asset over the long-term. The decade of the Marcellus' production history has enabled a wealth of data from which ML based models can be developed. Models like these can help inform the well designs for future well sites, reservoir management decisions, potential infill development, and even in retesting initial step-outs wells that occurred in drilling regions peripheral to core producing areas early in the play's development.

The Marcellus' geological characteristics have been described in detail by the likes of Milici and Swezey (2006), Engelder and Lash (2008), Boyce and Carr (2009), Zagorski et al. (2012), and the U.S. Energy Information Administration (2017) among others [94, 128, 70, 129, 130] and are therefore not described in length here. However, not uncommon to other continuous and expansive O&G plays, prominent reservoir properties that define the Marcellus boundaries include burial depth, pressure, TOC, thermal maturity, reservoir gross and net thickness, porosity, permeability, gas-in-place, fracturability (brittle vs. ductile), and the prominence or absence of natural fractures. Geologic properties are vastly diverse across the extent of the Marcellus due to the reservoir's inherent heterogeneity [71, 73]. Ranges of several of these reservoir properties across the producing sections of the Marcellus include thermal maturity between 1.23 to 3.5 percent $R_{o,}$ 1 to >18 percent TOC content by weight, 4 to 20 percent total porosity, 0.4 to >0.8 psi/foot pressure gradient, 3,900 to >8,500 total vertical depth, and 50 to 350 feet gross thickness [94, 131, 132].

Two prominent core areas exist in the Marcellus. Each core region has facilitated higher productivity when compared to other portions of the play. One core region spans across portions of southwestern Pennsylvania and northern West Virginia (southwestern core) and the second is located in northeastern Pennsylvania (northeastern core). The core regions are known to contain disparity in their controlling geologic characteristics that make each uniquely distinct. Relative to the northeastern core area, the southwestern core is characteristically higher in TOC, has a higher net-to-gross thickness ratio (due to absence of limestone intervals, and a more distal depositional environment resulting in a condensed section), a smaller gross thickness, is more porous and permeable, contains a lower pressure gradient on average, and is less thermally mature. In contrast, the Marcellus' northeastern core is substantially thicker, known to have a higher-pressure gradient, and is more thermal mature [74, 70, 75]. Production data has indicated that wells in the northeastern core, where the play is at its thickest relative to other areas are, in general, larger producers of natural gas. However, thermal maturity conditions favor portions of the southwestern core to be abundant in natural gas liquids. As a result, these areas of the play may be more enticing to operators given the potential economic benefits over the dryer gas areas prominent in the northeast. Aside from the core areas, the Marcellus' periphery remains largely underexplored and may contain segments of favorable production potential for future development [71].

3.4.2 Gradient boosted regression tree model overview

A GBRT-based ML model capable of predicting natural gas equivalent hydrocarbon production from horizontal wells in the Marcellus Shale was recently developed in our previous work [122]. The model predicts, either the Top 12-months or First 12-months production indicators, both of which have shown to that correlate to well-level EUR. The GBRT approach

proved to provide for an accurate predictive model for predicting both production indicators, but showed better performance when predicting the Top 12-month indicator. As noted in our previous study, the Top 12-month indicator effectively represents productivity potential for wells with or without disruptions to their production time-series profiles; whereas the First 12-month indicator is less effective for wells when interruptions to their production time-series from the distinctive "peak and decline" trends exist. The Top 12-month metric is calculated by summing the largest 12 monthly production values in gas equivalent (million cubic feet of gas equivalent [MMcfge][6]) for a given well irrespective of when those months occur during a well's productive lifetime.

For this study, the GBRT modeling approach and dataset presented in Vikara et al. (2020) were leveraged to assess productivity over the extent of the study domain by generating predictions for the Top 12-months production indicator given a few subtle adjustments to the overall model development approach. This model uses an input set that consists of multiple well completion attributes along with well spatial coordinate data; the latter serving as a proxy for spatially-dependent geologic conditions (Table 7). The Vikara et al. (2020) study conducted a feature engineering and selection approach which was used to refine the input parameter dataset for the final model formulation. The approach consisted of one-at-a-time parameter removal where training and validation datasets are altered to reflect the omission of a given parameter from all of the attributes evaluated. GBRT-based models were then re-trained on the modified predictor dataset and the overall prediction accuracy (R^2 and root mean squared error [RMSE]) was quantified against the validation dataset. The process was repeated by swapping the previously

[6] Determining the gas equivalent involves combining gas and/or oil production values for each well into a single unit by considering one barrel (bbls) of oil to contain six times the British thermal units as one thousand cubic feet of gas.

omitted parameter back into formulation for another. The approach was similar to recursive feature elimination, one parameter at a time sensitivity analysis [31], and the R^2-loss evaluation approaches that have been used by others [64, 38]. The result of the feature engineering and selection approach conducted by Vikara et al. (2020) established the final model formulations that offer the most accuracy and reliability, and are built on a refined but essential set of predictor parameters. The resulting parameter set consist of those presented in Table 7.

Table 7. Predictor and response variables for the GBRT-based predictive model.

Variable Category	Variable Type	Parameter	Mean	Std. Dev.
Well Design	Predictors	Water per perforated foot (bbls)	32	19.1
		Proppant per perforated foot (lbs)	1,475	866
		Additive per perforated foot (bbls)	1.54	3.81
		Perforated interval length (ft)	5,501	2,088
		Well trajectory azimuth (degrees)*	325	29.3
		Acre spacing (acres)	150	126
Geology Proxy		Surface hole latitude (decimal degrees)	40.643055	0.97
		Surface hole longitude (decimal degrees)	-78.721317	1.95
Productivity Indicator	Response	Top 12-months production (MMcfge)	1,637	1,084

*Well azimuth trajectory data was normalized between 180º and 360º to set all well data to consistent orientations

A feature importance evaluation was performed on the input parameters (data in Table 7) for the GBRT-based models conducted in the Vikara et al. (2020) study. The evaluation included analyzing the relative importance of predictor variables [90] to evaluate parameter importance hierarchy, and a partial dependence analysis [93, 85, 91] to evaluate functional relationships between predictor variables and response. Perforated interval length, the geologic proxy parameters of well surface longitude and latitude, well spacing, and water intensity were shown to

have the most notable effects of on the production response (Top 12 months) when both analyses are evaluated in tandem. The well design attributes consisting of perforated interval length, well spacing, and water intensity were shown to demonstrate positive linear relationships with production. Conversely, the effects of both latitude and longitude on production were highly non-linear. Partial dependence analyses focusing on the concurrent effects of latitude and longitude in the Vikara et al. (2020) study highlighted an emergence of the two core regions of the Marcellus (described earlier in Section 3.4.1) that elicited higher effects on productivity than areas in the play's periphery regions when the effects from well designs were marginalized.

A GBRT ML algorithm was implemented as part of model development. GBRT is a powerful statistical learning technique that can solve both classification and regression style problems. GBRT-based models also enable straightforward feature importance evaluation as well [133]. Our previous work has demonstrated that GBRT, when applied to complex unconventional O&G systems, performs quite well [122]. Additionally, the use of GBRT enabled unique feature importance evaluation to better understand the hierarchy of both well design and geologic reservoir quality variables on model response. GBRT produces a final prediction model in the form of an additive ensemble of weak prediction models, typically decision trees. The model is built in sequential fashion where new decision trees are fit to prior stage model residuals [68, 81, 80]. This "boosting" approach has been noted in many instances to provide improvements compared to other decision tree-based algorithms (like random forests which use a bagging approach) because of the way the algorithm sequentially addresses prediction shortcomings [68, 134, 135]. The final model is therefore a linear combination ensemble of every decision tree (Equation 2-3). Each decision tree within the ensemble contributes to solving for a response variable (y).

$$F(x) = \sum_{k=0}^{K} \beta_k h(x; a_k)$$ Equation 3-2

Per Equation 2-3, the GBRT algorithm aims to approximate a final model *F(x)* which minimizes a loss function against the training dataset through a weighted sum of basis functions *h(x; a_k)* called "weak learners," which take the form of the decision trees{*0, k ... K* number of trees} of a specific size [79]. Each weak learner within the ensemble is function of input variables (*x*) with parameters $a = \{a_1, a_2, \ldots a_K\}$. Expansion coefficients (β) provide the weighted sum contribution for each weak learner and are fit to training data along with parameters (*a*) in sequence. An initial weak learner model $F_0(x)$ is first established followed others in greedy fashion set that the specified number of trees and fit to prior stage residuals. Boosting algorithms, like GBRT, are becoming broadly utilized in several data-science applications due to noted improvements realized in model accuracy. However, they have been narrowly applied in studies evaluating unconventional O&G production with a few noted exceptions [55, 15].

The model development and validation approach implemented as part of the Vikara et al. (2020) study was applied here but adapted through two steps. The first step involved randomly subdividing the final project dataset into training, validation, and testing datasets through an 80/10/10 percentage-based split. A validation holdout approach [136, 137] was then implemented as part of model training, hyperparameter optimization, and performance testing. This adjustment from the previous 60/20/20 split enables a larger and more diverse training dataset while still providing a validation dataset (first holdout set) to confirm optimal hyperparameter settings as well as a testing dataset (second holdout set) used to evaluate the finalized model performance other holdout data. An exhaustive grid search approach was used in which different models were built on the training data for the distinctive hyperparameter combinations evaluated [89]. Trained

models were then used to make predictions against the validation dataset. The model resulting from the hyperparameter combination that yields the most accurate model while avoiding over or underfitting was selected as part of the finalized model formulation. The second step involved evaluating final model performance on the 10 percent subset holdout test data, as well as to provide additional confirmation that models were not over or underfit.

These modifications resulted in 1) confirmation that the optimal combination of GBRT-related hyperparameters for predicting the Top-12 months production indicator response are the same as the original Vikara et al. (2020) study (Boosting stages [i.e., number of trees] = 3,000; Minimum sample splits = 10, Maximum depth [i.e., tree size] = 14, and Learning rate = 0.005; least absolute deviation was used as the loss function for all hyperparameter combinations evaluated) and 2) an improvement in overall model performance when predicting against the holdout test dataset (R^2 = 0.82; root mean square error [RMSE] = 397) (Figure 19). Python version 3 and packages within the scikit-learn library [79] were leveraged as part of the ML workflow implementation.

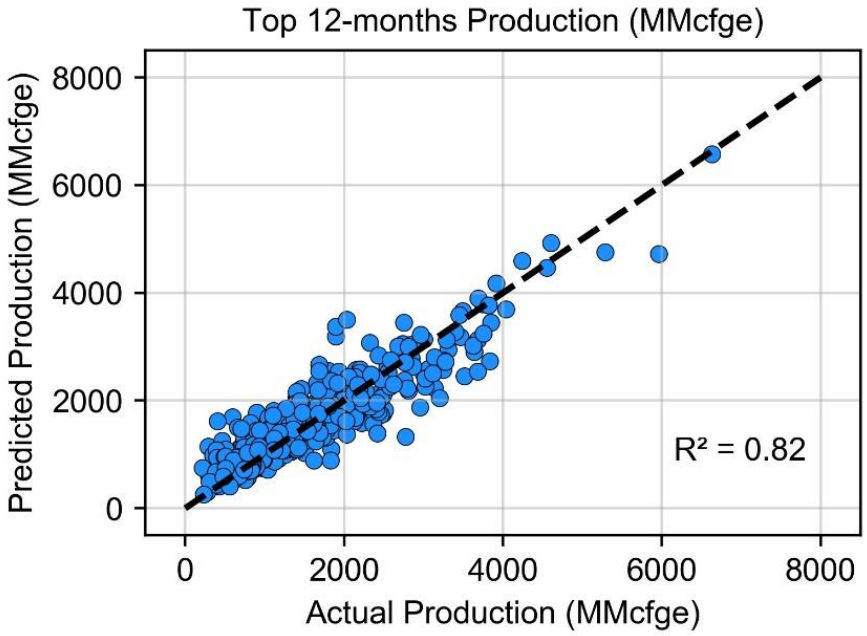

Figure 19. Scatter plot for evaluating model performance of the GBRT machine learning predicted values for Top 12-months production against the actual (i.e., observed) values for wells in the testing dataset.

3.4.3 Simulation approach and productivity contouring

The GBRT machine learning model was used to estimate the Top 12-month production productivity indicator over the entirety of the study area under two different well design scenarios. The resulting data was then plotted spatially and contoured against Top 12-month production estimates across the play. The contours enable spatial assessment of the productivity potential across the Marcellus' producing region. Top 12-month production estimates were determined through simulation at evenly-spaced points (called "pseudo wells") located at centroids of sub-grids within a larger cartesian grid framework fit to the study area outlined in Figure 18 using geographic information system (GIS) tools. The cartesian grid was trimmed at its peripheral to ensure simulations only where in-field well observations used to train the GBRT model had

occurred by: 1) fitting a convex hull envelope [138] around the perimeter of the study wells (Figure 18) and trimming pseudo wells outside the envelope boundary; 2) trimming pseudo wells outside of the Marcellus geospatial extent per Figure 18; and 3) trimming pseudo wells falling south of the southernmost limit of commercial production in the northeast Pennsylvania dry gas window (i.e., Marcellus Line of Death) [139, 140]. The resulting simulation grid included 27,631 pseudo wells with roughly 1 x 1 square mile spacing—where each pseudo well therefore contains a unique value for geologic proxy parameters latitude and longitude.

The two well design simulation scenarios evaluated were under "standard" well design and "tailored" well design configurations. Under the standard well design scenario, a single simulation was conducted at each of the pseudo well locations to generate a Top 12-month production value. Each pseudo well completion design was identical – affixed at the mean value for each of the GBRT model well design predictor parameters from the study dataset (Table 7). Latitude and longitude were allowed to vary based on the specific coordinates of each pseudo well. Therefore, the only production factors varying over the simulation grid relate to spatially-influenced geology variability. Other studies have demonstrated the utility of well locational data in data-driven O&G modeling [64, 59, 15, 55] as an effective strategy for handling spatial disparity driven by geologic heterogeneity over large study domains. The tailored well design scenario was intended to generate improvements in well design component combination for each pseudo well which maximized the Top 12-month production response given the pseudo well's location. Implementing this approach tests the hypotheses that 1) well designs must be custom to the corresponding geologic conditions present to maximize production potential [12, 13, 14, 141] and 2) a ML model capable of generalizing horizontal well behavior can provide rapid insights into such a pursuit. This step includes a brute force optimization approach performed by simulating each pseudo well under 700

randomly determined well design combinations (perforated interval length was not varied). A Monte Carlo Latin hypercube (LHC) sampling approach [142, 143] was used to generate controlled random samples of well design parameters within an operationally practical P_{20} to P_{80} range observed in wells from the study dataset. Perforated interval length was fixed at the dataset mean to compare resulting well productivity against the standard well designs. The well design combination resulting in the highest Top 12-month production estimate for the 700 simulations at each pseudo well was chosen as the best tailored well design and used for further evaluation (data available in Appendix [144]). The input variable ranges for both well design scenarios evaluated are summarized in Table 8.

Table 8. Input variable ranges used for the standard and tailored well design simulation scenarios.

Predictor Parameter	Standard Well Design Value (mean)	Tailored Well Design Range (P_{20} – P_{80})
Perforated interval length (ft)	5,501	5,051 (fixed at mean value)
Water per perforated foot (bbls)	32	22.4 – 41.6
Proppant per perforated foot (lbs)	1,475	1,045 – 1,930
Additive per perforated foot (bbls)	1.54	0.4 – 2.01
Well trajectory azimuth (degrees)	325	314 – 342
Acre spacing (acres)	150	68 - 182
Surface hole latitude (decimal degrees)	Specific to each pseudo well	Specific to each pseudo well
Surface hole longitude (decimal degrees)	Specific to each pseudo well	Specific to each pseudo well

Mean, $P_{20,}$ and P_{80} values used are explicit to the study well dataset for in-field deployed Marcellus Shale wells.

The Top 12-month production responses from both the standardized simulations and tailored well design simulations with LHC sampling are then plotted on maps and contoured as a function of predicted well productivity at each pseudo well location. This approach can highlight the mutability in productivity potential across the Marcellus' regional extent due to the inherent spatially-induced geologic variability, as well as changes in productivity potential from standardized to tailored well designs that maximize production given unique well placement. The

study area is then classified into distinct productivity bins or "grades" from low to high based on threshold cutoffs for the simulated productivity potential (using data generated from tailored well design simulation and LHC sampling only). This binning / grading approach concept has been implemented in various forms from several existing sources as a way to isolate, analyze, and compare different regions of a given play based on productivity [145, 146, 147, 114]. The approach implemented here is unique in that ML modeling served as the basis for estimating productivity potential. Geologic data (described in Section 3.4.4 from well wireline logs was analyzed as a means to evaluate the geologic properties spanning the Marcellus net thickness interval and determine where differences or similarities exist from one productivity bin to another.

3.4.4 Evaluation of well log data

Basic geologic parameters for the Marcellus Shale interval were determined from evaluation of wireline log data. The parameters of interest include bulk density (RHOB in g/cm^3), gamma ray (API), neutron porosity (NPHI in percent, fraction, or decimal), and deep resistivity (Ohm-m) – parameters common to the widely used Triple Combo well log tool string. These basic measurements capture many (but not necessarily all) of the geologic parameters believed to influence shale gas productivity as discussed in Section 3.3.

Well log data, collected typically in half foot resolution, was extracted at two intervals of interest: 1) the gross interval of Marcellus including interbedded limestones like the Cherry Valley; and 2) for the Tully Limestone. Insights from studies conducted by Zagorski et al. (2011), Arthur (2011), and Carter et al., (2011) were used to inform the picks for the Marcellus Shale and Tully Limestone intervals [74, 96, 75]. Gamma ray readings were normalized (referred to as Normalized Gamma Ray from here) using a single-point normalization approach against the Tully Limestone

as a reference for all of the well logs evaluated. Other data processing included removal of data in intervals associated with washout zones. Well log data were used to identify three other parameters of interest at each well: 1) the total vertical depth to the top of the Marcellus (in feet); 2) the gross thickness of the Marcellus (feet); and 3) the net thickness of the Marcellus absent limestone intervals (feet).

For each well log, data across the Marcellus net thickness interval pertaining to the six following geologic attributes were set aside for further analyses: total vertical depth, net thickness, normalized gamma ray, density, neutron porosity, and deep resistivity. Well log data is made available through Appendix [144]. It is worth noting that of the well logs available (51 in total), many were missing log data for one or more of the parameters of interest. Additionally, since the thickness of the Marcellus and interbedded Cherry Valley limestone varies across the play, the volume of data at each well extracted varies depending on: 1) The gross thickness of Marcellus, 2) The logging resolution at a given well and, 3) The availability of each logging string at each well. The following statistical analyses performed described in Section 3.4.5 were therefore typically conducted with highly unequal sample sizes depending on productivity bin(s) evaluated.

3.4.5 Statistical approaches applied

Multiple Kruskal-Wallis (KW) tests were conducted to evaluate the similarity or disparity of both the optimal well design choices (from the LHC sampling described in Section 3.4.3) and geologic properties (from the geologic assessment described in Section 3.4.4) within associated productivity bins. Kruskal-Wallis is a nonparametric statistical technique performed on the ranking order of observations from different datasets [148]. The KW test can be implemented to assess the probability that a random observation from each group are equally likely to be above or below a

random observation from a different group. This test was used because it does not require the evaluated groups to be normally distributed and the test is more stable to outliers. The independent variable evaluated was productivity bin, which included five levels [*A, B, C, D, E*]. The dependent variables are each of the simulation response data for the five well design attributes (with the exception of perforated interval length which was held constant) along with the well log data for the six geologic parameters evaluated. The equality of the five medians for each productivity bin was tested under the following null (H_0) and alternate (H_1) hypotheses list in Equation 3-3:

$$H_0: \tilde{x}_A = \tilde{x}_B = \tilde{x}_C = \tilde{x}_D = \tilde{x}_E$$
$$H_1: \tilde{x}_i \neq \tilde{x}_j \text{ for at least one pair } (i,j)$$

Equation 3-3

The subscripts in Equation 3-3 *i* and *j* correlate to any potential productivity bin levels combination under each well design or geologic property evaluated. The null hypothesis is rejected at a significance level of $\alpha = 0.05$. Kruskal-Wallis tests can provide insights into the overall significance of given factors (i.e., productivity bins) and the corresponding best well design attributes and geologic properties associated with each, but the test cannot inform exactly where differences lie.

Following KW tests, Dunn's tests [149] were used post-hoc to compare the respective pairs of medians for the given well design or geologic properties across each of the five productivity bins where null hypotheses are rejected. The overall significance level was assumed $\alpha = 0.05$ for testing pairwise comparisons for the well design and geologic properties evaluated in this study under the hypotheses presented in Equation 3-4. A Bonferroni adjustment factor was applied for comparisons using Dunn's test [150, 151].

$$H_0: \tilde{x}_i = \tilde{x}_j$$

$$H_1: \tilde{x}_i \neq \tilde{x}_j$$

Equation 3-4

Confidence intervals on the differences in all pairs of medians were also constructed at the 100(1 - α) level (i.e., 95% based on $\alpha = 0.05$). Specifically, Dunn's tests in this application will be able to determine where statistical similarities or differences exist across the best well design choices and geologic properties associated within each productivity bin—an approach that should help differentiate characteristics for controlling gas recovery across varying productivity ranges of the Marcellus. The statistical approaches described in this section are performed using Minitab 18 Statistical Software. A macro developed by Orlich (2010) [152] was used to perform Dunn's Test pairwise comparisons.

3.5 Results and Discussion

This section illustrates how the ensembled ML modeling and well design optimization approach described throughout Section 3.4 has been applied to the Marcellus. The results discussed throughout this section demonstrate that well designs tailored to their specific placement across the Marcellus can improve well productivity compared to standard designs that might be more commonly applied in the field.

3.5.1 Productivity contorting

The maps in Figure 20 show the predicted productivity potential under the standard well design and tailored well design simulation scenarios. The warmer color contours in Figure 20 could

be classified as higher productivity regions and the colder colors are essentially lower productivity regions. Simulation data between scenarios are quantified in Figure 21. Results of these simulations show the expansion of increased potential productivity (i.e., warmer colors) to new regions of the play coupled with a reduction in lower productivity regions (i.e., colder colors) using tailored well designs compared to standard well designs.

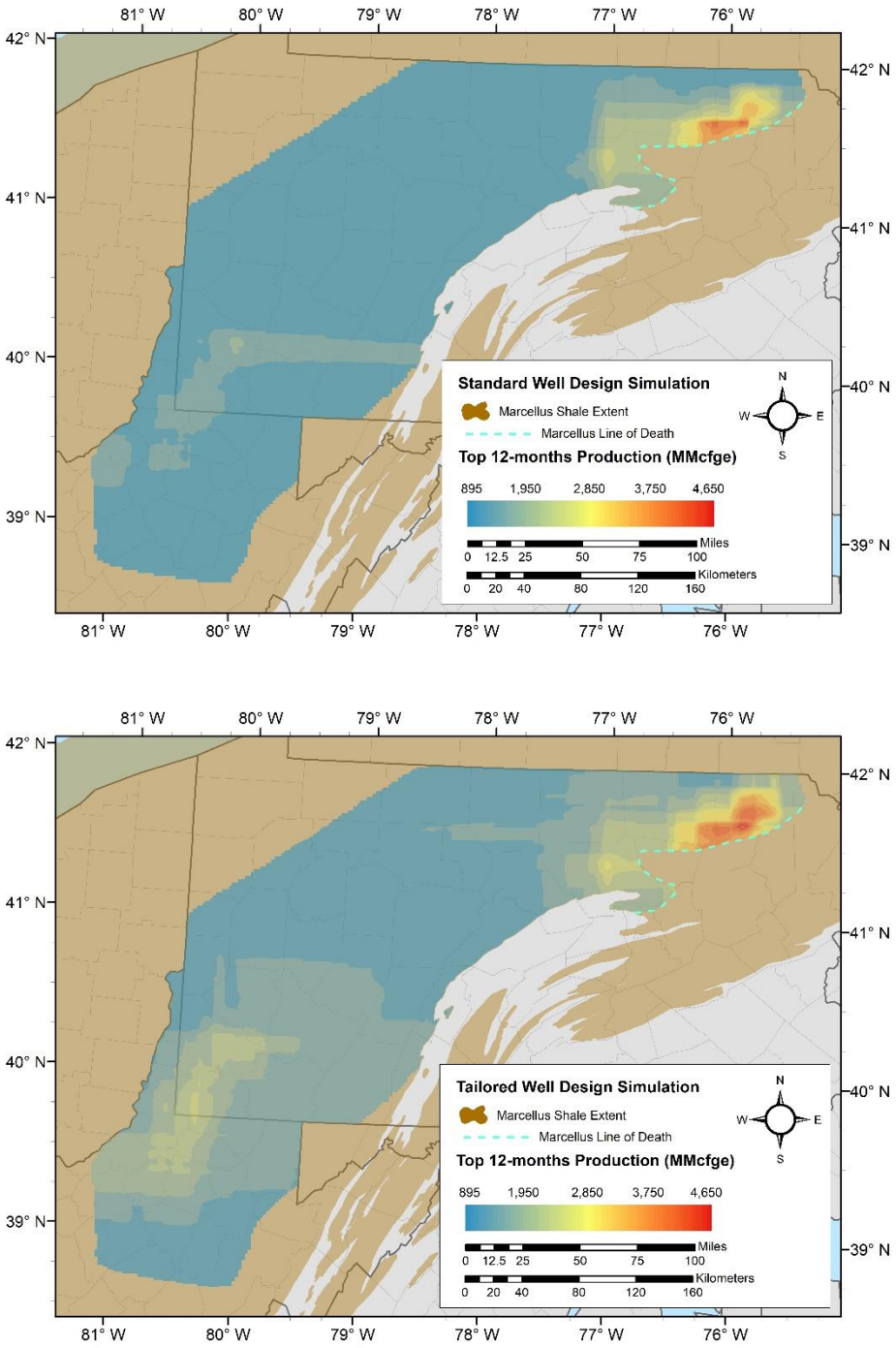

Figure 20. Contour maps depicting simulation results from the standard (top) and tailored (bottom) scenarios. The contours are derived from the Top 12-month production responses generated at pseudo well locations.

The application of tailored wells has resulted in a shift of the marginal mean of the entire pseudo well population Top 12-month production estimate from 1,162 MMcfge under standard well designs to 1,430 MMcfge – an overall increase of roughly 23 percent. The histogram presented in Figure 21A visually depicts the comparison of the pseudo well counts and resulting productivity estimates under each modeling scenario. The entire population of pseudo wells shifts towards larger productivity (x-axis in Figure 21A) when tailored well approaches are applied. Notable improvements in productivity from the standard to tailored well design scenario are realized in several regions across the play per Figure 20 and Figure 21B; particularly in the northeastern core (latitude = 41.75; longitude = -76.6) where substantial improvements are realized (> 1,000 MMcfge) and most of the southwestern region (longitude < -81).

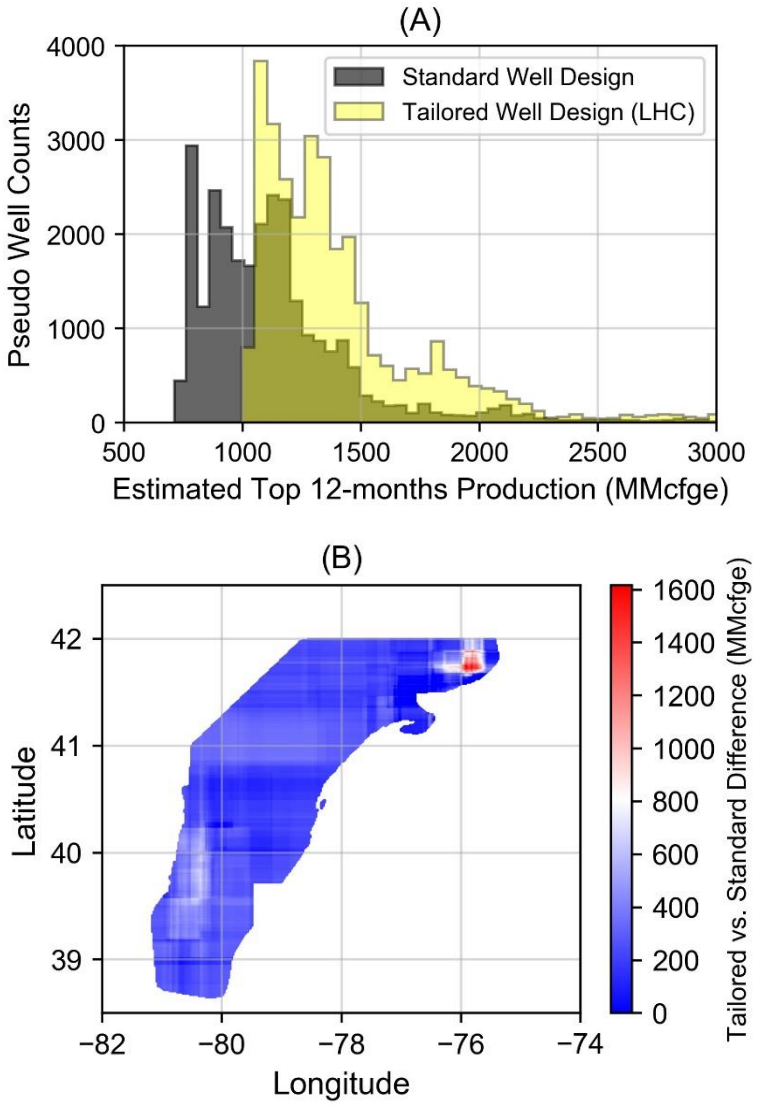

Figure 21. Comparison of productivity outputs from the tailored and standard well design simulation scenario: (A) histogram of pseudo well counts and estimates Top 12-months Production; (B) scatter plot quantifying the difference in Top 12-months production between scenarios.

Our previous study [122] explored the impact of varying only water and proppant intensity and observed the simulated Top 12-month production response for a handful of wells. This work concluded that well design choices pertaining to water and proppant deployed in the field fell short of more optimal combinations that would potentially improve gas productivity at each of their

corresponding well locations. Table 9 below provides a similar comparison of simulated tailored pseudo wells to actual in-field designs at randomly-selected locations within the extent of the study area – however additional well design parameters are adjusted than just water and proppant. In-field wells with similar gross perforated interval lengths (5,501 feet +/- 75 feet), other well design parameters within the P_{20} and P_{80} ranges for those simulated, and roughly proximal to those pseudo wells simulated were conditionally screened for comparison. This analysis provides added perspective on the impact of tailoring wells and their noted performance improvement to not just the modeled standardized designs, but to Marcellus wells recently deployed in the field.

Table 9. Characteristics from the four wells reviewed as part of the in-field vs. tailored well design comparison

Well Parameter	Well Location A Westmoreland County, PA		Well Location B Doddridge County, WV	
	In-field design	Tailored design	In-field design	Tailored design
Water per perforated foot (bbls)	26.8	31	32.3	41.3
Proppant per perforated foot (lbs)	1,200	1,613	1,180	1,786
Additive per perforated foot (bbls)	0.93	0.47	0.98	1.03
Perforated interval length (ft)	5,545	5,501	5,448	5,501
Well trajectory azimuth (degrees)	349	330	304	341
Acre spacing (acres)	110	125	134	175
First production year (year)	2013	NA	2015	NA
Surface hole latitude (decimal degrees)	40.508006	40.50843424	39.265388	39.26388838
Surface hole longitude (decimal degrees)	-79.545961	-79.54867426	-80.596312	-80.59495965
Top 12-months Production (MMcfge)	967	1,343	1,305	1,847
Well Parameter	Well Location C Susquehanna County, PA		Well Location D Bradford County, PA	
	In-field design	Tailored design	In-field design	Tailored design
Water per perforated foot (bbls)	28	41.1	29	38.9
Proppant per perforated foot (lbs)	1,284	1,911	1,311	1,811
Additive per perforated foot (bbls)	1.55	1.35	1.62	0.46
Perforated interval length (ft)	5,477	5,501	5,435	5,501
Well trajectory azimuth (degrees)	351	314	346	331
Acre spacing (acres)	134	113	140	156
First production year (year)	2011	NA	2012	NA
Surface hole latitude (decimal degrees)	41.738956	41.73035199	41.879733	41.87743468
Surface hole longitude (decimal degrees)	-76.029431	-76.02050261	-76.441567	-76.43415032
Top 12-months Production (MMcfge)	2,804	3,040	1,352	1,522

The comparison between the four wells chosen at random indicate that improvements in productivity can occur with subtle variations to in-field well designs. In each well, improvements were noted with increasing both the water injected per perforated foot in addition to proppant per foot intensity. However, additive, acer spacing, and well trajectory had differing effects on the resulting productivity response subject to the specific location of the well.

3.5.2 Grading regions and well log locations

The results from estimating productivity potential using the tailored well design simulation and LHC sampling step were differentiated into five distinct productivity bins or "grades" from high (A) to low (E). The five productivity bins cutoffs were semi-arbitrarily determined at the percentiles outlined in Table 10 with the objective of limiting the top two bins (A and B) at or above the 90^{th} percentile of the simulated pseudo wells from the results in the tailored well simulation scenario. The total count of pseudo wells in the highest productivity bin (Bin A), for instance, includes only the top two percent of the simulation results (in terms of the Top 12-month production indicator) evaluated across the extent of the study area. The second bin (Bin B) includes the segment of the study area from the 90^{th} to 97^{th} percentile productivity range. This grouping convention was intended to isolate the most prolific production regions in the area evaluated for comparison purposes in the following analyses (in Sections 3.5.3 and 3.5.4). The resulting spatial distribution of the subsequent productivity binning is depicted in Figure 22. Bin A is isolated in the Marcellus' northeastern core region. The majority of Bin B is also in the northeast core geographic region and includes a subset in the southwestern core. The other bins are split across the northeast core, southwestern core, and peripheral areas of the study region.

Table 10. Cutoffs of the simulated Top 12-month productivity indicator for the five productivity bins.

Low End Percentile	High End Percentile	Low End Productivity (MMcfge)	High End Productivity (MMcfge)	Bin
0.01	0.30	1,033	1,200	E
0.31	0.60	1,201	1,390	D
0.61	0.89	1,391	1,990	C
0.90	0.97	1,991	3,290	B
0.98	0.99	3,291	>3,882	A

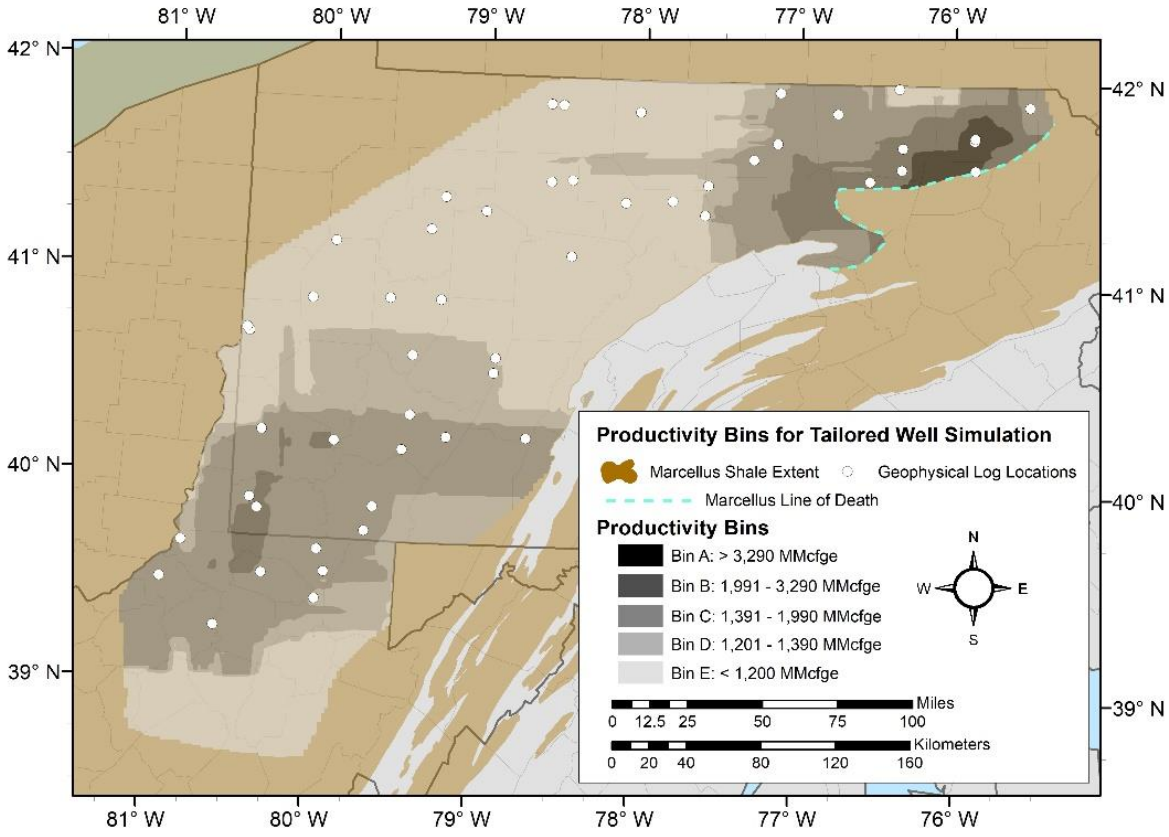

Figure 22. Contour map depicting 1) the extent of each of the five productivity bins based on simulation data from the tailored well simulation scenario and 2) the location of logging data analyzed.

The locations of the 51 geologic well logs analyzed in this study are also plotted in Figure 22. The well log data for the Marcellus interval compiled from these logs were aggregated based on their spatial placement to a corresponding productivity bin. As a result, the following quantity

of well logs available for analysis in each productivity bin were: 3 in Bin A; 4 in Bin B; 19 in Bin C; 4 in Bin D; and 21 in Bin E. Figure 23 depicts a well log interpretation example at depths within and proximal to the Marcellus interval in two different productivity bins. This figure highlights some of the noticeable differences in the geologic properties of the Marcellus from the highest productivity bin compared against one of the lower quality bins. For instance, the notable geologic features with the Marcellus in the Bin A well compared to the Bin D well demonstrate that larger net thickness (275 feet vs. 85 feet) and greater overall average neutron porosity across the Marcellus interval (20 percent vs. 14 percent) exist in this Bin A well example compared to the Bin D well example. These geologic properties, as discussed early in this chapter in Section 3.3, correspond to favorable hydrocarbon producing characteristics for shale reservoirs as documented previous in literature [114, 153, 131, 154, 147]. However, in contrast, the Bin A well example from Figure 23 is lower in other properties favorable to hydrocarbon production—particularly in average gamma ray (269 API vs. 320 API – not normalized in this example), average resistivity (97 Ohm-m vs. 164 Ohm-m), and a pore pressure proxy of true vertical depth (6,500 feet vs. 7,550 feet) compared to the Bin D well. The Marcellus Shale in both wells possess relatively the same density on average (2.54 g/cm^3 vs. 2.51 g/cm^3).

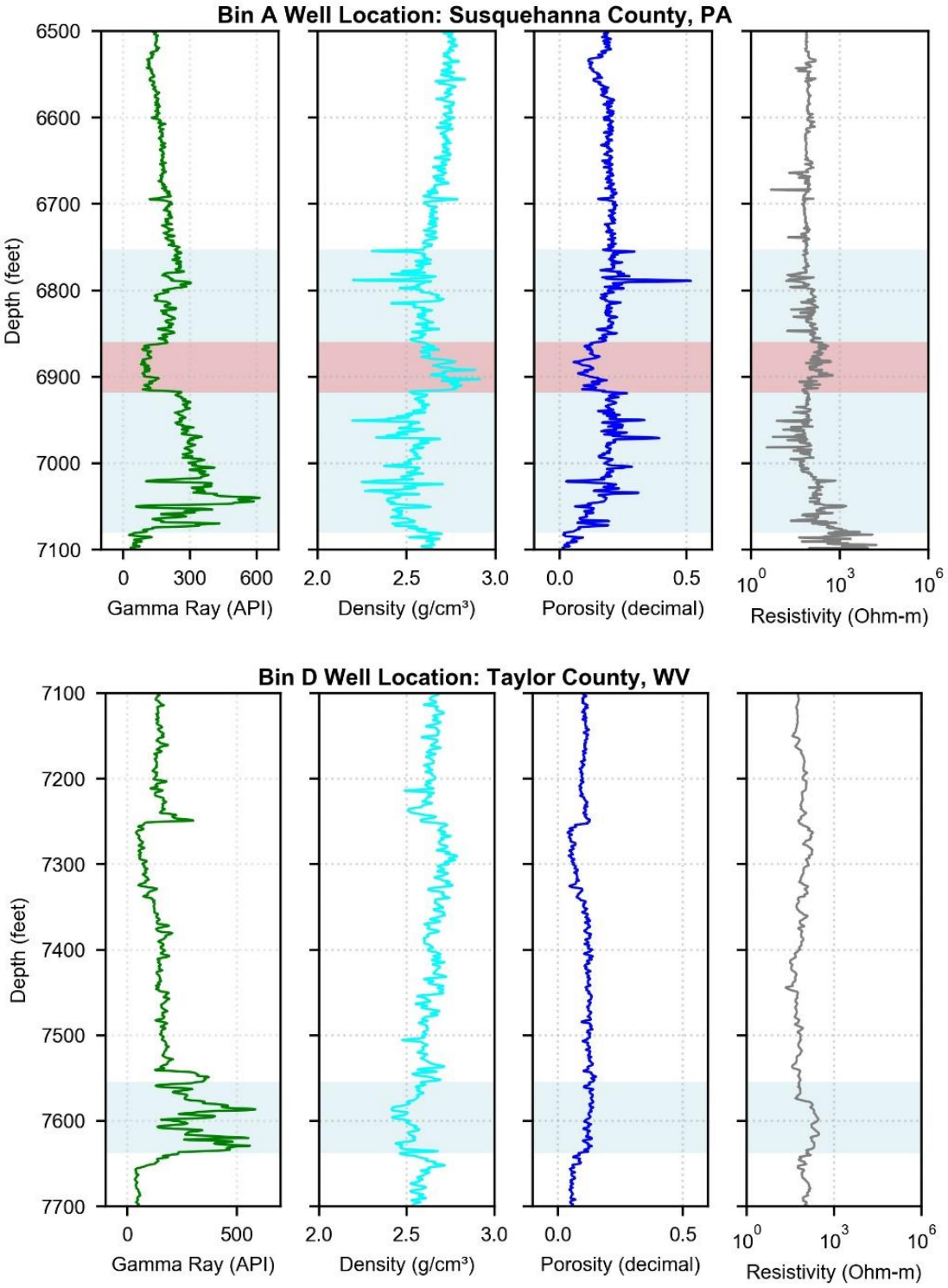

Figure 23. Examples of well log data from two distinct productivity bins. The top is from a well located in Bin A and the bottom is from a well located in Bin D. The net thickness of the Marcellus interval picks are highlighted in blue and Cherry Valley Limestone intervals (where present) are highlighted in light red.

Determination of the productivity bins is partially a result from the resulting GBRT modeling response for every pseudo well evaluated under the LHC sampling scenario. The productivity response at pseudo well location is inherently influenced by both geologic characteristics of the Marcellus (indirectly evaluated by the GBRT model through the use of latitude and longitude coordinates) and the best well design combination that result in the highest productivity. The initial assessment of two well logs from the two productivity bins in Figure 23 suggest that: 1) Higher productivity regions may not comprise of the most favorable geologic characteristics for natural gas potential (on average) for all controlling factors to peripheral, lower productivity regions; 2) A hierarchy of geologic characteristics, or a favorable combination of several characteristics, in terms of productivity potential may therefore exist and, 3) In order to maximize productivity, well design choices will likely need to be fit-for-purpose to specific geologic conditions. Sections 3.5.3 and 3.5.4 evaluate the distributions and statistical significance of several well design criteria and geologic properties resulting from the LHC sampling scenario as they relate to the five productivity bins. These sections provide insight into which properties are likely affecting the productivity of Marcellus wells and which bins have statistically unique well design criteria or geologic properties—information that should help differentiate key drivers to improved productivity and inform tailored well design choices for future wells.

3.5.3 Statistical evaluation of geologic characteristics

The box-and-whisker plots in Figure 24 depict the distribution of geologic parameters of interest in the Marcellus net thickness interval determined from evaluation of wireline log data as they relate to each productivity bin. For each geophysicial parameter, its mean, median, and

relative variance can be compared across productivity bins. A smaller variance for given parameter may suggest more geologic homogeneity exists (at least vertically across the net pay portion of the Marcellus) for that productivity bin, whereas greater variation suggests higher heterogeneity. Therefore, the mean or median values for any particular parameter must be evaluated in context of a given property's distribution disparity across the Marcellus interval when considering productivity potential and fit for purpose well design choices.

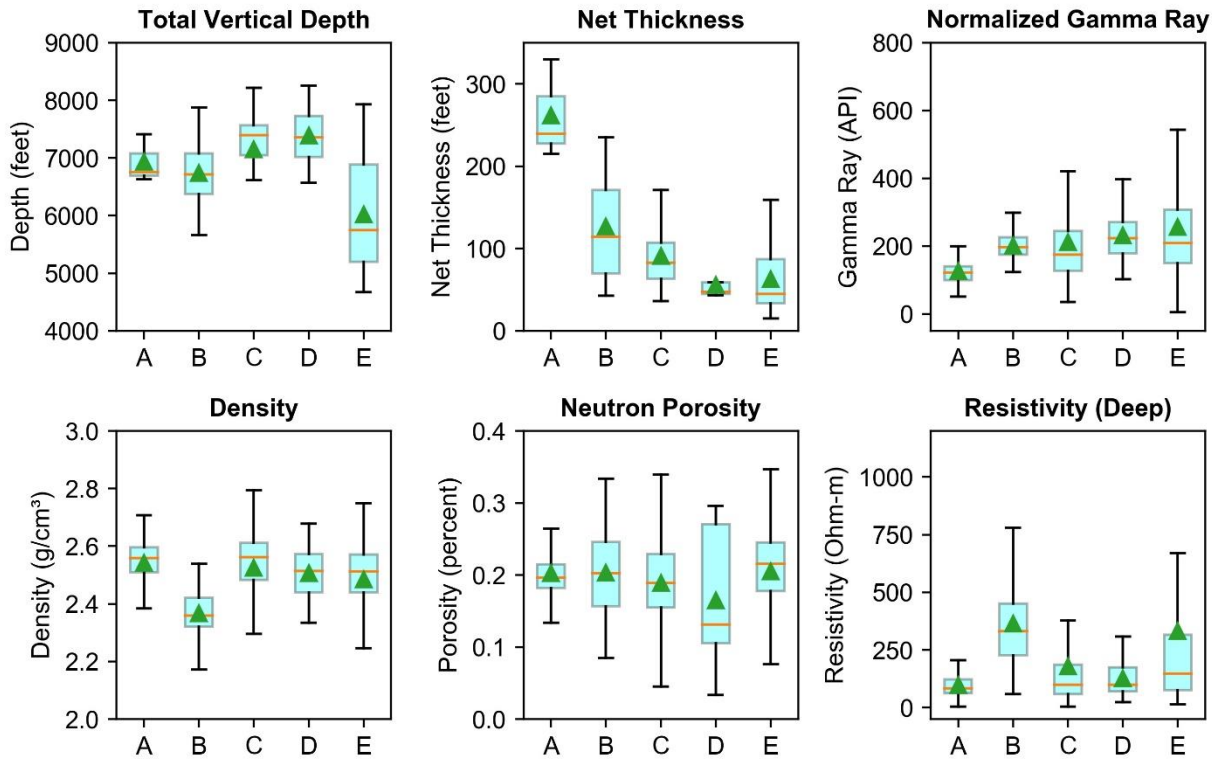

Figure 24. Box-and-whisker plot of geologic properties in the Marcellus net thickness interval for each productivity bin determined through the tailored well design scenario. The box extends from the 25th to 75th quartile values of the data, with a line at the median (50th quartile). The triangle is at the data mean. Whiskers extend to the range of the data at the 10th and 90th quantiles.

The KW results on the geologic data yielded significant variation for all parameters among rock quality ($\alpha = 0.05$). No parameter was determined to be insignificant on the rock quality groupings. Therefore, a Dunn's test was performed for all six geologic parameters within each productivity bin. The post hoc Dunn's test results (Table 11) showed which geologic parameters differed significantly within each rock quailty bin at $\alpha = 0.05$. Mean and median values, standard deviations, and 95% confidence intervals (CI) around the median are also presented for each geologic parameter in Table 11. Property values in Table 11 that do not share a Dunn's test group number [1 through 5] are considered significantly different from each other.

Table 11. Results from Dunn's test on geologic properties across productivity bins

Geologic Attribute	Summary Statistic	Bin A	Bin B	Bin C	Bin D	Bin E
Total vertical depth (feet)	Mean	6,931	6,742	7,146	7,387	6,018
	Median	6,754	6,713	7,397	7,361	5,748
	Standard Deviation	419	908	1,070	706	1,045
	95% CI	(6,630, 7,410)*	(5,665, 7,878)*	(7,050, 7,568)	(6,570, 8,255)*	(5,245, 6,795)
	Dunn's Test Group	1	1	1	1	2
Net thickness (feet)	Mean	261.5	126.6	90.8	55.9	62.9
	Median	240	115	83	48	45
	Standard Deviation	60.6	84.9	39.2	19.6	44.7
	95% CI	(215, 330)*	(43, 235)*	(63, 107)	(44, 85)*	(34, 85)
	Dunn's Test Group	1	2	2	2	2
Normalized gamma ray (API)	Mean	126	201	211	231	256
	Median	122	198	175	224	209
	Standard Deviation	36.9	40.4	140.1	77.5	154.1
	95% CI	(119, 124)	(186, 206)	(171, 179)	(219, 228)	(203, 214)
	Dunn's Test Group	3	1	2	1	1
Density (g/cm^3)	Mean	2.54	2.37	2.52	2.51	2.49
	Median	2.56	2.36	2.56	2.51	2.51
	Standard Deviation	0.086	0.097	0.140	0.079	0.141
	95% CI	(2.55, 2.57)	(2.35, 2.38)	(2.56, 2.57)	(2.50, 2.53)	(2.51, 2.52)
	Dunn's Test Group	1	3	1	2	2
Neutron porosity (percent)	Mean	0.203	0.203	0.189	0.165	0.205
	Median	0.196	0.203	0.189	0.132	0.216
	Standard Deviation	0.046	0.054	0.056	0.089	0.064
	95% CI	(0.195, 0.2)	(0.18, 0.23)	(0.187, 0.191)	(0.128, 0.135)	(0.214, 0.219)
	Dunn's Test Group	1,2	1	3	4	1
Deep resistivity (Ohm-m)	Mean	96.8	364.6	178.1	126.3	330.7
	Median	83.8	331.6	99	99	148.2
	Standard Deviation	62.4	177.6	371.3	67.7	664.9
	95% CI	(80.1, 88.3)	(310, 394)	(93.9, 102.4)	(90.5, 117.8))	(141.1, 160.4)
	Dunn's Test Group	4	1	3	3	2

* Indicates the CI's achieved based on data available. CI's are lower than 95% as a result; Bin A = 75% CI and Bin B = 88%. CI

The Dunn's test group number orders are based on the resulting highest group population value for that given parameter relative to the other Dunn's test groups. The resulting groups determined from the Dunn's test indicate that the higher productivity bins do not necessarily contain the most favorable geologic conditions under all circumstances. For instance, the highest productivity bin (Bin A) is in the 1st Dunn's test group for total vertical depth, net thickness, density, and neutron porosity; Bin A also contains lower standard deviations for each property relative to other bins, suggesting higher homogeneity may exist. However, Bin A is in the lowest groups for deep resistivity (group 4) and normalized gamma ray (group 3). This circumstance suggests that one or more of the geologic parameters pertaining to depth, thickness, or neutron porosity are contributing towards Bin A being the highest in terms of productivity potential (agnostic to pseudo well designs deployed) relative to the other bins despite the comparative shortfalls in normalized gamma ray (likely a function of Bin A being predominantly located in the overly thermal mature northeastern core area) and resistivity. In contrast, productivity bin E, the lowest productivity bin, is in Dunn's test group 1 for neutron porosity and normalized gamma ray – two of the properties strongly tied to in-place hydrocarbons. However, Bin E is relatively thinner and shallower than the other groups, and typically has a higher standard deviation across geologic properties compared to other bins, suggesting that larger relative heterogeneity exists. These findings may suggest that the actual rock quality in Bin E could be, in many instances, relatively high compared to the other groups. But because Bin E is thinner compared to other bins, there might not be as much resource in place from a volumetric quantitative perspective. In terms of exploration, the Bin E extent is quite large and less developed than core areas of the Marcellus; but could, perhaps, contain potentially high-productivity regions in isolation worthy of future prospecting efforts.

The noted disparity in geologic property medians and standard deviations across productivity bins indicates that there are various combinations of geologic drivers at play influencing productivity potential. Given those dissimilarities, well completion strategies would likely require targeted designs to specific geologic circumstances in order maximize productivity. In the prior discussion example, the best well completion approaches for Bin A rock, which is thick, deep, dense, and homogenous relative to other groups, would likely differ from thinner, more heterogenous Bin E rock. The resulting data for the best wellbore designs for pseudo wells in each productivity bin determined through LHC sampling are evaluated in Section 3.5.4.

3.5.4 Statistical evaluation of well design attributes

The box-and-whisker plots in Figure 25 depicts the distribution of the resulting aggregation of well design attributes that generated the highest Top 12-month production estimate at each pseudo well following LHC sampling and simulation as they relate to each productivity bin. The resulting Top 12-month production estimate from the pseudo wells that correspond to each bin are also presented for comparison across bins within Figure 25. For each well design attribute, its mean, median, and relative variance can be compared across productivity bins, highlighting: 1) ranges of the most favorable attribute setting that maximized production for pseudo wells common to each bin; 2) the intra-bin variability of a given attribute that results in the most productive pseudo wells; and 3) comparison and contrast of the most productive well design attribute combinations across bins. Figure 25 highlights the disparity and extent of the range in optimal well design attribute settings across bins that maximize the GBRT productivity response – a finding that emphasizes that the most prudent utilization of hydrocarbon resources is likely attained though fit for purpose well designs.

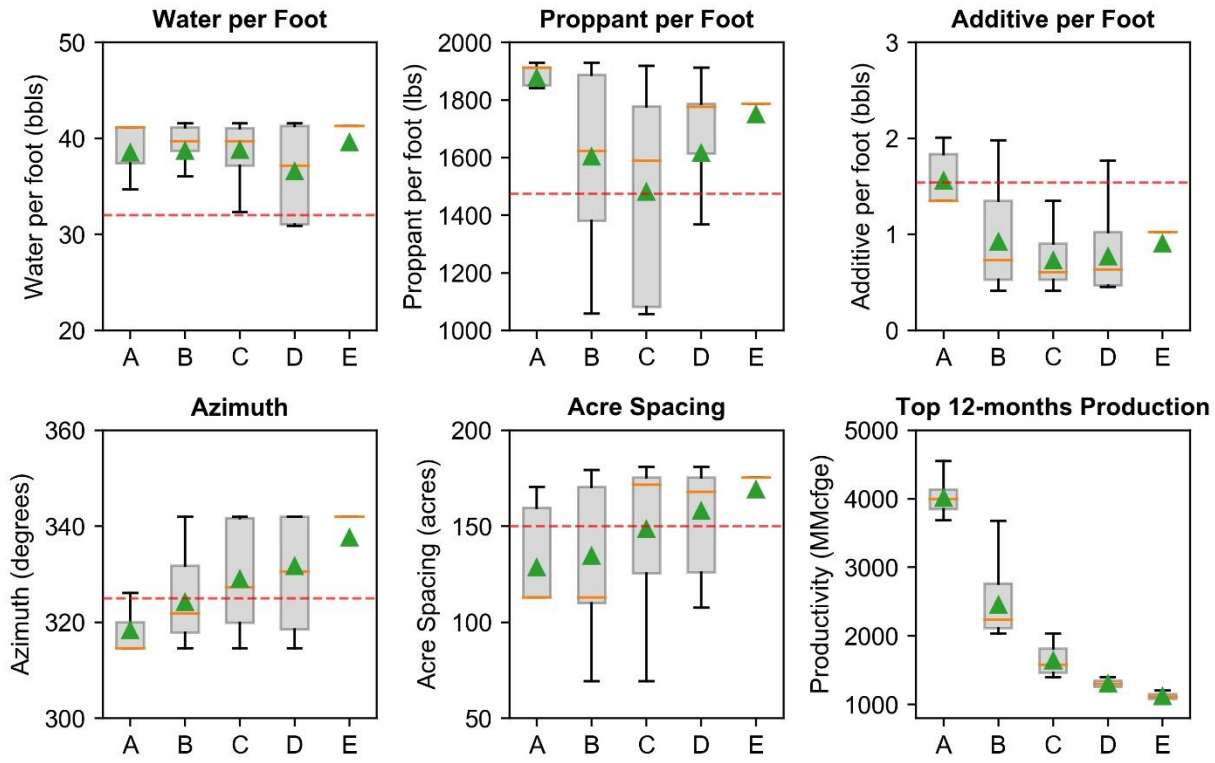

Figure 25. Box-and-whisker plot of best well design attributes across each productivity bin determined through the tailored well design scenario.[7] The box extends from the 25th to 75th quartile values of the data, with a line at the median (50th quartile). The triangle is at the data mean. Whiskers extend to the range of the data at the 10th and 90th quantiles. The red dashed line represents the mean values for each design attribute under the standard well design scenario.

The dashed red line in Figure 25 provides a reference point to the attribute settings used under the standard well modeling scenario for each parameter evaluated. Results presented in Section 3.5.1 (Figure 20 and Figure 21) comparing the standard and tailored well design indicated

[7] The strong convergence and minimal variance for well design settings noted in Bin E is attributed to the smaller relative contribution of real-world training data (Figure 2) given Bin E's regional extent (Figure 6) in the periphery of the Marcellus Shale producing area.

that higher well productivity was attained when using well design configurations determined through the tailored well design scenario. The tailored well design scenario was also found to potentially improve upon well performance for in-field well designs with similar perforated interval lengths and in proximity as those modeled through pseudo wells (Table 9). The results here demonstrate that optimized well design parameter settings can vary substantially relative to the dataset mean values used in the standard well design scenario depending on well placement within the extent of the area evaluated (i.e., productivity bin). For two parameters, water per foot and proppant per foot, the best attribute settings (referencing the mean and median values) are found at greater levels at all productivity bins relative to the standard well design scenario. However, substantial variation exists in the optimal proppant per foot settings in Bins B (northeast core and subset of southwest core) and C (fringe of northeast core and majority of southwestern core); possibly attributed to the extensive regional extent and associated geologic heterogeneity of those bins. The sensitivity of both water volumes and proppant intensity on resulting productivity are consistent with findings from our previous work using case studies to explore optimized well design choices compared to in-field designs [122]. Conversely, higher productivity typically occurs when additive per foot is set lower than the dataset means for all bins. Bins B through E settings are relatively consistent at below one bbl per foot. However, Bin A (northeast core) is shown to be more productive when more additive per foot is implemented, near 1.55 bbl per foot on average.

In the two other cases, acre spacing and azimuth, the optimal settings straddle the dataset mean with Bins A and B optimized at settings typically below the dataset mean, and Bins C, D and E near or above the mean. Resulting wellbore azimuth trajectory ranges from modeling appear relatively consistent to being perpendicular to the overall present-day tectonic stress field and J1

joint sets common to Devonian-aged reservoirs within the Appalachian Basin – a well design practice believed to maximize productivity in unconventional horizontal wells [155, 75, 156]. Present-day stress orientation may vary across the basin; therefore, it is expected that the optimal design attribute for each resulting productivity bin would vary accordingly given the resulting spatial constraints of the bin. Bin A and B that have a small geographically constrained footprint have a smaller ideal wellbore azimuth standard deviation; while bins C and D that range across both the north and south region of the study area have a large standard deviation of ideal wellbore azimuth. In terms of acre spacing, Bins A and B are showing that production in down-spaced wells relative to the dataset mean is not negatively impacted, whereas Bins C, D, and E are optimized when well spacing is near or above the dataset mean. Bins A and B predominantly cover the northeastern core of the Marcellus play. Operators prominent in that region have demonstrated that down-spacing wells has indeed not negatively impacted well productivity [157] – a notion that better utilization of the nation's gas asset can be developed more prudently and higher "play-level" recovery factors achieved by potentially down-spacing wells in those regions. Additionally, less child/parent well interactions may therefore be occurring and well completion upscaling (increased proppant and water use) may lead to productivity improvements without detrimental effects on neighboring wells. This is likely because the thicker Marcellus interval in the northeast core offers more vertical "stacked" targets per acre, at least for Bin A acreage. On the other hand, an operator predominantly active in the southwestern core (prominent in portions of the extent of Bins C, D, and E) has reported reduced normalized performance in down-spaced wells [158]. The southwestern core is more condensed (thinner with net-to-gross closer to one) than the northeast core, and the vertical stacking strategy is less applicable. Therefore, well design considerations,

particularly regarding overall hydrocarbon recovery at the field scale vs. individual well level, are much different in the southwest core compared to the northeast.

The KW findings on the resulting optimal well design attributes for each pseudo well yielded significant variation for all attributes with rock quality ($\alpha = 0.05$). No parameter was determined to be insignificant on the rock quality groupings. Therefore, a Dunn's test was performed on all five well design attributes within each productivity bin. The post hoc Dunn's test (Table 12) showed which well design attributes differed significantly within each productivity bin at $\alpha = 0.05$. Mean and median values, standard deviations, and 95% CI's around the medians are also presented for each well design attribute. Property values in Table 12 that do not share a Dunn's test Group number are considered significantly different from each other. Table 12 also features the number of pseudo wells within each productivity bin.

Table 12. Results from Dunn's test on best well design attributes from LHC sampling across productivity bins.

Well Design Attribute	Summary Statistic	Bin A	Bin B	Bin C	Bin D	Bin E
Pseudo wells	Count (*N*)	282	1,492	8,378	8,219	9,259
Proppant per foot (lbs)	Mean	1,877	1,602	1,481	1,616	1,749
	Median	1,912	1,622	1,590	1,777	1,787
	Standard Deviation	84	244	296	267	131
	95% CI	(1,912, 1,912)	(1,517, 1,656)	(1,569, 1,613)	(1,777, 1,777)	(1,787, 1,787)
	Dunn's Test Group	1	3	5	4	2
Water per foot (bbls)	Mean	38.5	38.7	38.8	36.6	39.6
	Median	41.1	39.7	39.7	37.2	41.3
	Standard Deviation	3.82	3.65	2.83	4.45	3.67
	95% CI	(41.1, 41.1)	(39.7, 39.7)	(39.7, 39.7)	(37.1, 37.4)	(41.3, 41.3)
	Dunn's Test Group	2,3	4	2	3	1
Additive per foot (bbls)	Mean	1.56	0.92	0.73	0.77	0.90
	Median	1.35	0.73	0.61	0.64	1.03
	Standard Deviation	0.26	0.49	0.29	0.32	0.23
	95% CI	(1.35,1.35)	(0.73, 0.73)	(0.61, 0.64)	(0.61, 0.9)	(1.03, 1.03)
	Dunn's Test Group	1	2	4	3	2
Well bore azimuth (degrees)	Mean	318	324	328	331	338
	Median	315	322	327	331	342
	Standard Deviation	5.9	8.5	9.7	10.4	8.5
	95% CI	(315, 315)	(322, 322)	(327, 327)	(331, 331)	(342, 342)
	Dunn's Test Group	5	4	3	2	1
Acre spacing (acres)	Mean	128	134	149	158	169
	Median	113	113	171	168	175
	Standard Deviation	21.2	31.5	35.5	21.2	15.2
	95% CI	(113, 113)	(113, 117)	(171, 171)	(168, 169)	(175, 175)
	Dunn's Test Group	4	3	2	2	1

Results indicate that wells in Bin A, on average, receive the most intensive designs (most proppant, second most water, most additive, and smallest spacing) relative to wells in the other bins. Despite the higher intensity completion settings, wells in Bin A can also handle tightener acre well spacing placement.

3.5.5 Reduced order multivariate predictive model

Since the GBRT machine learning model can make predictions at unique latitude and longitude coordinates given certain well design inputs, production estimates can be generated at locations where geologic properties are known. Then, the collection of the production prediction

data, well design parameters, and geologic properties can be correlated into a model that leverages actual geologic data as inputs. A reduced order predictive model was developed using a multiple linear regression approach which can estimate production at new Marcellus horizontal well sites without having to employ the GBRT machine learning model by the user. Additionally, this approach can couple well design and geologic data into a single analytical method so that they can be evaluated in tandem. The reduced order model is intended to be as simplistic as possible in its formulation and therefore easily interpretable by end-users; but must also be accurate in its generalization of estimating gas productivity. The model formulation is presented in Equation 3-5 where y_i denotes the Top 12-months production (MMcfge) at a given horizontal well (i) used in model fitting. The model was fit using a total of 16 input parameters related to well design and well log data geology attributes. The model formulation was fit using Least Squares to obtain parameter estimates (β) for each input parameter considered. The intercept (β_0) was assumed zero and the error (ε) assumed independent and normally distributed.

$$y_i = \beta_0 + \sum_{x=1}^{16} \{\beta_1 x_1 + \beta_2 x_2 \dots \beta_{16} x_{16}\} + \varepsilon \qquad \text{Equation 3-5}$$

A new training dataset was compiled by generating production estimates at specific spatial coordinates of the well logs available as part of this study. A similar LHC resampling approach to the one described in Section 3.3 was conducted here on the same well design input parameters and ranges (Table 8), however, gross perforated interval was also included and evaluated between 3,320 to 6,544 feet in length. The other 10 input parameters relating to geologic properties (which included mean values for the Marcellus net thickness interval, alongside the standard deviation of each property) were held constant based on observed data representative at each well log location.

A total of 31 of the available 51 well logs contain data for each of the six geologic properties evaluated. Data from those 31 well logs were used as part of the linear model development. LHS resampling using different well design choices were conducted at each of the 31 well log locations 1,000 times. A production estimate was generated using the GBRT model for each sample as the response. As a result, 31,000 specific training realizations were used for model fitting (data available through Appendix [144]). The model was fit as purely additive (no interactions considered) and under various polynomial orders in order to maintain simplistic formulation. A 3rd order polynomial result in the best fit.

The fitted model has a squared correlation coefficient is 0.733 and RMSE is 241, meaning the trained model should be reliable for production prediction in areas across the Marcellus where the specific geologic properties needed as inputs are known. This reduced order model is intended to provide a simplistic and efficient option for evaluating potential well performance founded on the specific design choices given known geologic conditions. The model should be helpful in informing future well design choices given access to relatively common geologic data. Additionally, its linear formulation also enables potential well design parameter optimization.

The resulting reduced model equation is segmented by three brackets in Equation 3-6 – the first of which includes parameters related to well design choices, the second to the mean values for the geologic properties evaluated, and the third related to the heterogeneity (in terms of standard deviation) of the corresponding geologic properties across the Marcellus interval.

$$\begin{aligned}\hat{y} = &\left[2.84\times10^{-9}{L_P}^3 + 3.55\times10^{-8}P^3 + 5.26\times10^{-3}W^3 + 2.15\times10^{-5}{A_2}^3 + 2.05\times10^{-5}{A_3}^3 - 4.66{A_1}^3\right] \\ &+ \left[1.37\times10^{-9}{D_1}^3 + 1.17\times10^{-4}{h_s}^3 + 8.03\times10^{-3}\emptyset^3 + 7.04\times10^{-9}{R_{deep}}^3 - 6.59\times10^{-7}G^3 - 66.3{D_2}^3\right] \\ &+ \left[5.75\times10^{-9}{R_{Stdev}}^3 + 1.57\times10^{-5}{G_{Stdev}}^3 + 12.5{\emptyset_{Stdev}}^3 - 2.61\times10^{-2}{D_{2Stdev}}^3\right]\end{aligned} \quad \text{Equation 3-6}$$

Where:

$\hat{y}$= Estimation of Top 12-months productivity indicator (MMcfge) for a new well
L_P = Gross perforated interval length (feet)
P= Proppant per foot (lbs / foot)
W= Water per foot (bbls / foot)
A_1= Additive per foot (bbls / foot)
A_2= Wellbore azimuth trajectory (degrees)
A_3= Acre spacing (acres)
D_1= Top of Marcellus Shale depth (feet below ground surface)
h_s= Marcellus net thickness (feet [net limestone units])
G= Normalized average Marcellus Shale gamma ray (API [normalized per Section 3.4.4])
D_2= Average Marcellus Shale density (g/cm^3)
ϕ= Average Marcellus Shale porosity (fraction [based on neutron porosity])
R_{deep}= Average Marcellus Shale deep resistivity (Ohm-m [via deep induction logging])
R_{stdev}= Marcellus deep resistivity standard deviation (Ohm-m)
G_{stdev}= Marcellus normalized gamma ray standard deviation (API)
ϕ_{stdev}= Marcellus porosity standard deviation (fraction)
D_{2stdev}= Marcellus density standard deviation (g/cm^3)

All model β coefficients are statistically significant ($p < 0.05$) in influencing the response variable $\hat{y}$, with the exception of Marcellus deep resistivity standard deviation (R_{stdev}). Model residuals were evaluated and are normally distributed (Appendix [144]). The model estimates are only relevant when input parameter ranges for well design features fall within those used to fit the model presented in Table 8 and for gross perforated intervals between 3,320 to 6,544 feet in length.

The positive or negative sign associated with the beta coefficients for each term in Equation 3-6 can suggest each parameters' impact on productivity. For instance, positive beta coefficients and associated terms improve the productivity response, whereas negative terms are potentially detrimental to the productivity. Data used to fit the linear model here were not standardized [159] so that input data remains in the units commonly attained in the field and are inherently in different scales—therefore the magnitude of each beta coefficient cannot be compared in pairwise fashion from one term to another to assess each parameter's influence. Notice the positive impact of the beta coefficients for geologic properties associated with depth, resistivity, porosity, and thickness;

the negative beta coefficients correspond to gamma ray, density, and the standard deviation term associated with density. This outcome coincides with the statistical findings regarding the standard deviation and confidence intervals of geologic properties and associated productivity bins (Figure 24 and Table 11) discussed in Section 3.5.3, along with the notion that high in-reservoir vertical heterogeneity (inferred from the Marcellus standard deviation terms) has a prominent impact on the overall hydraulic fracturing process, fracture propagation, and subsequent well productivity [160, 161, 162]. A potential contradiction here is the negative effect of measured normalized gamma ray to heuristic belief on gas/oil availability. The likely reason is the noticeable low gamma ray associated with the overly thermally mature Marcellus northeast core (Bin A and portions of Bin B) despite being a high productivity region.

3.6 Conclusions and Outlook

In this chapter, we have introduced a framework that ensembles a data-driven predictive model that estimates well-level productivity with a completion design optimization approach aimed at improving well production potential. This analytical approach developed leverages readily available datasets provide to be a fast and cost-effective evaluation tool that could complement existing reservoir management practices. The framework has been tested and results discussed when applied to the producing extent of the Marcellus. The insights gained through this work should be advantageous in both the 1) identification of high-priority drilling regions based on productivity potential as well as 2) informing the tailoring of future well designs given their placement in the Marcellus. These aspects are resolved through spatial ranking of regional productivity potential and demarcation via contour mapping, as well as identification of optimal

well design configurations within each ranked region driven by controlling geologic conditions. Simulation results from the LHC sampling and brute force well optimization approach show the expansion of increased potential productivity across the study area when wells were tailored based on placement as compared to standard well designs.

The framework developed does not consider infrastructure development issues or limitations, the impact of well design choices on costs, nor are larger macroeconomic drivers which may impact natural gas demand reflected. Therefore, the authors suggest that such a framework serve as compliment to reservoir and field management strategies moving forward – not a replacement for current approaches. Additional work is needed to integrate geological data in a reliable fashion on a larger scale into the front end (i.e., ML development component) of the modeling framework. This addition may expand the framework utility in evaluating geologic properties and well design attributes (as well as their interactions) that make up shale production controlling factors.

While this approach was applied across the Marcellus producing region, it could be modified and adapted to a more focused scale; either in the Appalachian Basin or elsewhere. The types of data sets used are likely commonplace for different gas or oil reservoirs and may be obtained from public sources. Furthermore, the study area was grouped based on simulated productivity bins, but the same framework could be applied on a different grouping configuration (i.e., regional delineations, geologic conditions delineation, political boundaries [state or county], etc.) and reevaluated statistically in a similar manner.

4.0 Application of a Deep Learning Network for Joint Prediction of Associated Fluid Production in Unconventional Hydrocarbon Development

4.1 Chapter Summary

Machine learning (ML) approaches have risen in popularity for use in many oil and gas (O&G) applications. Time series-based predictive forecasting of hydrocarbon production using deep learning ML strategies that can generalize temporal or sequence-based information within data has become an impactful research topic. Recent emphasis on hydrocarbon production provides opportunities to explore the use of deep learning ML to other facets of O&G development where dynamic, temporal dependencies exist and that also hold implications to production forecasting. This study proposes a combination of supervised and unsupervised ML approaches as part of a framework for the joint prediction of produced water and natural gas volumes associated with oil production from unconventional reservoirs in a time series fashion. The study focuses on the pay zones within the Spraberry and Wolfcamp Formations of the Midland Basin in the U.S. The joint prediction model is based on a deep neural network architecture leveraging long short-term memory (LSTM) layers. Our model is capable to both reproduce and forecast produced water and natural gas volumes for wells at monthly resolution and has demonstrated 90 percent joint prediction accuracy to held out testing data with little disparity noted in prediction performance between the training and test datasets. Additionally, model predictions replicate water and gas production profiles to wells in the test dataset, even for circumstances that include irregularities in production trends. We apply the model in tandem with an Arps decline model to generate cumulative first and five-year estimates for oil, gas, and water production outlooks at the well and

basin-levels. Production outlook totals are influenced by well completion, decline curve, and spatial and reservoir attributes. These types of model-derived outlooks can aid operators in formulating management or remedial solutions for the volumes of fluids expected from unconventional O&G development.

4.2 Introduction

The continued pursuit for reliable, affordable, and secure supplies of energy accentuates the necessity for continued research into ways to economically and efficiently access the vast amount unconventional natural gas and oil resources that exist. Over the last decade and a half, the application horizontal drilling techniques coupled with advanced, multi-stage hydraulic fracturing technologies has facilitated the widespread development of unconventional oil and gas (O&G) reservoirs (such as shale and tight oil reserves) [2]; resulting in a revolution in the energy landscape [3, 4, 5], particularly in the United States (U.S.).

Hydraulic fracturing methods make use of injected liquids under high pressure to generate breakages in subsurface formations and are usually implemented where low permeability conditions exist. The fracturing fluid is composed of a base fluid, typically water, constituting >98 percent of the total fluid volume [163] and the remaining contribution coming from proppant and chemical additives. The goal of the hydraulic fracturing process is to promote the generation of new fractures in the tight hydrocarbon-bearing rock formations inherently low in both permeability and porosity while simultaneously augmenting the size, magnitude, and connectivity of existing fractures to stimulate oil and/or gas flow to wells [164, 165, 166]. Once the hydraulic fracturing process is completed, the high in situ pressures within the reservoir as compared to the lower

bottomhole pressure in the wellbore (which can be managed via artificial lifting) prompts fluids to migrate towards the well and be produced at the surface. The fluid that returns to the surface may contain a combination of hydrocarbons (oil and/or gas) and water, in addition to injected chemical additives from the hydraulic fracturing process, as well as naturally occurring materials like brines, metals, and radioactive materials [167]. Each constituent requires some form of management, depending heavily on the intended endues of each, which may include sale to market as a commodity, reuse as part of site operations, or treatment and disposal.

Horizontal wells drilled and completed in shale gas and tight oil formations make up the preponderance of hydrocarbon production in the United States. Specifically, crude oil production from tight formations alone reached 6.5 million barrels per day in the U.S. through 2018, accounting for 61 percent of the total oil produced in the U.S. The U.S. Energy Information Administration (EIA) indicates that use of horizontal wells accounted for 96 percent of the overall U.S. crude oil production from tight formations by the end of 2018 [97]. A recent surge in the development of tight oil reserves located in the Permian Basin in western Texas and eastern New Mexico (41 percent of total tight oil production in the U.S. in 2018) has led to considerable growth in overall U.S. crude oil production [168].

While unconventional oil (and gas) resources remain critically important in the pursuit towards energy security, challenges persist in effectively forecasting their production potential. For instance, productivity in unconventional reservoirs is known to be responsive to the nature and effectiveness of the interactions between wellbore design, completion and stimulation processes and the inherent irregularities in reservoir conditions. As a result, fluid production responses can be highly disparate across: 1) An entire O&G play [169]; 2) wells on a given pad targeting the same formation; or even 3) the different perforation stages of single well's lateral component

[170]. Production forecasts hold implications on the strategic decisions made by the O&G sector. For instance, resulting production outlooks, depending on the long-term trajectories of fluid volumes produced, can prompt macro-scale consequences like potential fluctuations in oil and/or gas market prices and associated impacts on the environment [171]. Additionally, forecasts can influence micro-scale outcomes that ultimately shape a wide range of operating and maintenance scenarios for field operators or even effect company profit margins. Reservoir modeling and simulation are commonly used to inform decision makers regarding the potential production response and long-term performance of hydraulically fractured horizontal wells in unconventional reservoirs. These approaches can be costly in terms of the time and computational resources needed to execute effectively [47, 105]. Furthermore, difficulties exist in attaining sufficient levels of geological data at the well level [51] to sufficiently reflect the diversity in reservoir conditions needed to model fluid flow. This challenge intensifies when the interest spans to multi-well performance evaluation at the field-scale or larger.

Given the computational resources that are typically widely available and the emergence of O&G digital datasets that include features associated with well completion, stimulation, and production, many have taken to machine learning (ML) and data analytics as a compliment to existing approaches for O&G production analysis [40, 32, 38]. ML-based tactics can provide additional analytical functionality to traditional reservoir simulation methods. They have proven effective in accurately and reliably modeling circumstances involving highly complex systems where variable conditions are known to be prominent, not uncommon to wellbore/reservoir relationship interactions in unconventional O&G development. Additionally, they offer expeditious predictive capability, allowing users to quickly generate multiple realizations thereby enabling greater insight into the systems modeled [41].

A number of potential use cases exist where ML has been applied as part evaluating the effects of hydraulic fracturing designs on hydrocarbon production in unconventional reservoirs. As an example, several studies utilize static productivity indicators that reflects cumulative production under a fixed time duration (i.e., six months or one year) as response variables [15, 56, 172, 122, 55] to evaluate potential well response to various hydrofracking completion designs. The use of static response variables enabled straightforward evaluation of input feature impact rating and ranking, as well as sensitivity evaluation. The findings from these studies have proven insightful in identifying key production drivers representative to the study areas evaluated, as well as effective in approximating well productivity given well completion design choices. However, they are not directly translatable to well history matching, production forecasting, and facilitating data-driven production outlook scenarios [59, 60, 61, 10].

Many studies are taking focus on using ML for dynamic reservoir analysis by evaluating time series-based topics, like oil or gas production over the life of producing wells. These studies are leveraging empirical data that includes daily or monthly cumulative hydrocarbon production values over all or a portion of each well's productive life. Many of the relevant studies apply deep learning ML strategies in order to capture and generalize the intrinsic temporal or time sequence-based properties within the data. Findings from recent studies indicate that the deep learning approaches applied have been exceedingly effective at predicting dynamic production trends accurately on holdout data. The results of which suggests that these approaches hold substantial implications and potential viability in production forecasting.

To gain further comprehension on O&G-related time series analysis using ML, we provide a short review of relevant studies works that have focused on this topic. A study by Jie et al. developed two deep learning models to predict daily gas production from a single well completed

in the Sichuan basin in China [173]. The researchers developed artificial neural network- (ANN) based models using: 1) A fully-connected multilayer perceptron-based ANN with a single hidden layer and 2) a long-short term memory- (LSTM) based ANN with stacked LSTM layer architecture. Empirical data for daily gas production over a three-year period was used for analysis. The first 900 dataset observations were used for model training while and the last 100 observations were used for holdout model performance testing. Input data included the data features (assumed at daily resolution) of oil pressure, casing pressure, daily water production, cumulative gas production, cumulative water production, and water-gas ratio. Results indicated prediction error of 1.56 percent for the LSTM-based model and upwards of 9.66 percent for the MLP-ANN. Sagheer and Kotb implemented deep LSTM architectures to estimate monthly oil production for two oil fields; one was the Tarapur Block of Cambay Basin to the west of Cambay Gas Field in India and the other in the Huabei oilfield in China [174]. They demonstrate the predictive effectiveness in stacking LSTM layers as part of network architecture when long interval temporal dependencies may exist as compared to model performance when shallow neural network architectures are used. Additionally, the researchers noted that their LSTM-based model outperformed counterpart formulations explored that were based on deep recurrent neural networks (RNN) and Deep Gated Recurrent Unit models. The work performed by Huaibei Liu et al. included the development of an ensemble empirical mode decomposition (EEMD) based LSTM learning network capable of time series forecasting of oil production. Case studies were performed to empirical field from the SL and JD oilfields, China [175]. Their proposed EEMD-LSTM configuration outperformed other model types developed under ensembles between EEMD and MLP-based artificial neural networks and EEMD with support vector machine.

Collectively, these studies demonstrate the utility and capability of deep learning-based ML (with noted effectiveness of LSTM) for time series hydrocarbon production prediction. The knowledge gained through these works provides both a foundation as well as an opportunity to extend these approaches to other aspects critical to O&G development where: 1) Dynamic, temporal dependencies exist; 2) said aspects possess significant connotations to production forecasting; and 3) that have not been extensively explored in previous research. An obvious need that meets these criteria would be to possess the ability for assessing the potential volumes of the associated water and natural gas produced in tandem with crude oil. Many operators targeting oil-rich unconventional reservoirs are faced with the challenge of managing large volumes of water and natural gas that are often co-produced. Limited natural gas processing and pipeline takeaway capacity can force operators to resort to venting or flaring produced natural gas.

Venting is the direct release of natural gas produced from O&G operations to the atmosphere. Flaring involves the controlled combustion of produced natural gas at the wellhead, converting methane to carbon dioxide and water vapor. From an environmental standpoint, flaring is less detrimental than venting given that carbon dioxide is 25 to 28 times less impactful as a greenhouse gas than methane over a 100-year period [176, 177]. According to the EIA, the quantities of natural gas vented or flared from O&G wells in the U.S. reached record levels in 2019 averaging 1.48 billion cubic feet per day (Bcf/day) (1.3 percent of the total natural gas volume produced) [178]. Texas and North Dakota contributed nearly 85% (1.3 billion cubic feet per day [Bcf/day]) of all reported flaring and/or venting (only Texas contributed to gas venting) of produced natural gas. Produced water is often managed via disposal through deep well underground injection.

The injection of large volumes of waste water from O&G operations has been strongly correlated to the increased frequency of occurrence of induced seismic events including magnitude 2+ earthquakes, particularly in Oklahoma, Ohio, Arkansas, West Virginia, and Texas [179]. Literature suggests that many are working to generate solutions and reuse options for associated gas and water production [180, 181, 182] – but a need exits to be able to effectively quantify and forecast produced volumes of both natural gas and water to best inform the development of management or remedial solutions as well as grasp the potential environmental implications for planned O&G development [183].

For this study, we propose a combination of supervised and unsupervised ML approaches as part of a framework that can reliability estimate both produced water volumes and natural gas associated with oil production in a time series fashion. This type of predictive modeling capability is expected to be useful towards 1) informing well operators as part of developing strategies to ensure the effective management, treatment, or potential reuse based on the volumes and quantities of produced fluids, and 2) supplementing hydrocarbon production outlooks with additional fluid volumes in time series fashion. Study efforts focus on the Permian Basin region of the U.S. The region holds enormous consequence regarding domestic oil and gas production. According to a report by the Texas Independent Producers & Royalty Owners Association, yearly crude oil production in the Permian Basin has grown by 1.2 billion barrels since 2009, resulting in a 371% increase in oil output over the last ten years [184]. This overall growth has enabled the Permian to become the world's top-producing oil field [185]. While the region itself major producer of both oil and gas, the basin currently faces several challenges. These include: 1) Steeper well decline rates and lower initial production (IP) values as development is moving to non-core regions; 2) associated natural gas production has outpaced pipeline takeaway capacity, which has led to an

increase in flaring and venting practices; and 3) produced water volumes and associated management costs are both on the rise [186, 187, 183]. Combined, these impacts threaten to potentially lower the Permian's overall production potential while consequently increasing the environmental burden associated with O&G operations. Therefore, an opportunity exists to propose research targeted towards these specific challenges and would provide beneficial outcomes to both potentially improving recovery and estimation of the types and volumes of fluids produced at the well level – each of which require specific management strategies and bear potential environmental implications.

4.3 Data, Study Area, and Methods

The focus of this study is to generate a ML-based prediction model capable of time series joint prediction of associated natural gas and water that are produced alongside oil as part of unconventional hydrocarbon development (three-stream production example presented in Figure 26). Secondarily, this study aims to demonstrate the utility of such a model as a compliment to existing O&G operational management strategies.

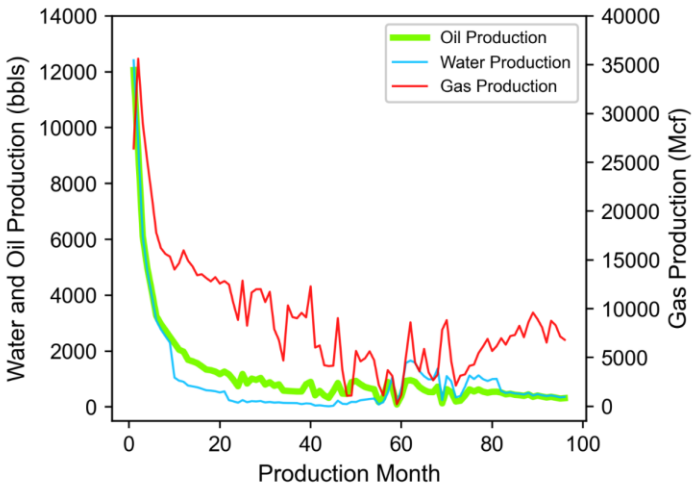

Figure 26. Example of oil, water, and natural gas production data for a horizontal well in northern Reagan County, Texas producing from Wolfcamp A and placed at a total vertical depth of 7,713 feet below ground surface.

The model is based on a deep neural network architecture leveraging LSTM layers in order to accommodate time-dependent conditions in the data and be proficient towards multi-output prediction. The model development workflow, described throughout the following subsections, is interconnected with several data preprocessing steps that includes data sub-division, engineering of new features, outlier removal, data standardization, and feature selection. The model would have the functionality to not only replicate well production history (the primary focus of many existing time series O&G analyses), but also enable fluid production forecasts that extend past observed production timeframes for existing wells, as well as be used to predict fluid volumes in time series fashion at new (i.e., theoretical) well sites. Additionally, the ML-based model proposed here is intended to be applicable across multiple producing reservoirs, focusing on the "Wolfberry" pay zones (highlighted in Upper Spraberry through Cisco/Cline [Wolfcamp D] reservoirs in Figure 27). Such a model will help provide a data-driven approach for a more holistic evaluation towards field development where multiple producing reservoir options are co-located. In the current low-

price environment for oil and gas, operators must be informed to the best extent possible of potential risks and opportunities they may face over both the short and long term [188]. The inherent challenges facing the Permian suggests that field development decision making is complex. Overall, this study proposes a modeling tool that works towards helping inform complex field decision choices by scaling up model outputs via a single predictive model.

4.3.1 Study Area

The study area for this work focuses in the Midland Basin, one of the major sub-basins of the larger Permian Basin. The Permian Basin (Permian) is an extensive sedimentary basin and major and O&G-producing region geographically located in West Texas and the neighboring areas of southeastern New Mexico. The Permian spans roughly 75,000 square miles and comprises greater than 7,000 fields in West Texas alone [189]. The Permian has been important in the U.S. energy economy for nearly a century. According to the EIA, the Permian has produced hydrocarbons for approximately 100 years and has supplied more than 35.6 billion barrels of oil and roughly 125 trillion cubic feet of natural gas (data as of January 2020). The Permian accounted for approximately 35 percent of the total U.S. crude oil production and over 13% of the total U.S. natural gas production in 2019 [190]. It is expected to remain one of the largest hydrocarbon-producing regions in the world with remaining reserves on the order of 46 trillion cubic feet of natural gas and over 11 billion barrels of oil [191]. The Permian contains several sub-basins and platforms that include the westernmost Delaware Basin, Central Basin Platform, and the easternmost Midland Basin [192]. The extent of the Central Platform and Midland sub-basins as well as the eastern edge of the Delaware Basin is shown in Figure 28.

The Midland Basin is the eastern subbasin of the larger Permian Basin and is bordered by carbonate platforms like the Central Basin Platform, Eastern shelf, and Northern shelf. The basin is at its deepest on its western edge and shallows to the east. Its western delineation is marked by folding and faulting on the eastern edge of the neighboring Central Platform. It is bounded to the east by the Eastern shelf, considered a somewhat arbitrary description that represents the shallowing in burial depth from the western edge [193]. The Northern shelf limits the basin's extent to the north. Towards its southernmost portion, basin's formations start to thin towards the Ozona Arch – an extension of the Central Basin Platform. [192].

<table>
<tr><th>Era</th><th>Period</th><th>Epoch</th><th>Local Series</th><th>Stratigraphic / Formation Name</th><th>Reservoir Operational Name</th></tr>
<tr><td rowspan="12">Paleozoic</td><td rowspan="8">Permian</td><td rowspan="2">Guadalupian</td><td rowspan="2">Ward</td><td>San Andreas</td><td>San Andreas</td></tr>
<tr><td>San Angelo / Glorieta</td><td>San Angelo / Glorieta</td></tr>
<tr><td rowspan="4">Leonardian</td><td colspan="2">Clearfork</td><td>Upper Leonard</td></tr>
<tr><td rowspan="3">Wichita</td><td>Upper Spraberry</td><td rowspan="3">Spraberry</td></tr>
<tr><td>Lower Spraberry</td></tr>
<tr><td>Dean</td></tr>
<tr><td>Lower Leonard</td><td>Wolfcamp</td><td>Wolfcamp A</td></tr>
<tr><td rowspan="2">Wolfcampian</td><td colspan="2" rowspan="2">Wolfcamp</td><td>Wolfcamp B</td></tr>
<tr><td>Wolfcamp C</td></tr>
<tr><td rowspan="4">Pennsylvanian</td><td>Virgilian</td><td colspan="2">Cisco / Cline</td><td>Wolfcamp D</td></tr>
<tr><td>Missourian</td><td colspan="2">Canyon</td><td>Canyon</td></tr>
<tr><td>Des Moinesian</td><td colspan="2">Strawn</td><td>Strawn</td></tr>
<tr><td>Atokan</td><td colspan="2">Atoka / Bend</td><td>Atoka / Bend</td></tr>
</table>

Figure 27. Stratigraphic description for a subset of the Midland Basin, Texas. The producing reservoirs of interest to this study are highlighted. This figure was generated from collective content compiled from lithostratigraphic interpretations of the Permian Basin from several literature sources **[194, 195, 196, 197, 198, 192, 190]**.

The Lower Permian aged (Leonardian epoch) Spraberry and Dean formations are made up of interbedded turbidite sands, laminated siltstone, carbonate, and organic-rich shales [196]. The Spraberry consists of upper- and lower-unit intervals [199, 200] (certain interpretations include a middle Sprayberry and Jo Mill as well [201, 202]) – the Dean formation is located stratigraphically beneath the Lower Spraberry. Each stratigraphic unit is distinguished by its lithologic composition. For instance, each of the three formations consists of thick sequences of fine-grained sandstones and siltstones that lie on top of an equally thick lower unit made up of black shales and dark carbonates [203]. The formations are known to be generally under-pressured (averaging 800 – 900 psi [5.4 – 6.1 MPa]) with matrix porosity ranging from 6 to 15 percent, matrix permeability below 10 md, and are highly naturally fractured [204, 205, 206]. The average true vertical depth to the top of the Upper Spraberry unit is roughly 6,800 feet across the Midland Basin. The complete section from the top of the Upper Spraberry to the base of the Dean ranges in thickness between 1,200 and 1,870 feet [204]. Similar to other unconventional hydrocarbon plays, productivity in the Spraberry fluctuates across the basin [207].

The early Permian aged (Wolfcampian-Leonardian epoch) Wolfcamp is described as a mixed siliciclastic-carbonate succession with stacked stratigraphic units comprising of cyclic gravity flow deposits – each separated by mudstone and siltstone [190]. The Wolfcamp is described by Sutton [204] as a dual-lithology system consisting of organic-rich shale with interbedded limestone. Lower reservoir quality portions of the Wolfcamp are associated with the presence of grainy carbonate facies, whereas higher reservoir quality portions have been tied to the occurrence of siliceous mudstones [208]. The entire section of the Wolfcamp ranges in porosity between 2 and 12 percent with average permeability near 10 millidarcies (mD) [190]. The formation varies substantially across the Midland Basin in terms of depth, thickness, and lithologic

composition. The Wolfcamp is at its deepest near the center of the Midland Basin, measuring approximately 12,000 feet deep. It shallows substantially towards the edges of the basin, varying in depth from 4,000 to 7,000 feet [204]. The thickness of the entire section of the Wolfcamp averages around 1,800 feet. The Wolfcamp is extensive throughout the Permian Basin and is considered one of the most abundant unconventional O&G plays worldwide.

The Wolfcamp formation has been appealing to O&G operators given its stacked configuration, in which multiple thick hydrocarbon-producing zones exist in sequence [209]. The stacked intervals of the Wolfcamp formation are called benches – from shallow to deep they are referred to as A, B, C, and D. Each bench has shown to be different in terms of its overall lithology, fossil content, total organic carbon content, and thermal maturity [210]. Saller et al. (1994), Blomquist (2016), and Peng et al. (2020) provide detail on the geologic composition of the Wolfcamp and various benches within and therefore the differentiation is not described at length here [211, 212, 213]. Recent development efforts in the Midland Basin are preferentially targeting the more oil-rich Wolfcamp A and B (roughly 95 percent of total Wolfcamp production) opposed to the more gas-rich Wolfcamp benches C and D [214, 210].

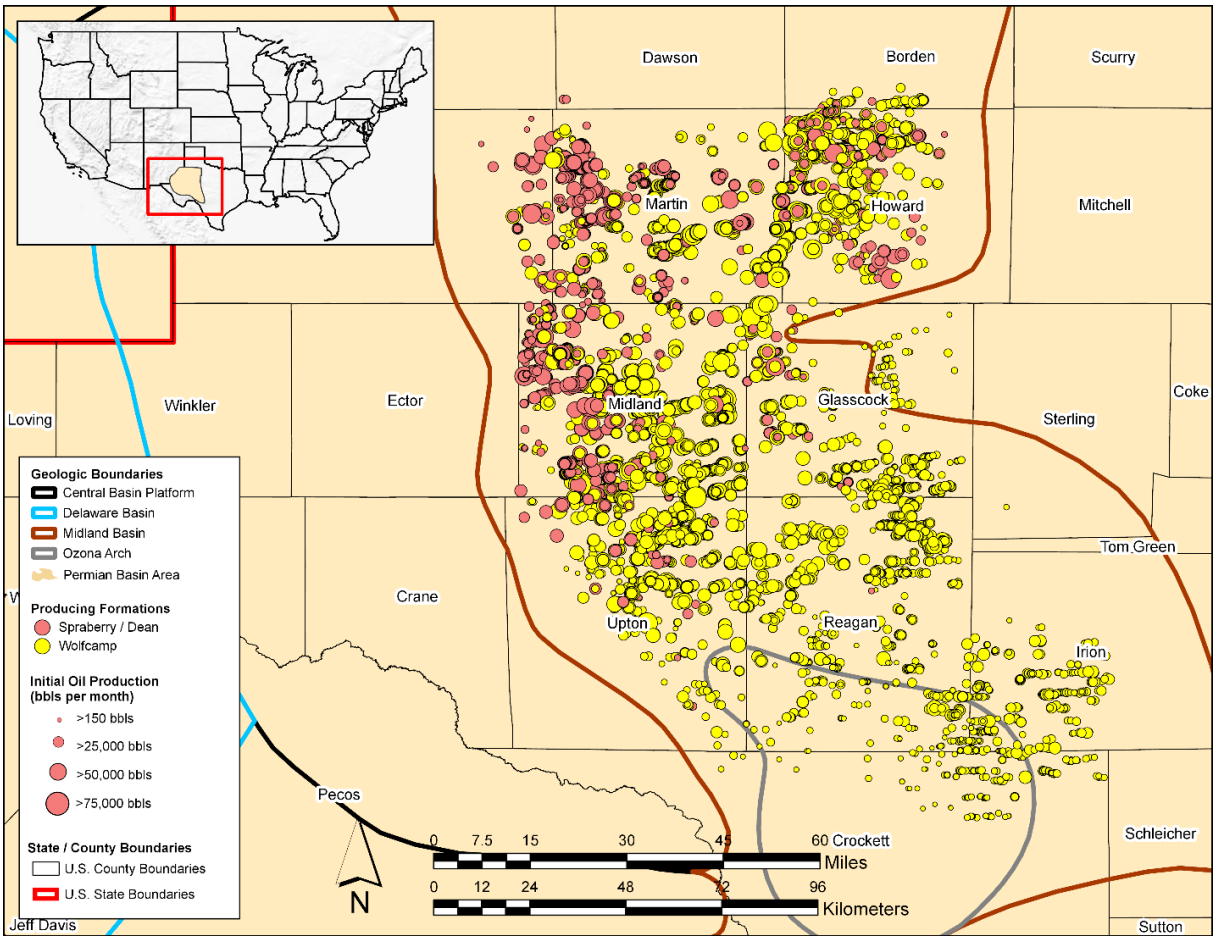

Figure 28. Map of the study area in the Midland Basin, Texas. Well data used for the study was acquired from DrillingInfo / Enverus **[215]**. The geographic information system (GIS) layers applied to support the generation of this figure were acquired from the University of Texas at Austin **[216]** and United States Geological Survey **[217]**.

The Permian region and associated sub-basins have been known to produce large volumes of natural gas and water that are co-produced with oil. A study by Kondash et al. has noted that Permian Basin wells have increased the water used per well as part of hydraulic fracturing operations from 30,800 barrels per well in 2011 up to 267,325 barrels per well in 2016 – a 770 percent increase [218]. The flowback and produced water volumes during that same timespan had increased over 400 percent; averaging 56,610 barrels per well in 2011 to over 232,700 barrels per

well in 2016. Specifically, in the Midland Basin, waste water disposal volumes derived from O&G operations have steadily increased since 2011, reaching approximately 4.5 billion barrels per day in 2017 [219].

In 2017, flaring and venting of natural gas in the Permian basin in Texas and New Mexico was estimated at nearly 300 million cubic feet per day (MMcfd), roughly 4.4 percent of the total gas produced that year. In that same year, the Midland Basin produced approximately 1,019 billion cubic feet (Bcf) of natural gas, and flared 24 Bcf of that total (2.35 percent of all gas produced) [220]. In 2019, flaring and venting of natural gas in the Permian reached an all-time record high based on the year's third quarter estimates, averaging 752 MMcfd (275 Bcf total) [221]. The Midland Basin portion of 2019 flaring ranged from approximately 150 to 290 MMcfd [222].

Well data leveraged for this study (described further in Section 4.3.2) are grouped based on the associated targeted producing reservoirs listed in Figure 27. Wells are tabbed as either "Spraberry / Dean" or "Wolfcamp" dependent upon their associated Stratigraphic / Formation Name. The wells used as part of this study are plotted in Figure 28; they are colored based on their associated producing formation and sized based on each well's initial oil production (in barrels [bbls] / month).

4.3.2 Study Data Overview and Data Processing

Much of the well completion and production-related data used for this study is acquired from the O&G data vendor DrillingInfo / Enverus [215]. Other features were derived through feature engineering to further supplement the available feature dataset. The dataset contains features related to well production performance attributes, Arps decline curve attributes [223], well completion attributes, and spatial and reservoir attributes – all specific to horizontal production

wells spanning the Spraberry / Dean and Wolfcamp producing intervals (highlighted in Figure 27) in the Midland Basin with drilling initiation dates within the January 1, 2010 to June 30, 2020 timeframe. The dataset includes a combination of static (well data that does not change over the well's productive lifetime) and dynamic features (well data with temporal dependencies – mostly three-stream production data) for the wells meeting these screening criteria. This database query yields data for approximately 6,480 wells in total in which each well has data reported for all features of interest (both static and dynamic features) and duplicate entries are omitted. No attempts at data interpolation with respect to missing values occurs in this study.

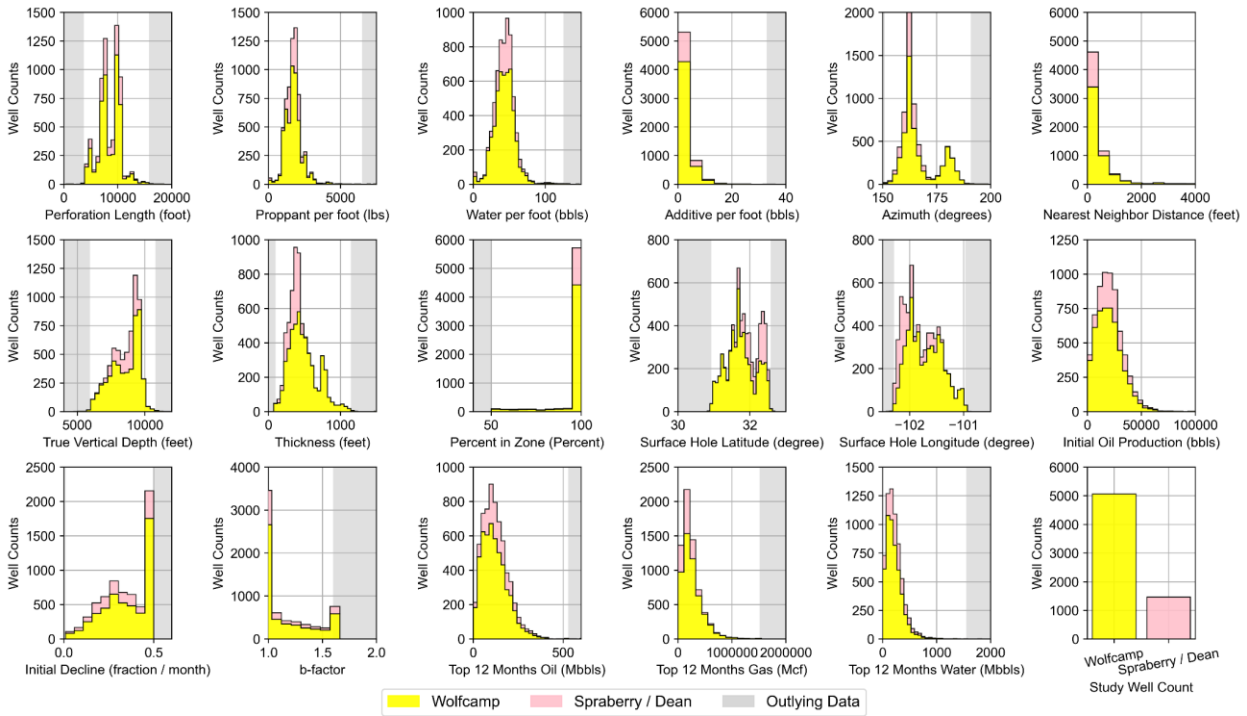

Figure 29. Distribution of static features for each well in the study dataset.

The distributions of the static study features of interest are evaluated to screen and remove potential outlying well data and refine the overall dataset. Their distributions are presented in

Figure 29. Data outside of +/- 3 standard deviations from a given feature's mean value (grey margins within subplots in Figure 29) are considered outlying and possibly highly influential on ML model response [224, 225]; even if distributions are not explicitly normally distributed. All outlying data is removed from the static and dynamic contributions to the dataset (approximately 270 wells had features meeting outlying criteria). The resulting dataset consists of 6,210 wells in total extending across 12 Texas counties, the extent of which is plotted in Figure 28 and the descriptive statistics for features from these wells are summarized in Table 13.

Table 13. Summary of the study dataset features evaluated.

Dataset Features	Data Group	Static	Dynamic	Mean	Median	Standard Deviation
Monthly Oil (bbls)	Well Performance Attributes		X	4,863	2,429	6,448
Monthly Gas (Mcf)[8]			X	12,500	7,906	13,846
Monthly Water (bbls)			X	8,510	3,572	13,496
Top 12 Months Gas (Mcf)		X		251,286	207,532	182,648
Top 12 Months Oil (bbls)		X		124,320	114,314	70,210
Top 12 Months Water (bbls)		X		226,856	197,664	157,721
EUR Gas (MMcf)		X		1,732,470	1,171,682	1,722,215
EUR Oil (bbls)		X		449,302	380,333	326,663
Initial Oil Production (bbls)[9]	Decline Curve Attributes	X		20,807	19,675	11,593
Initial Decline (fraction / month)		X		0.35	0.36	0.13
b-factor		X		1.2	1.0	0.2
Timestep Cumulative (months)			X	25.3	21	18.8
Perforation Length (foot)	Well Completion Attributes	X		8,480	8,302	1,959
Proppant per foot (lbs)		X		1,732	1,718	548
Water per foot (bbls)		X		43	44	14
Additive per foot (bbls)		X		2.9	2.4	2.4
Azimuth (degrees)[10]		X		166	163	8
Nearest Well Distance (feet)		X		438	231	838
Percent in Zone (percent)		X		97	100	10
True Vertical Depth (feet)	Spatial and Reservoir Attributes	X		8,571	8,828	993
Thickness (feet)		X		460	415	188
Surface Hole Latitude (degrees)		X		31.8253	31.7971	0.4093
Surface Hole Longitude (degrees)		X		-101.7740	-101.8346	0.3204

[8] Mcf = thousand cubic feet

[9] DrillingInfo / Enverus quantifies initial oil production as the cumulative production volume observed during a given well's first full month of production [330].

[10] All wellbore azimuth trajectories based on true north = 0 degrees.

The features within each data group from Table 13 have a specific role as part of the hydraulic fracturing and oil / gas production process. The breadth of data features available within the study dataset affords the opportunity to explore a multitude of aspects related to unconventional oil and gas production in the Midland Basin. Data groupings and their associated features are briefly described in the following bullets:

- ***Well Performance Attributes:*** These features relate to fluid production for wells in the study dataset. The dynamic features within the data group represent summation of the three-stream (oil, gas, and water) empirically-derived monthly values at the well level provided by DrillingInfo / Enverus. Data for these dynamic features is available for each month in a given well's productive lifetime. Therefore, the volume of this data varies across wells depending on when they began production and how long wells are kept online. The "Top 12-months" static features for oil, gas, and water were derived via summation of the 12 largest observed values for each well based on monthly dynamic feature data. This approach has been implemented in our prior work [169, 122] and has proven to effectively represent productivity potential for unconventional wells that may or may not have been subject to disruptions to their production time series profiles. Both the Top 12-months Oil and Gas features correlate strongly to well level estimated ultimate recover (EUR) as indicated in Figure 30. The static EUR features represent an estimation of the technically recoverable reserves at the well level. They are calculated by DrillingInfo / Enverus [226] using a combination of historic production data and a combination of Arps decline curve models [223].

- ***Decline Curve Attributes:*** These features are inherent to decline curve analyses based on the Arps decline curve model [223]. The Arps model can be used to evaluate oil and / or gas declining production rates over time. Time-dependent reduction in hydrocarbon production can be attributed to reduced reservoir pressure as well as the relative change in the volumes of the produced fluids. The approach can also be used to forecast hydrocarbon production into the future. The Arps approach is based on fitting a mathematical decline model (either exponential, hyperbolic, or harmonic) to empirical observations of an asset's (i.e., well) performance history [227]. Well features related to initial (oil) production, the initial decline, and degree of curvature (b-factor) are the parameters related to the Arps model. Values for these features for each well in the study dataset have been determined by Drillinginfo / Enverus [226]. The DrillingInfo / Enverus approach solves for the most appropriate Arps model parameters that minimize the sum of squared errors based on empirical production values for a given well [226]. DrillingInfo / Enverus restricts b-factors between 0 and 2. The b-factor is typically greater than 1 in unconventional shale plays given the inherent low permeability rock matrix and resulting extended duration of transient flow [228]; potentially a derivative of the bulk of empirical observations with shorter producing timeframes [229].
- ***Well Completion Attributes:*** These features pertain to each well's design and completion attributes as it relates to well placement, orientation, and hydraulic fracturing design. The major hydraulic fracturing design features include the length of the perforated interval contacting the reservoir and the volume of proppant,

water, and additive used for hydraulic fracturing normalized to a per foot of perforated interval basis. Proppant includes solids that may vary in size, shape or material type. They typically consist of sand or engineered materials (i.e., resin-coated sand or high-strength ceramic materials like sintered bauxite) and are used to keep reservoir fractures open and conductive following hydraulic fracturing [230]. Additives may serve a variety of functions, with examples including the assurance of effective transport of water and proppant downhole and throughout the reservoir, as well as to ensure sustained hydrocarbon recovery after hydraulic fracturing. Specific components can tend to vary from one well to another and from operator to operator. However, example constituents include acids, friction reducers, biocides, pH adjusters, scale inhibitors, iron stabilizers, corrosion reducers, gelling agents, and cross-linking agents [48, 231]. Other important well design characteristics captured in the dataset relate to the wellbore lateral orientation, spacing distance to nearby wells, and the portion of the horizontal perforated length within the targeted producing reservoir zone of interest. The directional alignment (reflected by azimuth) is often a design choice by field operators; one that is driven by the natural orientation of in situ stresses in targeted reservoir producing zones. Horizontal segments of wells that are drilled along the minimum horizontal stress often produce transverse fractures following horizontal fracturing. This form of fracturing may improve drainage efficiency. As a result, well laterals oriented properly on azimuth given natural in situ stress regimes may experience higher productivity [48, 163]. Well azimuth was approximated based on the geographic orientation between each well's surface hole latitude and longitude

and lateral toe latitude and longitude. Well spacing may provide insight into the field operator's anticipated drainage area based on the applied water and proppant intensity. Additionally, spacing-related data can be helpful in determining if closely-spaced wells suffer from possible interference from hydraulic fracturing operations (i.e., frack hits) or effects from parent / child well interactions [232, 233] from nearby wells. We approximated the nearest well distance for each well in the dataset using the haversine formula and bottom hole latitude and longitude coordinates to its closest well neighbor prior to any dataset reduction. Percentage in zone is a metric which provides an indication of the wellbore geo-steering efficiency of the horizontal lateral component. DrillingInfo / Enverus provides this data readily for each well. Wells with a high portion of their perforated segment in the targeted producing zone are more likely to be better producers than those wells expected to deviate substantially off target. Each feature in this data group is treated as static. In actuality, many of these features, like proppant, water, and additive per foot, could essentially vary over the life of any given well due to refracturing campaigns.

- ***Spatial and Reservoir Attributes:*** The features included attempt to best approximate the variability that may exist in the geologic conditions which influence hydrocarbon prominence and producibility that span the reservoirs of interest across the study domain. True vertical depth and thickness (i.e., reservoir thickness) are provided from DillingInfo / Enverus for each well. However, other relevant geologic characteristics that are known to influence hydrocarbon production, like total organic carbon, porosity, hydrocarbon and/or water

saturation, thermal maturity, reservoir pressure, existence of fracture networks, and capacity of the reservoir(s) to be hydraulically fractured [111, 107, 112, 102], are not directly or readily available in bulk. Additionally, many of these features are dynamic in nature and change over the duration of hydrocarbon production (such as fluid saturation and pressure in the reservoir), while others essentially remain static (such as porosity and thermal maturity) [234]. Each well's locational data (surface latitude and longitude) is used as a contingency means to approximate geologic conditional variability known to vary spatially across the study area – an approach widely used in other ML-based model development efforts occurring over large spatial horizons [64, 59, 15, 55].

A correlation matrix using Pearson's Product-Moment Correlation is presented in Figure 30 which provides quantitative indication of the linear relationship between each of the various static features of interest. A Pearson Correlation value of 0 or proximal to 0 indicates that no linear relationship exists between the two variables. Positive values indicate increasing linear relationships (1 represents a perfect positive relationship), whereas negative values signify decreasing linear relationships (with −1 representing a perfective negative relationship).

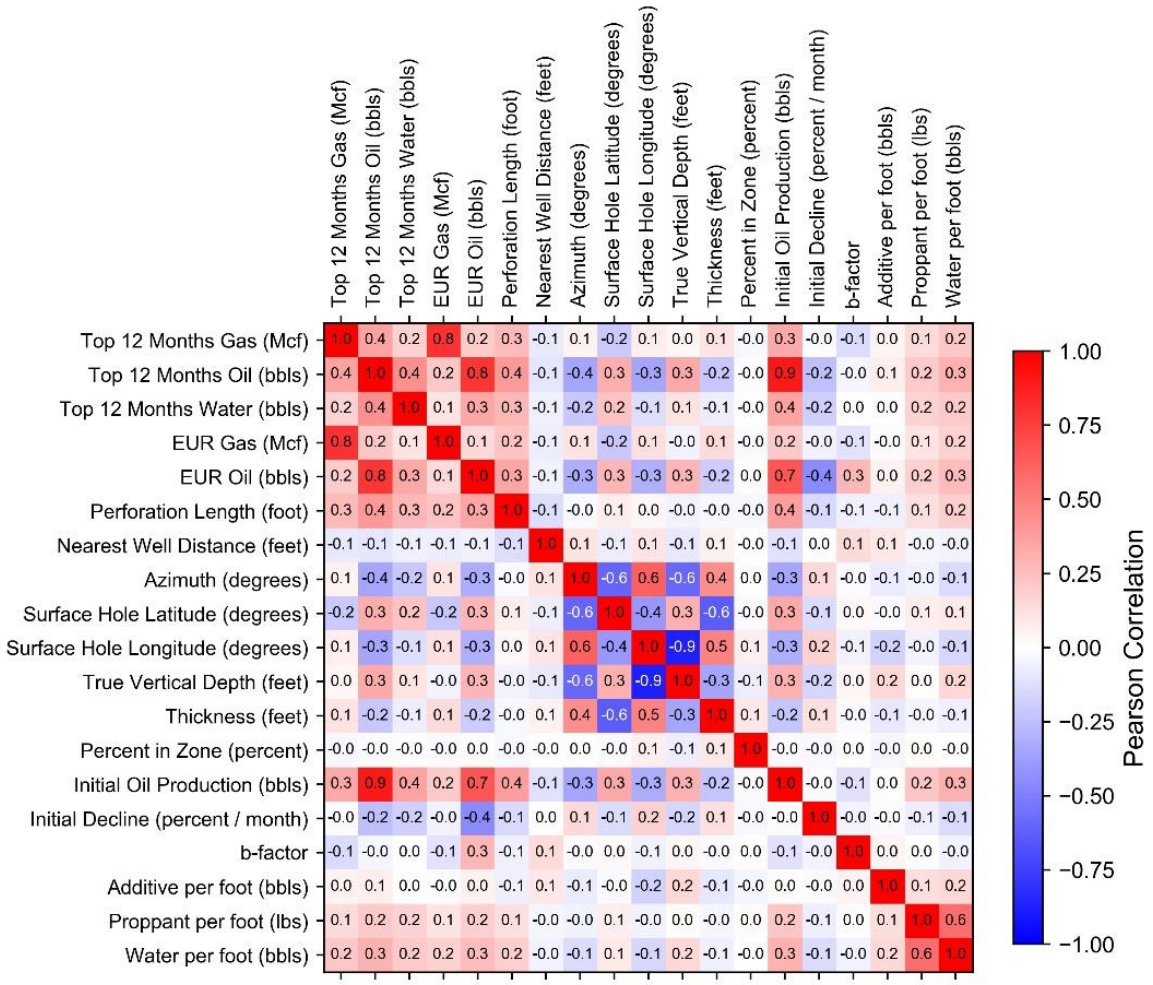

Figure 30. Pearson correlation matrix for the static dataset features evaluated.

The analysis represented in Figure 30 is informative specifically because: 1) This helps summarize the emerging patters that exist given the large volume of data features available; 2) it suggests how attributes correspond to other attributes, as well as with potential model outputs; and 3) it serves as a diagnostic check on data quality to ensure data features are related in a fashion that is intuitive and confirmatory based on heuristic understanding of the Midland Basin.

The Pearson Correlations alone highlight a number of noteworthy trends. For instance, Figure 30 shows several positive relationships between many of the well performance attributes

representing fluid production with well completion attributes specific to hydraulic fracturing design. The attributes of top 12-month oil, water, and gas, as well as the estimated EUR per well for both oil and gas are all positively correlated with increasing values of perforation length, proppant, and water per foot. These relationships suggest greater production results from well completion and hydraulic fracturing design upscaling; a concept noted by others [15, 56, 122]. Additionally, the decline curve attributes show correlation to both the well performance and well completion attribute features. Initial oil production is mostly positively correlated to these attributes, while initial decline (for oil), as expected is negatively correlated. The b-factor component is mostly uncorrelated to all features in the dataset with the exception of a positive correlation to oil EUR, and therefore holds influence over a well's longer-term productive profile. Finally, worth note are the correlations associated with reservoir thickness and true vertical depth based on well location in the basin. Moving west to east in the basin (based on surface hole latitude), Figure 30 suggests the reservoirs become both shallower and thinner. In contrast, reservoirs trend thicker and deeper when moving south to north (based on surface hole longitude). These correlations are as expected based on interpretations of Midland Basin reservoir depth and thickness isopaches and interpretations generated by the EIA [235, 192], Hamlin and Baumgardner [199], and Blomquist [213]. Based on this analysis, the dataset following outliers removed appears representative and suitable for use in ML model development.

4.3.3 Data Preprocessing Prior to Model Training and Testing

An important data preprocessing step is applied that scales attribute data in order to 1) afford equal consideration to all attributes considered, 2) improve training efficiency and, 3) increase numerical stability of the resulting models [236]. Data scaling is widely common in many

ML applications. The data scaling approach is implemented to both the static and time series parameters prior to use in the following feature selection and ML model development steps (described in Section 4.3.4 and 4.3.5). For the feature selection and clustering, input and response features are standardized to Z-values (Z) per Equation 4-1. For model training regarding the time series joint associated fluid production model, all features are scaled between 0 and 1 using linear mapping via Equation 4-2:

$$Z = \frac{x - \mu}{\sigma} \quad \text{Equation 4-1}$$

$$x_{normalized} = \frac{x - min_x}{max_x - min_x} \quad \text{Equation 4-2}$$

Where x represents feature values, μ is the feature mean value, σ is the feature standard deviation, min_x and max_x represent the respective minimum and maximum values for each dataset feature. The Z-score standardization step in Equation 4-1 rescales data for each parameter to a standard normal distribution with a mean of 0 and a standard deviation of 1. The data transformation from Equation 4-2 is used as a variant to the zero mean, unit variance standardization from Equation 4-1. The authors have gleaned from recent experience the effectiveness of 0 to 1 scaling in deep learning ML applications [237, 238, 239, 240] and are therefore applying it here. Predictions using finalized ML models are rescaled to their normal unit ranges.

Following data standardization and/or normalization, project dataset features are apportioned and merged into distinct dataset aggregates for use dependent upon the associated

project objective. The data features that are carried forward are largely dependent on the results from the feature selection, described in Section 2.4.

4.3.4 Feature Selection Approach

Features (i.e., variables) that are strongly correlated are therefore linearly dependent and may have almost correspondingly similar (if positively correlated) or opposing (if negatively correlated) effects on dependent variables of interest. The Pearson correlation metric (presented in Figure 30) is limited to assessing linear relationships concerning two features. However, important functional relationships between two or more features may exist which may not be linear in nature. This can be true even if Pearson correlation coefficients are close or equal to 0 [241]. As a result, Pearson correlation may be insufficient for informing model feature selection if used in isolation.

Feature selection involves a systematic process to down-select a subset of the most relevant features within the study dataset that strongly contribute to the ML model prediction response. Utilizing fewer features (and eliminating redundant or non-informative features) enables ML algorithms to train faster and more efficiently. Additionally, the use of fewer parameters can reduce dataset complexity, thereby decreasing the likelihood of ML algorithms overfitting to irrelevant input features and negatively impacting model prediction performance [242]. This study utilizes recursive feature elimination with cross validation (RFECV) as a feature selection approach. The objective was to establish a final set of input features that would be commonly applied as part of both the clustering evaluation and the development of the time series joint associated fluid production model.

The feature elimination component of the RFECV process searches for a subset of features by starting with all features in the training dataset and fitting a ML algorithm which is used as the

estimator [243, 242]. The estimator is trained on the original set of features considered. A total of 14 input features (i.e., *x* data) are included in this study which comprise variables associated with the "Well Completion Attributes," the "Spatial and Reservoir Attributes," and the Top 12-months Oil listed in Table 13, as well as two categorical variables that label the production wells evaluated based on their producing reservoir group – either the Wolfcamp or Spraberry / Dean formations. Two features are used as responses (i.e., *y* data) which comprise of the Top 12-month Water and Top 12-Month Gas. Static data (e.g., Top 12-month Water or Gas) was used as part of the RFECV instead of dynamic time series data (e.g., Monthly Water) in order to enable more efficient training of the estimator model. The importance of each feature is acquired (via beta coefficients for linear estimators or feature importance attributes common to tree-based models) following model training. The feature(s) with the lowest importance are then pruned from original set of features [244, 245]. The procedure is recursively repeated on the pruned set and resulting model accuracy is calculated for each iteration. The feature elimination process continues until a single feature remains. The desired number of features can then be established [246, 245]; typically set at the number of features that maximizes model performance, or where the inclusion of additional features does not substantially improve model performance.

Random forest (RF) is used as the estimator in the RFECV process for this study. RF-based models are considered advantageous in RFECV [247]; most notably because they possess the ability to measure the importance of each feature [87] based on mean decrease impurity (described effectively by Hur et al. [248]). Prior to use in RFECV, the RF estimator's hyperparameters are tuned via *k*-fold cross-validation using five folds. In this process, four folds of the training dataset are amassed to train models, and the remaining fifth fold is used to test (i.e., validate) the performance of resulting prediction models. The step is repeated so that each fold is ultimately

used once for model validation while the other k - 1 folds constitute the training set [249]. An exhaustive grid search occurs as part of the cross-validation loop to tune hyperparameters. The RF estimator formulated on all 14 input data features is built on four folds training data for distinctive hyperparameter combinations evaluated [89] as part of the grid search. Trained models were then used to make predictions against held out fifth fold validation data. The process is repeated for each combination of hyperparameters evaluated. The RF-specific hyperparameters tuned as part of cross-validation includes 1) the number of trees in each forest ensemble and 2) the minimum number of samples needed to split an internal node. The maximum depth corresponding to each tree (i.e., limits the number of nodes in each tree) was unbounded. The RF hyperparameter combination that provides for the best prediction accuracy while avoiding over or underfitting is used for RFECV.

The RFECV process also involves *k*-fold cross-validation using five folds. For each of the five RFECV fold iterations, 14 RF models are generated with the feature subset size decreasing from 14 to 1. The subset size is based on the number of input features used as part of model training. Resulting prediction model performance is evaluated by explained variance (Equation 4-3) which can effectively evaluate the multi-output response nature of the RF estimator.

$$explained\_variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}} \quad \text{Equation 4-3}$$

where $\hat{y}$ is the predicted value, y is the observed value, and Var is the variance (or square of the standard deviation). The goal of the RFECV process is to identify the formulation of the RF estimator with the feature set size that maximizes the explained variance relative to the other 13 estimator feature set combinations. The selected feature set can then be utilized as the input

features for performing the clustering analysis as well as for the time series-based joint associated fluid production model.

4.3.5 Machine Learning Model Development and Evaluation

This section describes the various ML approaches implemented as part of this study, the contribution of each towards the study objectives, and how their performance accuracy is quantified. The ML approaches utilized include both supervised and unsupervised methods, as well as the use of deep learning. Static data features that remain following RFECV step are incorporated in ML-based workflows. Python (version 3) and packages within the scikit-learn library [79] and Keras [250] are leveraged as part of the ML workflow implementation.

4.3.5.1 Clustering Evaluation

The majority of the static features within the study dataset undergo evaluation via *k*-means clustering [251], an unsupervised ML approach, prior to the development of the joint associated fluid production model. This step is intended to identify congregations of closely related wells based on their well completion, decline, well performance, and spatial and reservoir attributes (Table 13). The goal of this step is to be able to harvest Arps Decline properties (b-factor, initial production, and initial decline discussed) and well completion attributes representative of given clusters; from which oil production forecasts can be generated at the well level.

The *k*-means clustering process aims to determine the optimal number of clusters based on the input dataset features incorporated. Assuming dataset A of V-dimensional entities $a_i \in A$, for $i = 1, 2, \ldots N$, with N being the number of data entities in the dataset, *k*-means creates K number non-empty separate clusters $S = \{S_1, S_2, \ldots S_K\}$ proximal to centroids $C = \{c_1, c_2, \ldots c_K\}$, by

iteratively minimizing the sum of the within-cluster sum of squared distances (W_K, Equation 4-4) between each centroid and the data entities associated [252].

$$W_K = W(S, C) = \sum_{k=1}^{K} \sum_{i \in S_k} d(a_i, c_k) \qquad \text{Equation 4-4}$$

The term $d(a_i, c_k)$ in Equation 4-4 is the distance between data entity a_i and the associated centroid location c_k. In this study, *k*-means analysis is performed over a range of $K = 1$ through 30.

Two heuristic algorithms are applied to determine the optimal number of clusters – the Elbow method [253] and Hartigan's Rule [254]. The Elbow method can be used to visually evaluate W_k as a function of the number of clusters. The optimal number of clusters occurs at the point in which adding another cluster does not result in a substantial improvement to W_k. However, determining the optimal number of clusters through a visual determination approach like the Elbow Method can be highly subjective to the evaluator's judgement. Hartigan's Rule provides an alternative cluster determination approach and is based on comparing the resulting Hartigan's Index, which is a ratio between the Euclidean within-cluster sum of squared error based on *k* number of clusters (i.e., W_k) to that based on $k + 1$ clusters (W_{k+1}). The rule utilizes the notion that when clusters are effectively separated, Hartigan's Index ($H(K)$) becomes ≤ 10 and is taken as *k* to be the optimal number of clusters (Equation 4-5).

$$H(K) = \left(\frac{W_k}{W_{k+1}} - 1\right)(N - K - 1) \qquad \text{Equation 4-5}$$

The optimal number of clusters will be determined based on the resulting *H(K)* for each *K* = 1 through 30 evaluated. The Elbow Method will be applied in tandem to provide a visual heuristic complement to the resulting optimal *K* derived from Hartigan's Rule.

4.3.5.2 Time Series Joint Associated Fluid Production Model

For forecasting in time series circumstances, a deep learning neural network based on Long Short-Term Memory was developed for the joint prediction of associated water and natural gas production as part of oil production operations (referred to as the joint associated fluid production model [model]). The model objective is to possess the capability to reproduce as well as forecast water and natural gas volumes produced at a given well at monthly resolution based on the well's: 1) Monthly oil production volume; 2) explicit spatial and reservoir attributes (limited to the Spraberry / Dean and Wolfcamp Formations) in the Midland Basin; 3) specific well completion attributes; 4) producing month number (i.e., Timestep Cumulative data per Table 13), and 5) prior three-stream (oil, gas, and water) production volumes relative to current time (t) = month_{t-1}, month_{t-2}, month_{t-3}, and month_{t-4}.

LSTM are variants of Recurrent Neural Networks (RNN) which include memory functions that enable networks to learn long-term dependencies. The conceptual basis behind RNN is to utilize information where sequential dependencies exist so that output response is influenced by prior, yet relevant elements in sequence. The inherent RNN "memory" feedback component provides differentiation from "feedforward" neural networks (e.g., multilayer perceptron) where input data are independent from one another and strictly flow from input to output [255]. As a result, RNNs are effective in evaluating sequences of data, but are subject to gradient vanishing and struggle to handle longer-term sequential dependencies [256]. LSTM is a choice RNN-based

architecture for dealing with these two noted shortcomings under circumstances where temporal dependencies that span several time steps.

The LSTM concept was first introduced by Hochreiter and Schmidhuber in 1997 [257] and subsequently expanded and adapted by other since. LSTMs utilize a memory cell structure (Figure 31) to handle long-term dependencies in time series datasets [258]. The long-term memory component is reflected in the cell state (C_{t-1}). LSTM memory cells have the ability add or omit information to the cell state (i.e., $C_{t-1} \rightarrow C_t$), but only does so through carefully regulated structures called gates. Network gates consist of sigmoid or hyperbolic tangent (tanh) activation coupled with pointwise multiplication operations.

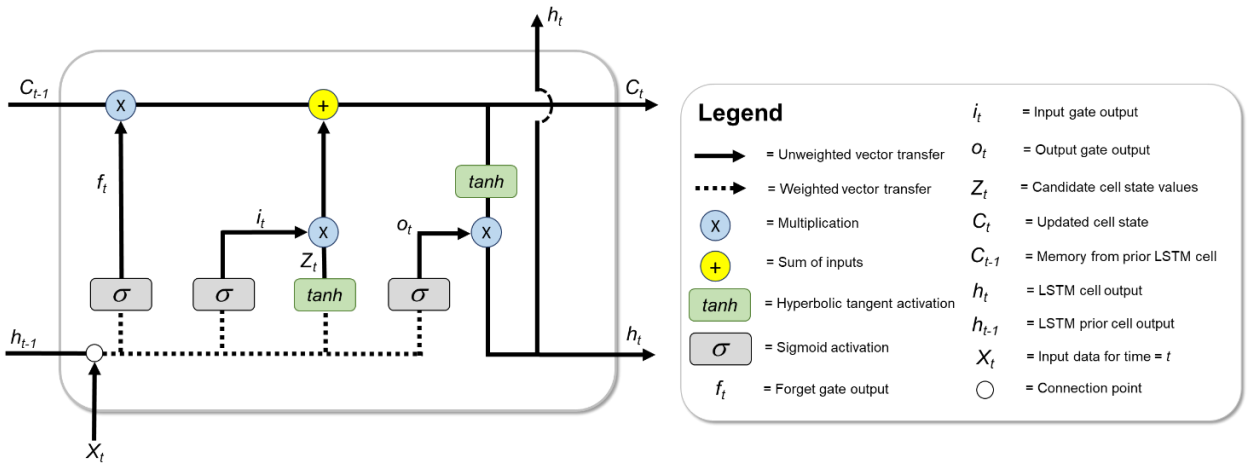

Figure 31. Example schematic of an LSTM cell. Figure concept is adapted from Kwak & Hui **[259]**, Olah **[260]**, and Poornima & Pushpalatha **[261]**.

Given the input data vector at time step t (X_t) and the previous time step LSTM cell output (h_{t-1}) instituted, the hidden state output for current LSTM cell (h_t) is calculated per the sequence discussed in the following bullets [257, 262]:

- First, the forget game (f_t) is utilized to determine information that becomes omitted away from the cell state. New information introduced to the LSTM memory cell via h_{t-1} and X_t undergoes sigmoid transformation, the result of which is output between 0 (becomes fully omitted) and 1 (becomes fully included) for each number in the cell state C_{t-1} per Equation 4-6.

$$f_t = \sigma(U_f X_t + W_f h_{t-1} + b_f) \quad \text{Equation 4-6}$$

- The second step involves determining new information to be stored in the cell state; this step occurs through two separate parts. The input gate (i_t) applies sigmoid activation to h_{t-1} and X_t and is used to inform values that will be updated in the cell state (Equation 4-7). Additionally, tanh activation generates a vector of new candidate values (Z_t), which could be included in the cell state per Equation 4-8.

$$i_t = \sigma(U_i X_t + W_i h_{t-1} + b_i) \quad \text{Equation 4-7}$$

$$Z_t = tanh(U_z X_t + W_z h_{t-1} + b_z) \quad \text{Equation 4-8}$$

- The prior cell state C_{t-1} is updated with new information to a new cell state C_t, via Equation 4-9:

$$C_t = f_t C_{t-1} + i_t Z_t \quad \text{Equation 4-9}$$

- The final step generates output (h_t) from the memory cell. The output is a function of the cell state C_t filtered via tanh activation as well as output from the output gate (o_t). The mathematical expressions for these steps are presented in Equation 4-10 and Equation 4-11.

$$o_t = \sigma(U_o X_t + W_o h_{t-1} + b_o) \quad \text{Equation 4-10}$$

$$h_t = o_t \times \tanh(C_t) \quad \text{Equation 4-11}$$

- The equation variables pertaining to U and W include, respectively, the weights to the recurrent (h_{t-1}) and input data (X_t) vectors. The b term is the bias for each gate.

Table 14. Summary of network architecture for the joint associated fluid production model.

Layer Type	Activation	Output Shape	Trainable Parameters
LSTM	Sigmoid	(None, 1, 48)	14,016
LSTM	Sigmoid	(None, 1, 96)	55,680
Dense	Relu	(None, 1, 96)	9,312
Dense	Linear	(None, 1, 2)	194

Model architecture (Table 14) and hyperparameter settings were ultimately determined via trial and error opposed to a more systematic approach like cross-validation (CV) with grid-search. The deep learning-based model requires a fairly extensive training duration (trained on a personal computer requiring approximately five seconds to train per epoch), therefore a holistic grid-search approach with CV to refine hyperparameter settings was not considered practical. Ultimately, the model network consists of three hidden units comprised of two stacked LSTM layers in a recurrent network fashion with sigmoid activation and one dense layer with rectified linear activation (Relu).

The stacked LSTM architecture is used given the noted successes demonstrated from comparable studies like Sagheer & Kotb, Utgoff & Stracuzzi, and Jie et al. that found improved modeling generalization with deep, stacked structures over shallower architectures [174, 263, 173]. The hidden layer sizes are set to vary as a function of the input size (input shape = 24 features) by 2x and 4x accordingly. The output layer enables regression-based prediction and is a dense layer with linear activation consisting of two neurons; one handling the predicted response for natural gas production and the other handling the predicted response for water production. All neurons are fully connected between model layers.

The inclusion of the dynamic well performance attributes of monthly oil, gas, and water results in a dataset size with 230,178 observations at monthly resolution. The portion of the project dataset used as part of the joint associated fluid production model development was randomly segmented into training, validation, and testing datasets through an 80/10/10 percentage-based split. This approach implements a training, validation, and testing split that respects the temporal order of observations from the project dataset by keeping the entire productive timeframe for a given well intact. For instance, 10 percent of the dataset wells (based on American Petroleum Institute well ID number) were selected at random to isolate a test dataset. All associated static and dynamic data is cross-referenced to each well for use in model development. The same process is conducted on the remainder of the dataset to isolate an additional 10 percent to serve as a validation dataset. The data from the remaining 80 percent of the wells is used for training as part of model training.

Early stopping is applied as an additional regularization step to combat overfitting. This approach monitors the predictive performance of the model for every epoch during training against predictions on the held-out validation set (22,989 observations) as a proxy for generalizing error.

Model training is discontinued when validation error is minimized conditional to the use of a patience tolerance of 25 epochs. Model weight optimization is determined under mini-batch gradient decent using the "Adam" adaptive learning rate optimization algorithm [264], a batch size = 500, and epochs = 1,000. The learning rate is set at 0.001. Keras default settings for first and second-momentum estimate decay rates as well as epsilon were used as part of Adam implementation. Once trained, model performance accuracy is evaluated on the 10 percent subset holdout test data (23,282 observations). This step also provides additional confirmation that models were not over or underfit. The performance metrics used as part of model training, early stopping, and testing evaluation are discussed in Section 4.3.5.3.

The model is easily employed to replicate a given well's historic water and gas production with the use of required input data for the given month of interest. To generate prediction forecasts for future time instances, we employ a recursive prediction approach as explained by Ji et al [265]. This strategy involves implementing the model in a $t + 1$ one step ahead prediction functionality under multiple iterations through the desired prediction horizon ($t + h$); where the prediction for the prior month (t) is used as an input for making a prediction for the following month ($t + 1$). Assuming well completion attributes do not change over time, these input features can be simply carried forward for all timesteps predicted. However, oil production is a dynamic, time-dependent input and required for forecasting water and gas volumes. Therefore, oil production forecasts that serve as inputs to the model must be derived from another means; potentially reservoir simulation output, a separate ML oil production predictive model, or even though analytical methods proposed by Fetkovich [227] and Arps [223].

4.3.5.3 Model Performance Evaluation

Our model performance was evaluated for the supervised learning-based joint associated fluid production model in two specific instances; 1) during model training against both the training and validation data sets and 2) through analysis goodness-of-fit for simulated predictions against the test dataset.

During model training, mean squared error (MSE) is used as the loss function. Performance of the model is quantified by MSE at each epoch against both the training and validation datasets; the later provides an overall generalization error estimate as well as an indication to potential overfitting if training and validation MSE's begin to diverge substantially [266]. MSE measures the mean squared difference between predicted values and the actual, ground-truth value. The metric is always non-negative and lower values (closer to zero) would suggest higher model performance. MSE is mathematically represented in Equation 4-12:

$$MSE = N^{-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad \text{Equation 4-12}$$

where N represents the length of the dataset, y_i is the observed value, and $\hat{y}_i$ is the simulated or predicted response value.

The finalized joint associated fluid production model prediction performance is evaluated by making predictions against the test dataset. A combination of MSE, root mean squared error (RMSE), and R^2 are used to evaluate model performance accuracy. RMSE correspond to the mean error between predicted and observed values and reflects the variance of errors independent of sample size. Like MSE, smaller RMSE values are associated with reduced mean error between

predicted and ground-truth data compared to model predictions where higher RMSE values occur [87]. RMSE provides a compliment to MSE and R^2, one expressed in the units of the response variable(s) of interest. The R^2 metric signifies the degree of correlation between simulated and observed values and is defined as the regression sum of squares ($SS_{Regression}$) divided by the total sum of squares (SS_{Total}). R^2 values are proportional to the data being evaluated and range between 0 and 1 – higher values represent smaller variations between the ground truth data and predicted values and lower values may suggest little to no correlation exists. RMSE and R^2 are described mathematically in Equation 4-13 and Equation 2-8 respectively:

$$RMSE = \sqrt{N^{-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad \text{Equation 4-13}$$

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1}(y_i - \bar{y})^2} \quad \text{Equation 4-14}$$

The overbar above variables Equation 2-8 indicates the mean value for the complete dataset of ground truth observations considered.

4.3.6 Oil Forecasting

Monthly oil production estimates are needed in order to predict the associated gas and water production for wells in the study area using the LSTM-based deep learning time series joint associated fluid production model. We utilize the Arps decline curve model [223] to enable oil

forecasts, either for 1) new (theoretical) wells where no historic production exists or 2) to extend historical production for existing wells. The Arps hyperbolic decline model, common for lower permeability shale production [267], is applied per Equation 4-15 to forecast oil production at the well level:

$$q = \frac{q_i}{(1 + bD_i t)^{\frac{1}{b}}} \qquad \text{Equation 4-15}$$

where q is the monthly oil production (bbls / month), q_i is the initial oil flow rate (bbls / month), b is the decline component which is dimensionless, D_i is the initial decline constant (fraction/ month), and t is the production month (month).

The Arps models have shown to provide for reliable hydrocarbon history matches (even in cases with b > 1) and affords simplicity in their use [268]. However, the hyperbolic model can tend to over approximate reserves when extrapolated without constraints to long-term transient flow considerations [269, 267]. Therefore, in this study, Equation 4-15 is only applied to forecast oil in short durations (limited to 60 months).

4.4 Results and Discussion

The findings from this study suggest that the approach outlined in Section 4.3 provides for a capable time series ML-based predictive model that can be used to either reproduce or forecast cumulative volumes of natural gas and water produced alongside oil at the well level. Based on the spatial extent of the dataset used for this analysis, the model is limited to application in the Midland Basin. However, the model can be applied to the producing intervals of both the Spraberry

/ Dean and Wolfcamp formations. The following subsections outline key results as part of model development, evaluation, and application.

4.4.1 RFECV Feature Selection Results

The hyperparameter combination selected via grid search cross-validation for the RF estimator used as part of RFECV included a formulation of 5,050 trees and a minimum of two samples to split an internal node. Figure 32 depicts the predictive performance of the RF estimator based on the number of features employed as part of training and testing via cross-validation. For this study, explained variance for each model iteration across the range of features selected are normalized relative to the feature inclusion resulting in the highest score. Once the number of features is reduced below six, the estimator's predictive performance begins to substantially decrease as more features are omitted as part of estimator training. In contrast, estimator performance gains are marginal at best when the number of features included in training are greater than six; with an optimal range between six and 11 features.

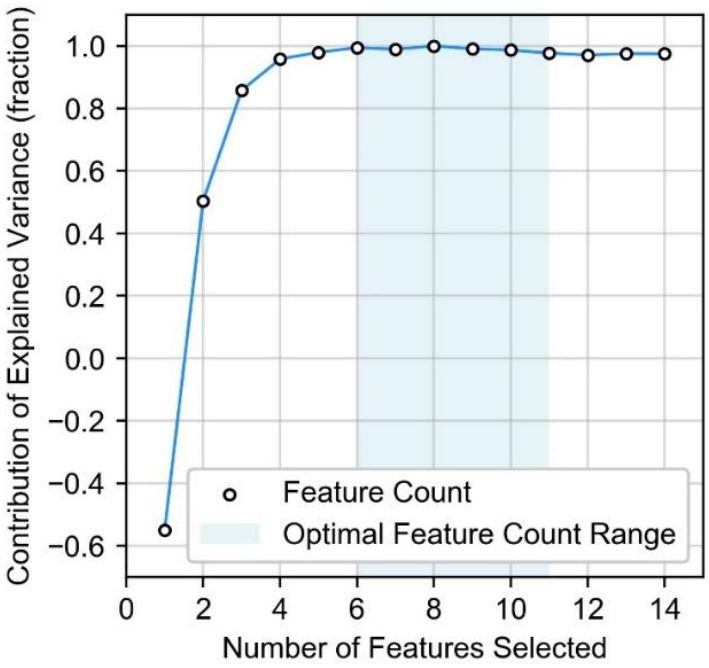

Figure 32. Effect of feature inclusion relative to the highest feature count score.

The ranking importance of each feature based on the estimator formulation with all 14 features included as part of training is presented in Figure 33. The ranking is based on the "relative" importance of each feature to that of the feature with the highest importance. The values for importance for each feature are normalized relative to the most import feature then scaled by 100. As a result, the feature with the highest importance has a value equal to 100, and all others less than 100. Examination of the feature importance ranking and magnitude indicates that oil production (reflected as Top 12 Months Oil) is the most important estimator feature for joint prediction of Top 12 Months Water and Top 12 Months Gas (static proxies for Monthly Gas and Monthly Water dynamic data features). The Top 12 Months Oil static data feature serves as a proxy for Monthly Oil, which is a dynamic model that changes with time. The following three features (latitude, longitude, and true vertical depth) specify the three-dimensional coordinates for

well horizontal placement within the basin. This finding suggests that the associated geological characterizes of producing reservoirs that also vary spatially and with burial depth are important contributors to the associated fluid response. Feature ranks five through seven (perforation length, water per foot, and proppant per foot) are noteworthy well completion design attributes.

The feature importance values in Figure 33 are used in concert with RFECV results from Figure 32 to inform the feature selection process. As a result, 11 static features are selected and three omitted from consideration for analysis moving forward. This down-selection includes omission of the features with the three lowest values of importance; which include percent in zone and the two categorical variables demarcating wells completed in either the "Spraberry / Dean" or "Wolfcamp" formations. The removal of three features and inclusion of the remaining 11 coincide with the RFECV upper bound feature range count presented in Figure 32 where explained variance remains high.

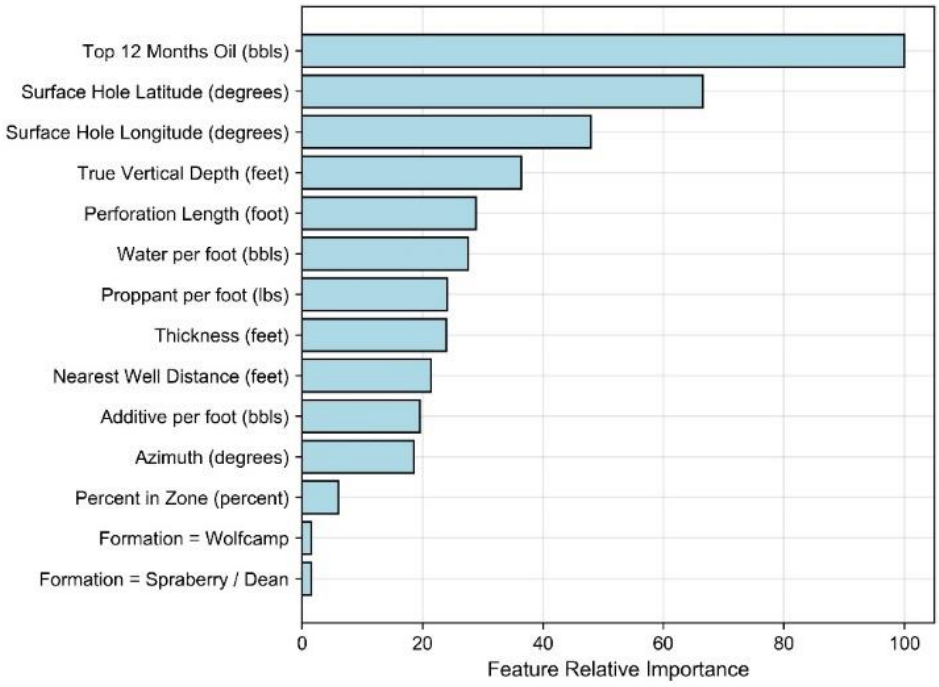

Figure 33. Summary of feature importance for the RF estimator used as part of RFECV.

The feature selection step helps establish final sets of input features that can be applied as part of both the clustering evaluation and the development of the time series joint associated fluid production model. Informed from the findings from RFECV and importance evaluation, two distinct dataset aggregates (in addition to the set used for feature selection) are created; one for clustering and another for the joint associated fluid production model training and testing.

Table 15 highlights the specific dataset features that make up each dataset aggregate. These data features are used for each of the following associated subsequent project tasks described in Sections 4.3.5.1 (clustering) and Section 4.3.5.2 (the joint associated fluid production model).

Table 15. Summary of feature inclusion for the various dataset aggregates. Each feature is demarcated for inclusion into the associated dataset aggregates as an input feature (*x*) or a response feature (*y*).

Dataset Features	Data Group	Feature Selection	Clustering	Deep Learning Time series
Monthly Oil (bbls) (*t* through *t-4*)	Well Performance Attributes			*x*
Monthly Gas (Mcf) (*t* through *t-4*)				*y*
Monthly Water (bbls) (*t* through *t-4*)				*y*
Top 12 Months Gas (Mcf)		*y*	*x*	
Top 12 Months Oil (bbls)		*x*	*x*	
Top 12 Months Water (bbls)		*y*	*x*	
EUR Gas (MMcf)				
EUR Oil (bbls)				
Initial Oil Production (bbls)	Decline Curve Attributes		*x*	
Initial Decline (fraction / month)			*x*	
b-factor			*x*	
Timestep Cumulative (months)				*x*
Perforation Length (foot)	Well Completion Attributes	*x*	*x*	*x*
Proppant per foot (lbs)		*x*	*x*	*x*
Water per foot (bbls)		*x*	*x*	*x*
Additive per foot (bbls)		*x*	*x*	*x*
Azimuth (degrees)		*x*	*x*	*x*
Nearest Well Distance (feet)		*x*	*x*	*x*
Percent in Zone (percent)		*x*		
True Vertical Depth (feet)	Spatial and Reservoir Attributes	*x*	*x*	*x*
Thickness (feet)		*x*	*x*	*x*
Surface Hole Latitude (degrees)		*x*	*x*	*x*
Surface Hole Longitude (degrees)		*x*	*x*	*x*
Wolfcamp (yes / no)		*x*		
Spraberry / Dean (yes / no)		*x*		

4.4.2 Cluster Analysis

The results from the *k*-means clustering analysis are exhibited in Figure 34. Clustering results are presented in the context of both the Elbow method and Hartigan's rule; both of which are used in tandem to select a representative number of well clusters from the study dataset where adding another cluster does not result in any substantial improvement to within-cluster sum of squared error. The visual heuristic results for the Elbow method suggest an appropriate cluster count falls somewhere between roughly 17 and 21 clusters (Figure 34A). The Hartigan Solution in Figure 34B explicitly identifies 18 clusters as optimal, and that adding the 19^{th} cluster (where 19 is the $k + 1$ cluster where the Hartigan Index ratio between k and $k + 1$ is ≤ 10) results in negligible reductions to within-cluster sum of squared error.

Wells within in the study dataset were mapped to their corresponding cluster and then plotted to inspect clustering distribution across the study area (Figure 35). An initial observation is that the resulting distribution of well clusters appears influenced by more so than just three-dimensional placement characteristics. For instance, clusters five and 14 (dark green and dark purple respectively) span a large area and occur over a variety of burial depths. Although the specific reasoning for cluster assignment is not analyzed in detail as part of this study, it is likely that non-spatial features related to well completion design or potentially well performance were influential for the commonalities of wells in these clusters. However, in certain cases, wells within certain clusters are in close spatial proximity. This seems true for cluster eight (red) in the southern portion of the basin as well as cluster 15 (light yellow) in the northeast portion of the basin. Table 18 presented later in this study provides a summary of descriptive statics for wells making up each cluster.

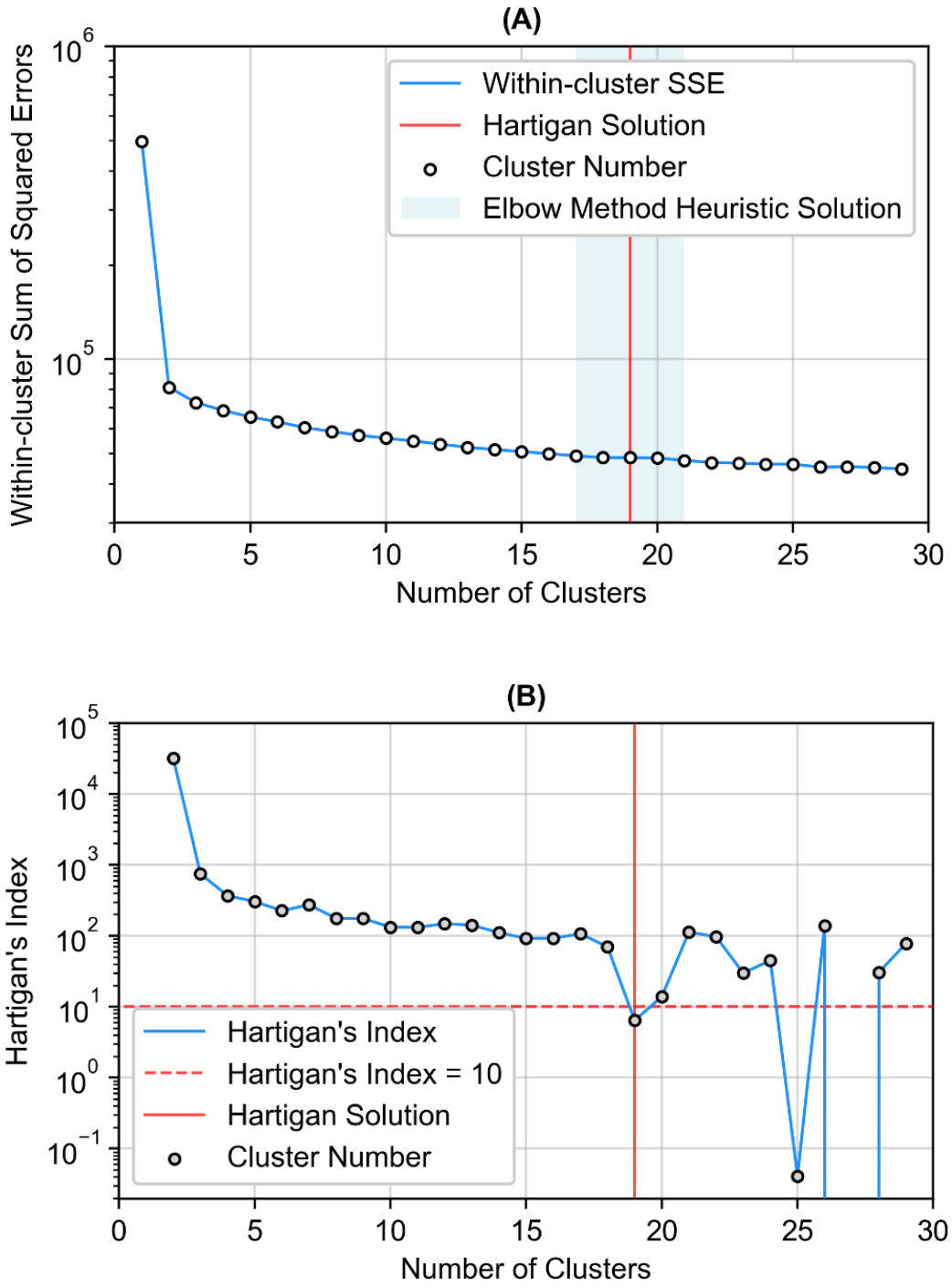

Figure 34. Elbow diagrams from *k*-means clustering analysis. The top figure (A) represents the total within-cluster sum of squared errors based on the number of clusters evaluated. The lower figure (B) shows the resulting Hartigan's Index as a function of the numbers of clusters evaluated.

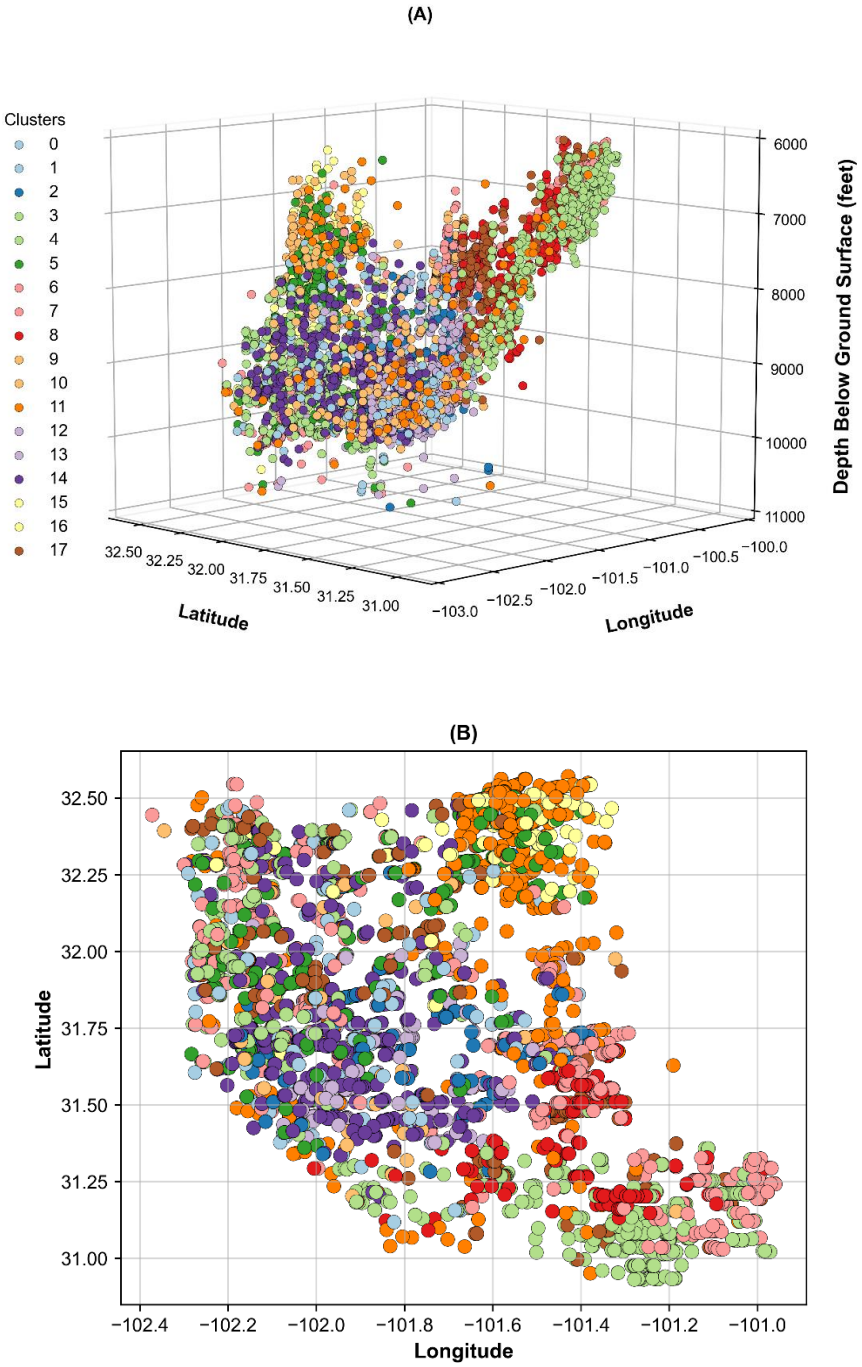

Figure 35. Well data demarcated by color corresponding one of the 18 clusters (labeled 0 – 17 based on Python's zero-based indexing). The top (A) is a three-dimensional representation of well data location which features placement along burial depth. The bottom (B) is a top-down depiction featuring well location by latitude and longitude coordinates only.

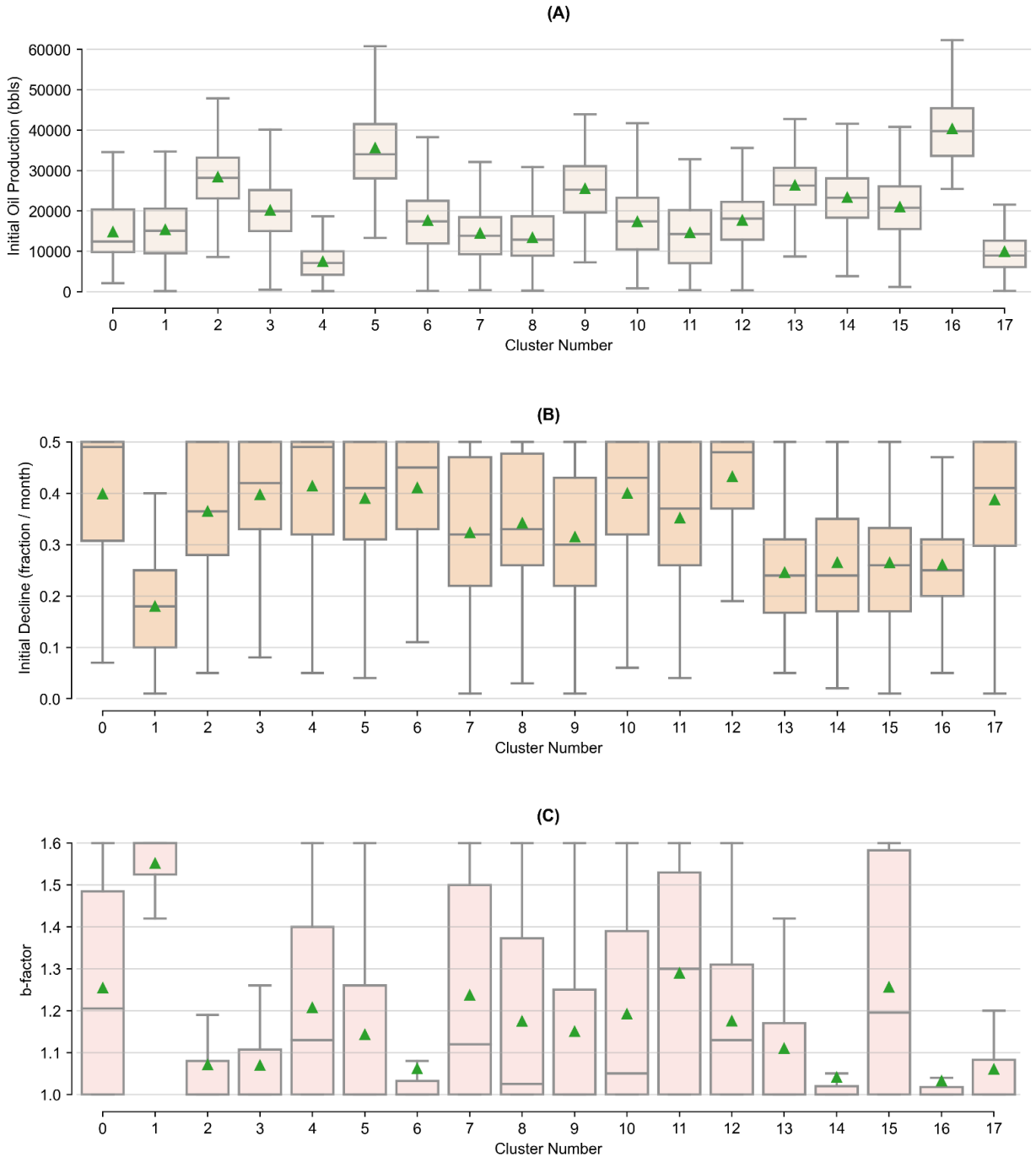

Figure 36. Box-and-whisker plots of Arps decline curve attributes calculated for wells within each cluster; including (A) initial oil production, (B) initial decline, and (C) b-factor. Boxes extends from the 25th to 75th quantile values of the data. A line occurs at the median (50th quantile). Green triangles occur at the mean value. Whiskers extend to the minimum and maximum values of the data absent outliers.

Arps decline properties can be extracted that are representative of the wells common to each cluster. These properties can then be used to forecast oil production at the well level using the Arps model (Equation 4-15). Figure 36 shows the distribution of the Arps decline properties for each cluster. Based on the distribution of these properties across clusters, oil production trends, and therefore associated gas and water, are expected to vary across clusters as well.

Multiple one-way Analysis of Variance (ANOVA) were conducted to evaluate the similarity or disparity of the Arps decline properties within across each cluster as a way to statistically infer and differentiate variability in oil production trends across clusters. ANOVA is a parametric statistical technique used to compare different datasets—specifically equality associated with their means and the relative variance between them [270, 271, 272]. In this case, the independent variable evaluated was the cluster number, which included 18 levels [0 through 17]. The dependent variables included initial oil production, initial decline, and b-factor. Null hypotheses are rejected at a significance level of $\alpha = 0.05$. ANOVA can provide insights into the overall significance of the well clusters and corresponding Arps decline properties, but the test cannot inform exactly where differences lie. Following ANOVA, Tukey's Test [273, 270] are used post-hoc to compare pairs of means for Arps decline attributes for which null hypotheses are rejected across each of 18 well clusters. The overall significance level is assumed $\alpha = 0.05$ for testing pairwise mean comparisons.

Table 16. Descriptive statistics and results from Tukey's test on decline curve attributes across well clusters.

Cluster Number	Initial Oil Production (bbls)				Initial Decline (fraction / month)			b-factor		
	Count	Mean	Stdev	Tukey's Group	Mean	Stdev	Tukey's Group	Mean	Stdev	Tukey's Group
0	84	14,816	7,459	H, I, J, K	0.40	0.12	A, B, C, D	1.25	0.24	B, C, D, E
1	259	15,364	7,653	J, K	0.18	0.09	H	1.55	0.08	A
2	246	28,382	7,826	C	0.36	0.12	D, E	1.07	0.14	I, J
3	594	20,148	7,935	G	0.40	0.11	B	1.07	0.13	I, J
4	460	7,481	4,835	M	0.41	0.12	A, B	1.21	0.22	C, D, E, F
5	574	35,577	10,694	B	0.39	0.11	B, C, D	1.14	0.20	G, H
6	328	17,625	7,588	H, I	0.41	0.10	A, B	1.06	0.13	I, J
7	609	14,442	7,392	K	0.32	0.14	F	1.24	0.25	B, C
8	230	13,408	7,594	K	0.34	0.12	E, F	1.17	0.22	D, E, F, G
9	173	25,506	8,124	D, E	0.32	0.13	F	1.15	0.23	F, G, H
10	515	17,353	8,606	H, I, J	0.40	0.11	B	1.19	0.23	E, F
11	101	14,630	8,813	I, J, K	0.35	0.13	C, D, E, F	1.29	0.24	B
12	485	17,666	6,449	H	0.43	0.08	A	1.18	0.18	F, G
13	304	26,324	6,777	C, D	0.25	0.11	G	1.11	0.18	H, I
14	554	23,346	7,579	E, F	0.27	0.13	G	1.04	0.09	J
15	160	20,971	8,156	F, G	0.26	0.12	G	1.26	0.25	B, C
16	346	40,342	8,293	A	0.26	0.10	G	1.03	0.08	J
17	188	9,959	5,386	L	0.39	0.11	B, C, D	1.06	0.11	I, J

ANOVA results yielded significant variation for all Arps attributes among well cluster as a condition, $p < 0.05$. No Arps attribute was determined to be insignificant based on well cluster groupings. Therefore, a Tukey's test was performed for each of the three Arps attributes across the 18 well clusters. The post hoc Tukey's test (Table 16) highlights which Arps decline attributes differed significantly within each cluster at $\alpha = 0.05$. Property values in Table 16 that do not share a Tukey's Group are considered significantly different from each other. The Tukey's Group lettering [A through L] are order based on the highest mean value for that given attribute relative to the other Tukey's Groups. Tukey's test results indicate that out of 18 different clusters, there are 12 statistically different initial oil production groupings, only eight statistically different groups exist for initial decline, and 10 statistically different b-factor groupings. From an Arps model perspective, higher oil productivity is tied to larger values of initial oil production and b-factor and

smaller values of initial decline. The analysis of variance and Tukey's pairwise comparison tests are performed using Minitab 18 Statistical Software.

4.4.3 Joint associated fluid production model training and performance

The predictive performance of the model as a function of training epoch is presented in Figure 37. The figure depicts the associated model loss (as MSE where model predictions, training, and validation data values are in normalized form between 0 and 1) following the update of network weights prompted by new estimates of the error gradient following each training epoch. Given the consistency of the trends in validation and training loss, the model appears to demonstrate a suitable fit to the training data with no suggestion of over or underfitting, indicating the model's overall effectiveness at generalizing associated fluid production. The application of early stopping ended model training after 325 epochs, resulting in a minimal generalization gap between training ($2.56e^{-4}$ MSE) and validation ($2.63e^{-4}$ MSE) performance.

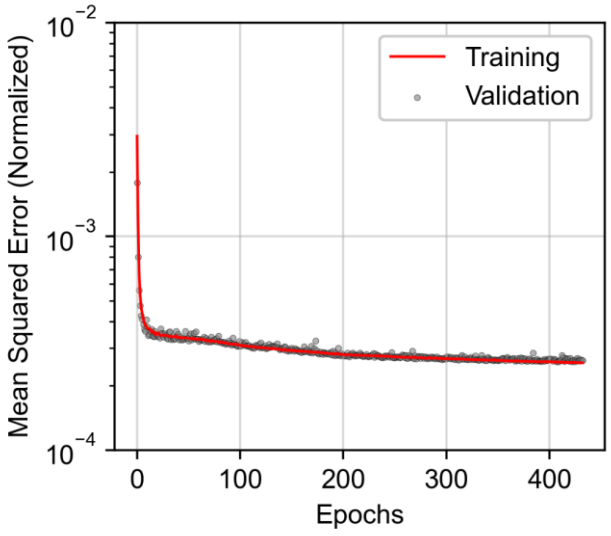

Figure 37. Learning curves for the joint associated fluid production model over training epochs.

The model's predictive performance summary against both the training and test dataset set is compared in Table 17. Performance metrics presented in Table 17 are based on the response data transformed from its normalized state per Equation 4-2 back into its original units (Mcf and bbls). Overall, there is little disparity for model performance between the training and held-out test data, as well as marginal difference in the model's ability to predict either water or gas.

Table 17. Model results for prediction on the training and test dataset.

Predicted Value	Training Data			Test Data		
	R^2	MSE	RMSE	R^2	MSE	RMSE
Monthly Gas (Mcf)	0.905	$1.83e^7$	4,273	0.900	$1.93e^7$	4,391
Monthly Water (bbls)	0.890	$2.00e^7$	4,482	0.904	$1.84e^7$	4,294
Joint Prediction (Monthly Water and Gas)	0.900	$1.91e^7$	4,379	0.900	$1.87e^7$	4,343

The prediction performance is visually compared with observed data from the test dataset in Figure 38. The parity plots (Figure 38 A and C) provide a visual depiction of the model's prediction to actual observed water or gas production on a monthly basis. The R^2 metric (listed in Table 17) is used to quantify the correlation of actual to predicted monthly production data as part of the comparison in Figure 38. Model performance that would perfectly generalize production trends would have an R^2 of one, and all data would fall exactly along the black dotted lines (i.e., 1-to-1 match) provided for reference. The model's joint predictive capability is fairly strong overall; however, the model is slightly more accurate at predicting monthly water production on holdout data compared to gas (this trend is inversed on training data predictions). Data is color coded by producing formation and sized by the production month to provide visual indicators for potential glaring trends in residual patterns. Fortunately, none seem to exist given that no

irregularities in model residuals for either formation occur based on upon visual inspection of the Figure 38 parity plots.

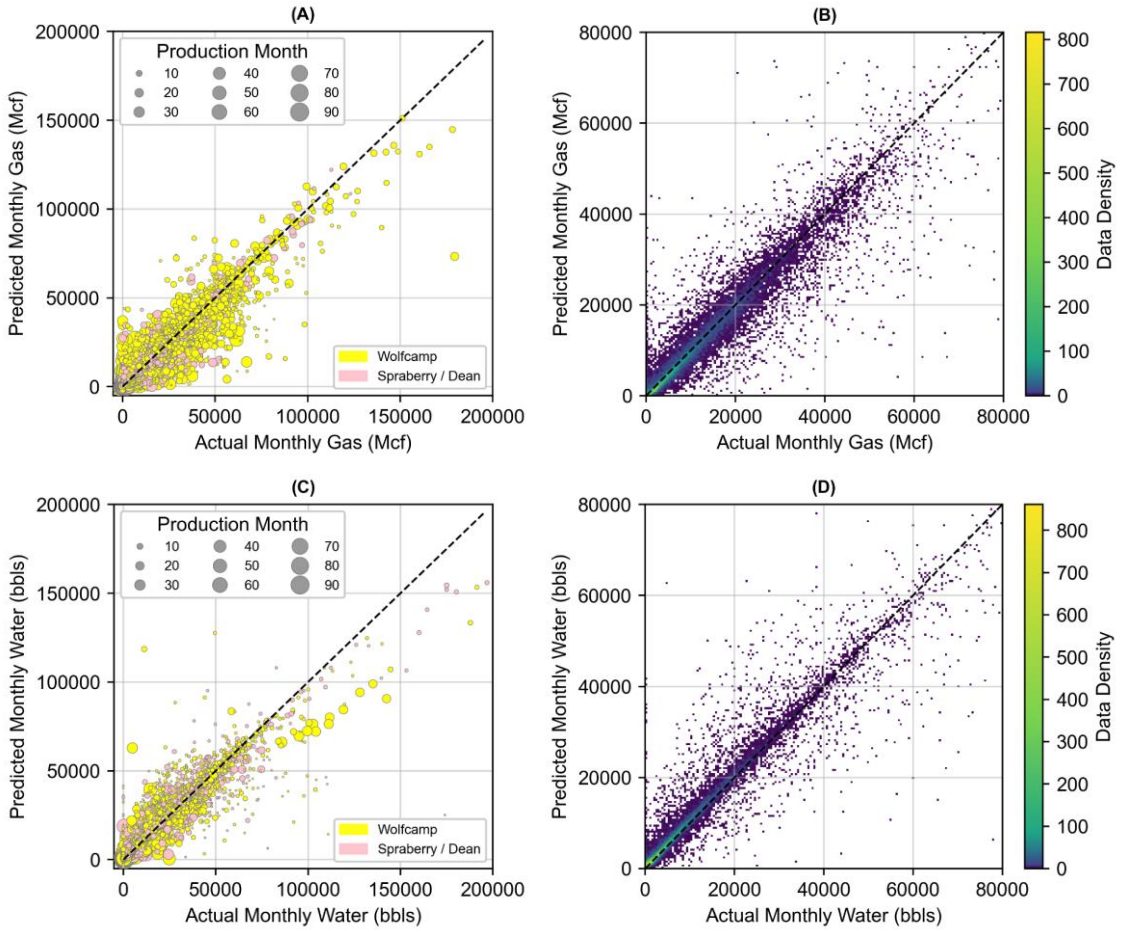

Figure 38. Parity plots of model performance comparing predicted values for monthly gas (A) or water (C) against actual values (i.e., observations) for wells in the test dataset. Additionally, the density of data within plot area pixels is provided for gas (B) and water (D).

Figure 38 (parts B and D) also features visual depictions of the density of data within each pixel of the *x* and *y* plotting area. Pixel coloration is based on the amount of data at a given *x* and *y* pixel. Viewed in isolation, the parity plots can be a bit challenging to assess the distribution of data around the 1-to-1 line given the large volume of data presented within and the spread

throughout the plotting space. The density plots emphasize where higher aggregations of data fall and where model residual (variation from the 1-to-1 line) are most prominent. The majority of monthly gas and water predictions compared to test data actuals fall along the 1-to-1 line and residuals appear evenly distributed at all fluid production volumes. Density plots are zoomed in to focus on the 0 to 80,000 bbls or Mcf fluid volume range where the majority of test data occurs.

Figure 39 shows replication of the production history for water and gas for four different randomly selected wells within the test dataset. Predictions using the joint associated fluid production model stop when known production observations end. Solid lines in Figure 39 depict actual production data for oil (green), water (blue), and gas (red) from each of the four wells. Red and blue dots indicate prediction responses for LSTM-based joint associated fluid production model. For reference, a brief review of each well evaluated in Figure 39 is provided in the bullets below:

- **Well 1:** Located in central Martin County producing from the Lower Spraberry with an 8,409-foot perforated length, and placed at a total vertical depth of 9,334 feet below ground surface.
- **Well 2:** Located in northern central Midland County producing from the Wolfcamp B with a 7,142-foot perforated length, and placed at a total vertical depth of 9,673 feet below ground surface.
- **Well 3:** Located in southeastern Midland County producing from the Wolfcamp B with a 6,722-foot perforated length, and placed at a total vertical depth of 9,383 feet below ground surface.

- **Well 4:** Located in western Martin County producing from the Wolfcamp C with a 4,855-foot perforated length, and placed at a total vertical depth of 10,031 feet below ground surface.

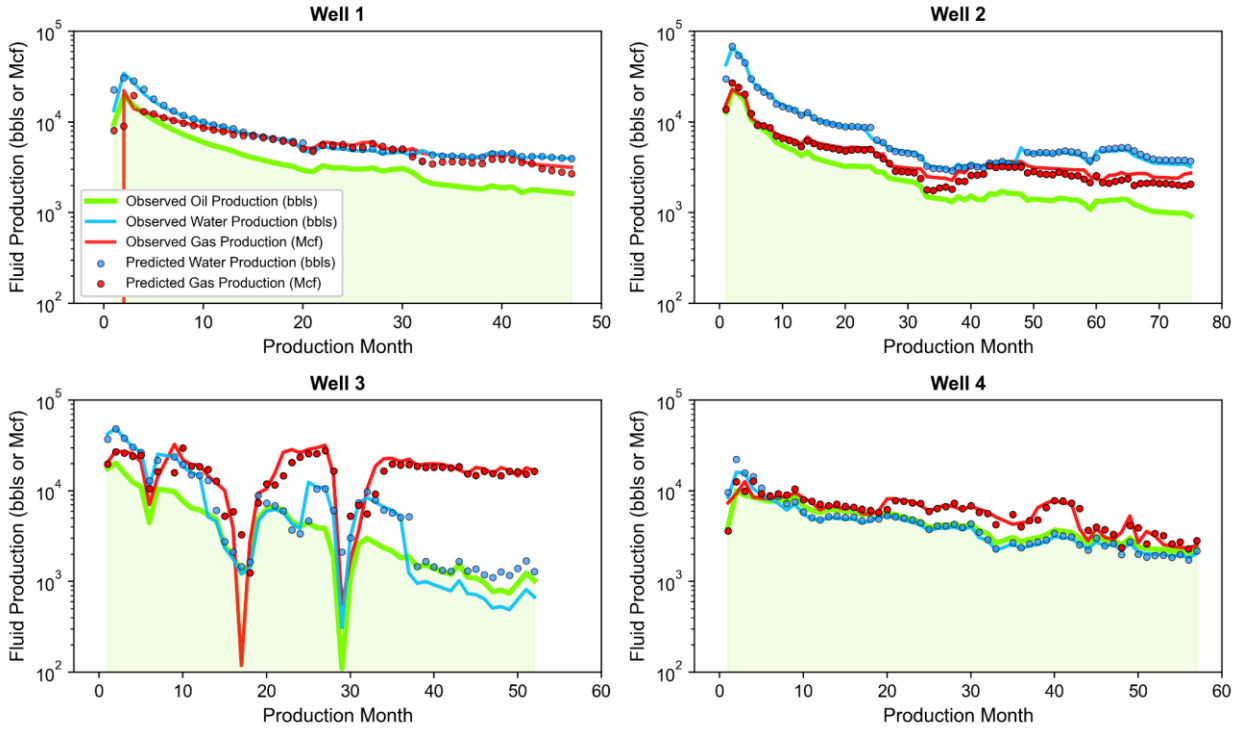

Figure 39. Replication of production history using the joint associated fluid production model for four test dataset wells.

Prediction results in Figure 39 are encouraging given the favorable replications of water and gas production profiles, even when circumstances that include irregular production trends exist. Worth noting is that the actual production trends for oil, water, and gas for each of the four wells evaluated are dissimilar in nature, yet the model is effective in replicating production profiles. Noted discrepancies in predictions to actual monthly flows seem to most commonly occur when highly transient (i.e., spikes or rapid falloffs) events transpire. However, given that the model

input features are heavily dependent on prior timestep flows for oil, water, and gas, the model appears to adjust to transient events in making next timestep predictions.

Results to this point have been based on comparison of model prediction to replicate known production flows from wells within the test dataset. However, one of the functionalities of a time series-based model lies in the ability to forecast into the future where no observations exist. We implement the model under a recursive multi-step forecasting strategy as a way to predict gas and water production trends past existing wells' known producing timeframes, as well as for generating production outlooks for new, theoretical well sites. Under this strategy, the model is used to make a prediction at time *t*, then the predicted values are appended to the input dataset to serve as prior month flow input data for predicting at time *t+1*. Oil predictions via the Arps model are incorporated as part of the input dataset to enable prediction at time *t, t+1,* through *t+h* where *h* = the total producing months prediction horizon. This process is repeated in a recursive manner until the *t+h* is reached. A simple exponential forecast smoothing function [274] is applied $t > 36$ months where the *t+1* prediction is a sum of model's *t+1* estimate plus the prior *t* value in a weighted 60/40 percentage contribution. Past the $t > 36$ producing timeframe, observed monthly water and gas values for wells in the study dataset are commonly in the range (or lower) of the model prediction error (see Table 17). The smoothing approach ensures stability in the forward predictions.

Since the joint associated fluid production model is a purely data driven model, it may be limited at making sound predictions for 1) circumstances where low quantities of data to train models exists and 2) timeframes that extend beyond the production durations for wells in the training data. Over 80 percent of the wells in the study dataset have well production timeframes less than 60 months (Figure 40). After 60 months, the volume of well data becomes sparse,

especially for Spraberry / Dean wells. Additionally, as discussed in Section 4.3.6, the application of the Arps model over the long-term with high b-factors using the hyperbolic model may overestimate hydrocarbon production. Plus, the recursive prediction strategy can suffer from error accumulation and propagation, particularly when the forecasting horizons increase [275, 276]. These potential limitations serve as the basis for setting our constraint to limit forecasts to shorter-term predictions.

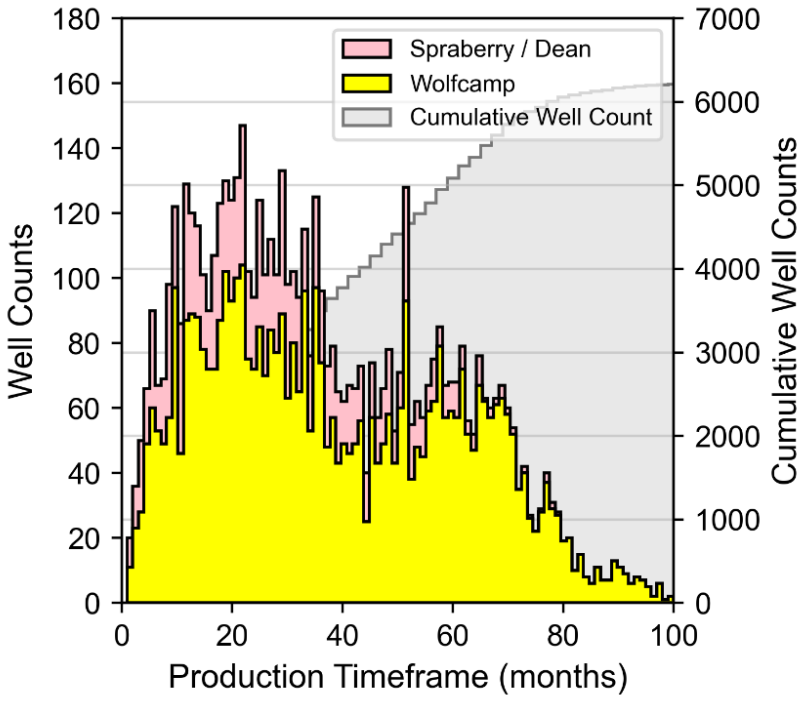

Figure 40. Stacked (left y-axis) and cumulative (right y-axis) histograms of well counts within the study dataset based on the production timeframe for each well.

Results in Figure 41 show forecasted production for oil, water, and gas for four different wells; three of which (Wells A, B, and C) are existing wells selected from the test dataset and the fourth (Well D) is a theoretical well based on the dataset mean values for input features common to Cluster 13 (see Table 18). Cluster 13 was selected as an example for analysis here because it

contains a realitvely large mean initial oil production and encompases a substantial portion of the well count from the study dataset; the majority of which are Wolfcamp wells. Forecasts using the joint associated fluid production model intentionally stop at 50 months under all cases regardless of well production history. Solid lines in Figure 41 depict actual production data for oil (green), water (blue), and gas (red). Red, green, and blue dots represent the montly forecasts for the Arps (oil) and joint associated fluid production model (water and gas).

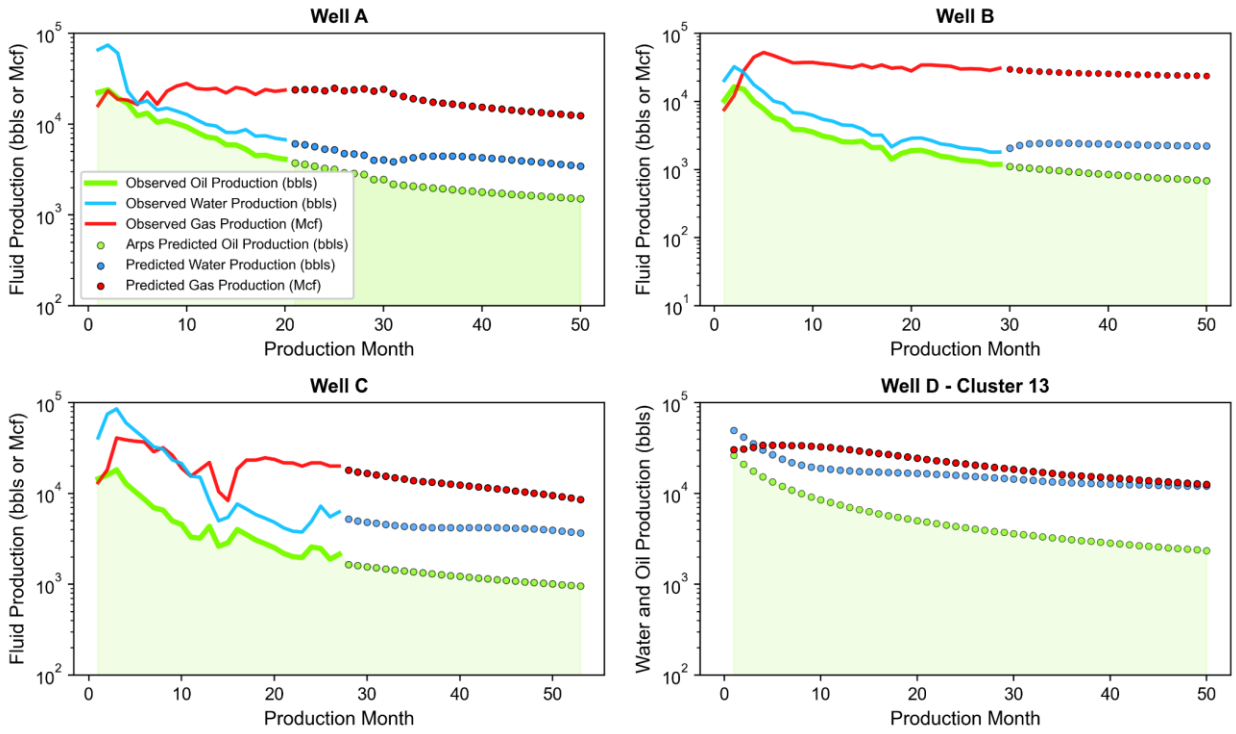

Figure 41. Gas and water prediction forecast using the joint associated fluid production model leveraging oil forecast outlooks generated from the Arps model.

For reference, a brief review of each well evaluated in Figure 41 is provided in the bullets below:

- **Well A:** Located in northern Upton County producing from the Wolfcamp A with a 7,745-foot perforated length, and placed at a total vertical depth of 9,476 feet below ground surface.
- **Well B:** Located in western Irion County producing from the Wolfcamp B with a 10,114-foot perforated length, and placed at a total vertical depth of 6,709 feet below ground surface.
- **Well C:** Located in southern Glasscock County producing from the Wolfcamp A with a 10,261-foot perforated length, and placed at a total vertical depth of 7,976 feet below ground surface.
- **Well D:** Theoretical well representative of Cluster 13 (see Table 18 in Section 4.5 for specifics) based on a 9,870-foot perforated length, an initial monthly oil production of 26,324 bbls, and placed at a total vertical depth of 9,128 feet below ground surface.

4.5 Oil, Gas, and Water Production Outlook

The joint associated fluid production model has been applied in combination with the Arps model to generate oil, gas, and water production outlooks at the well level representative of each of the 18 cluster groups identified in Section 4.3.5.1. In this example, outlooks provide cumulative first and five-year estimates for production totals based well completion, decline curve, and spatial and reservoir attributes at the mean for each of the 18 clusters. The suite of data presented in Table 18 is a digest of attribute statistics (most notably mean, standard deviation, and interquartile range [IQR]) as well as cumulative production estimates from the combination of the Arps and joint

associated fluid production models for each cluster. The collective data compiled within Table 18 is intended to serve as a guiding resource for assessing the potential volumes of produced fluids associated with oil production in the Midland Basin based on well completion design considerations and placement within the basin.

The predictions for each cluster appear aligned to typical volumes of in-field production trends for wells in the Midland Basin. For instance, our predicted production totals in Table 18 when compared in the context of water-to-oil and gas-to-oil ratios appear in range with those reported in literature [277, 188, 278, 279]. The calculated ratios are reflective of industry trends for the first year of production (ranging from approximately 1.4 to 2.8 bbls/bbls for water-to-oil [mean of 1.94] and 1.2 to 3.5 [mean of 2.0] thousand cubic feet [Mcf] / bbl for gas-to-oil). Cumulative produced water and gas (to-oil) estimates after 5-years or production are in the ranges reported by Kondash et al. and Kim respectively [280]. Additionally, the predictions capture increasing gas-to-oil ratio trends as wells becomes older [281]; not uncommon to unconventional plays, particularly when production causes reservoir pressures to fall below the bubble-point [282].

Table 18. Inventory of descriptive statics, 1st year, and cumulative 5-year production estimates for wells within each Midland Basin Well Cluster.

Data Group	Dataset Feature	Statistic	Midland Basin Well Cluster Number																	
			0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Well Completion Attributes	Perforation Length (foot)	Mean	6,782	8,791	9,593	7,928	7,719	10,061	9,177	7,139	9,663	8,307	7,677	7,253	7,225	9,870	8,762	9,448	9,972	7,417
		Stdev.	1,990	1,770	1,319	1,665	1,563	1,502	1,856	1,565	1,763	1,814	1,970	2,103	1,794	1,155	1,469	1,892	1,333	1,605
		IQR	2,985	2,704	1,132	2,378	1,065	756	2,581	1,545	1,756	2,711	2,933	4,212	2,361	563	2,313	2,612	740	1,549
	Proppant per foot (lbs)	Mean	1,818	1,698	1,764	1,659	1,303	1,845	1,938	1,477	2,283	3,281	1,728	1,609	1,441	1,752	1,812	1,787	1,828	1,342
		Stdev.	570	391	331	349	413	389	428	395	394	775	464	535	547	336	359	412	490	394
		IQR	487	468	319	430	495	367	459	517	546	872	677	759	687	305	413	507	407	532
	Water per foot (bbls)	Mean	46.8	45.6	50.7	40.6	28.2	47.8	49.6	37.2	51.4	71.4	39.4	40.6	36.6	49.4	48.8	44.9	48.0	32.8
		Stdev.	15.3	10.3	9.0	12.2	8.3	10.1	10.9	10.6	9.3	19.7	11.2	14.9	14.0	8.0	10.7	12.5	10.2	8.6
		IQR	9.8	12.4	11.6	15.5	9.1	13.0	13.1	13.2	8.7	23.2	13.8	17.3	18.9	8.7	9.6	15.5	10.6	7.8
	Additive per foot (bbls)	Mean	12.1	2.8	2.9	4.1	1.8	2.6	3.5	3.2	2.1	4.9	2.1	4.2	2.1	2.2	2.3	2.1	2.9	1.9
		Stdev.	3.9	1.8	1.5	3.1	1.3	1.6	3.3	1.8	1.2	2.9	1.5	3.4	1.3	1.4	1.5	1.5	1.7	1.2
		IQR	3.9	2.5	2.0	4.9	2.1	2.3	2.5	2.6	1.3	3.0	1.8	3.8	1.6	2.0	2.3	2.0	1.9	2.0
	Azimuth (degrees)	Mean	165.1	162.4	162.6	162.6	180.3	162.6	178.9	162.6	180.8	163.2	162.6	168.0	162.2	162.9	162.4	163.3	162.8	180.8
		Stdev.	7.0	3.7	3.7	3.6	3.4	3.3	5.9	3.3	3.1	5.5	2.9	8.8	3.3	3.7	3.4	3.3	3.6	2.1
		IQR	3.2	4.2	3.4	4.1	4.3	4.0	5.1	2.3	4.1	4.6	2.5	17.6	3.6	3.5	3.8	3.2	4.4	2.2
	Nearest Well Distance (feet)	Mean	844	288	254	261	550	259	523	382	388	395	392	5,658	303	328	243	486	278	343
		Stdev.	1,013	384	307	324	473	269	441	556	428	542	473	1,942	354	308	306	764	270	438
		IQR	1,185	307	277	267	519	295	413	390	453	419	404	2,770	285	386	259	511	316	438
Decline Curve Attributes	Initial Oil Production (bbls)	Mean	14,816	15,364	28,382	20,148	7,481	35,577	17,625	14,442	13,408	25,506	17,353	14,630	17,666	26,324	23,346	20,971	40,342	9,959
		Stdev.	7,459	7,653	7,826	7,935	4,835	10,694	7,588	7,392	7,594	8,124	8,606	8,813	6,449	6,777	7,579	8,156	8,293	5,386
		IQR	10,635	11,155	10,173	10,197	5,895	13,455	10,748	9,233	9,916	11,762	12,799	13,555	9,324	9,246	9,785	10,744	11,795	6,533
	Initial Decline (fraction / month)	Mean	0.40	0.18	0.36	0.40	0.41	0.39	0.41	0.32	0.34	0.32	0.40	0.35	0.43	0.25	0.27	0.26	0.26	0.39
		Stdev.	0.12	0.09	0.12	0.11	0.12	0.11	0.10	0.14	0.12	0.13	0.11	0.13	0.08	0.11	0.13	0.12	0.10	0.11
		IQR	0.20	0.15	0.22	0.17	0.18	0.19	0.17	0.26	0.22	0.21	0.18	0.25	0.13	0.15	0.18	0.17	0.11	0.21
	b-factor	Mean	1.25	1.55	1.07	1.07	1.21	1.14	1.06	1.24	1.17	1.15	1.19	1.29	1.18	1.11	1.04	1.26	1.03	1.06
		Stdev.	0.24	0.08	0.14	0.13	0.22	0.20	0.13	0.25	0.22	0.23	0.23	0.24	0.18	0.18	0.09	0.25	0.08	0.11
		IQR	0.50	0.08	0.08	0.11	0.40	0.26	0.04	0.50	0.39	0.26	0.39	0.54	0.31	0.17	0.02	0.59	0.02	0.09
Spatial and Reservoir Attributes	True Vertical Depth (feet)	Mean	8,924	8,964	8,947	9,310	7,112	8,811	7,150	9,020	7,460	9,078	7,883	8,340	9,238	9,128	9,123	7,751	8,963	7,523
		Stdev.	752	630	626	470	741	785	620	771	577	609	568	1,101	465	424	511	727	587	723
		IQR	798	848	884	563	963	1,296	1,018	1,174	686	784	527	1,950	555	540	673	956	962	961
	Thickness (feet)	Mean	443	398	471	320	774	375	633	374	553	503	384	463	541	653	380	369	356	477
		Stdev.	157	137	115	96	146	108	207	134	191	209	103	183	145	176	103	111	86	151
		IQR	209	148	137	148	59	136	338	185	361	244	132	224	168	289	119	110	115	254
	Surface Hole Latitude (degrees)	Mean	31.64	31.92	31.70	32.08	31.15	32.08	31.38	31.98	31.32	31.83	32.23	31.80	31.68	31.60	31.91	32.33	32.09	31.39
		Stdev.	0.3	0.3	0.2	0.3	0.1	0.3	0.2	0.3	0.1	0.3	0.3	0.5	0.2	0.2	0.3	0.2	0.2	0.2
		IQR	0.4	0.5	0.2	0.5	0.2	0.5	0.4	0.6	0.2	0.6	0.5	0.9	0.2	0.3	0.4	0.2	0.4	0.3
	Surface Hole Longitude (degrees)	Mean	-101.9	-101.9	-101.8	-102.1	-101.3	-101.9	-101.3	-101.9	-101.4	-101.9	-101.6	-101.7	-101.9	-101.8	-102.0	-101.6	-101.9	-101.4
		Stdev.	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.3	0.2	0.2	0.1	0.3	0.1	0.1	0.2	0.2	0.2	0.2
		IQR	0.4	0.3	0.3	0.2	0.3	0.4	0.3	0.5	0.1	0.3	0.1	0.6	0.2	0.2	0.2	0.3	0.4	0.2
	Wolfcamp	Count	68	168	223	315	456	419	326	445	230	137	356	89	459	301	321	88	227	188
	S.berry / Dean	Count	16	91	23	280	4	155	2	164	0	36	159	12	26	3	223	72	119	0
Production Forecast per Well	Cumulative Oil (Mbbls)	1st year	77	111	147	100	38	181	86	82	73	141	89	80	87	160	135	129	237	50
		5-years	154	282	275	183	74	346	156	169	145	281	173	168	167	328	265	279	465	91
	Cumulative Gas (Bcf)	1st year	0.15	0.20	0.35	0.14	0.10	0.27	0.29	0.14	0.26	0.26	0.12	0.11	0.17	0.39	0.23	0.16	0.38	0.15
		5-years	0.23	0.56	0.79	0.24	0.13	0.58	0.48	0.26	0.47	0.51	0.17	0.14	0.25	1.23	0.46	0.31	1.04	0.20
	Cumulative Water (Mbbls)	1st year	144	212	261	182	108	269	154	164	152	235	217	123	183	325	222	315	336	107
		5-years	262	710	710	316	239	685	334	315	376	502	414	190	302	998	483	913	952	213

Table 19 highlights several major takeaways from the digest presented in Table 18; particularly the cluster groups estimated to have the highest or lowest totals for 1) oil, gas, and water production per well, as well as 2) associated fluids normalized to a barrel of oil produced. The results indicate that Cluster 16 is the best oil producer for the first producing year and through 5 years of production. Cluster 16 also is noted to be comparatively efficient versus other clusters based on the associated fluids volumes produced with oil. Cluster 4 is the lowest oil producing cluster and highly inefficient regarding the associated fluids volumes produced with oil. Cluster 13 produces some of the largest volumes of associated water and gas, especially in the first year. As a result, it is one of the most inefficient clusters in terms of oil to gas and oil to water production ratios. Clusters 11, 15, and 16 are noted as relatively more "efficient" cluster than several others based on their higher oil to gas and oil to water producing ratios; both in shorter and longer projections. Overall, clusters 4, 8, and 13 appear to be the least efficient regarding associated fluid production normalized to oil.

Table 19. Summary of the highest and lowest predicted production totals and associated cluster groups.

Metric	Oil Production		Natural Gas Production		Water Production		Oil to Gas		Oil to Water	
	Mbbls	Cluster	Bcf	Cluster	Mbbls	Cluster	Mbbl / Bcf	Cluster	Mbbl / Mbbl	Cluster
Highest 1st year	237	16	0.39	13	336	16	806	15	0.71	16
Highest 5 years	465	16	1.23	13	998	13	1,200	11	0.88	11
Lowest 1st year	38	4	0.10	4	107	17	281	8	0.35	4
Lowest 5 years	74	4	0.13	4	190	11	265	13	0.31	4

We performed one last analytical case study using the data in Table 18 to generate production volume outlooks in regards to associated fluid production in the Midland Basin. Specifically, first and 5-year production outlooks are generated at the basin-level under three

development scenarios that comprise of a new fleet of wells built on different contributions of wells common to certain cluster groups. The scenarios include:

- **Scenario 1:** high efficiency development – 33.3 percent of wells based on Cluster 11, 33.3 percent of wells based on Cluster 15, and 33.3 percent of wells based on Cluster 16
- **Scenario 2:** low efficiency development – 50 percent of wells based on Cluster 4; 50 percent of wells based on Cluster 8
- **Scenario 3:** diversified development – contribution of wells from each cluster randomly assigned under equal probability per cluster

An average of 1,842 wells have been spud per year in Spraberry / Dean and Wolfcamp formations in the Midland Basin from 2017 through 2019 based on the study dataset. The generated outlooks under each of the three scenarios evaluated are therefore based on a theoretical new well fleet of 1,842 wells in each scenario. Production outlooks for oil, water, and gas volumes produced from the new well fleet in the 1st year and through five years of production are shown in Figure 42.

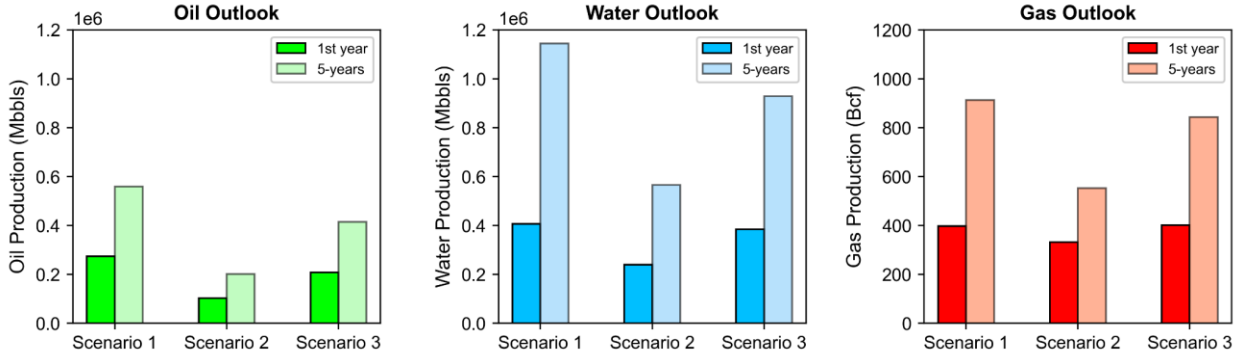

Figure 42. Oil, water, and gas production volumes under three different development scenarios for the Midland Basin. Each scenario assumes 1,842 new wells drilled and completed.

First year production volumes range from approximately 102,000 to 273,000 Mbbls oil, 239,000 to 406,000 Mbbls water, and 332 to 402 Bcf of gas across the three scenarios constructed. Production volumes through five years extend from 201,700 to 559,000 Mbbls oil, 566,000 to 1,145,000 Mbbls of water, and 552 to 913 Bcf of gas. Results emphasize the notion that development choices regarding well design and placement (varied here by clusters implemented) have considerable implications on resulting fluid production outlooks. Worth noting is that under Scenario 1, where well deployment is limited to the highest oil to associated fluid clusters, the largest volumes of associated fluids are produced compared to other scenarios. Furthermore, fluid volumes are likely to scale accordingly based on the number of wells that come online, which could be reflected in other deployment scenarios. Based on the approximate percentage of gas flared to gas produced in the Midland Basin per Leyden (2.35 percent of total), roughly 13 to 21 Bcf of gas would be flared over the five-years of production based on the results presented in Figure 42.

While this is a relatively straightforward example, it is nonetheless effective for quantifying produced volumes of both natural gas and water based on potential O&G development considerations. The outlooks can aid operators when formulating management or remedial solutions for the volumes of fluids expected. However, this analysis only includes production outlooks for the new wells considered and does not incorporate legacy production from wells producing prior to the installation of the new well fleet or those wells that come online afterwards. Production outlooks for natural gas or oil are highly dependent on a multitude of factors, including the typical production profiles of individual wells over time, the cost of drilling and operating those wells, the prospective economic return generated by those wells, the prevailing economic conditions related to O&G supply and demand, the intensity in which new wells are drilled,

completed, and turned online, and the available prospective area remaining for a given play [61, 11, 283, 284]. Forecasting associated water and gas would also be subject to similar factors. Therefore, alternative scenario formulations could be used to reflect different basin development outlooks than the one's analyzed here.

4.6 Conclusions

In this chapter, we have introduced a data-driven modeling framework that combines supervised and unsupervised ML approaches. The intent of the supervised learning component was to produce a deep learning-based model with the capability to generate reliable estimates of produced water and natural gas in a time series manner based on well completion and placement decisions. The unsupervised learning aspect established groupings of related wells, enabling a straightforward method to deduce Arps Decline, well completion, and reservoir and spatial attributes characteristic of each cluster group. The ensemble of the supervised and unsupervised elements of this work facilitates a means to forecast oil, water, and natural gas production at the well level as influenced by specific development considerations. Well level three-stream production volumes can be used to scale up outlooks at the pad, field, or basin-level (as demonstrated in Section 4.5). The framework has been applied to the producing extent of the "Wolfberry" within the Midland Basin. However, since the overall analytical approach is based on readily available datasets common to public sources, it could be easily modified for use in other mature unconventional O&G producing regions.

Major environmental concerns regarding shale O&G development are associated to water usage, induced seismicity via wastewater disposal, and flaring (and possible venting) of produced

natural gas. The framework presented in this study can be leveraged to help support the formulation of management and/or remedial strategies based on the volumes of fluids expected from unconventional O&G development operational conditions. Study results have highlighted the variability in noted water and gas volumes produced depending on wellbore design and placement considerations – a finding which suggests forecasting is a nontrivial task. Table 18 provides quantitative insight that can reduce the burden in estimating associated fluid production for future wells. Data compiled in Table 18 summarizes the potential volumes of produced fluids associated with oil production across the study area given well completion design considerations and placement within the basin. These data can be used to build out three-stream fluid production outlooks for the Midland Basin. Forward-looking production outlooks for oil, water, and natural gas as highlighted in Figure 42 are highly dependent on the nature of well design and placement considerations of the subsequent fleet of wells (as well as legacy production from existing wells). However, many of these design choices that would determine the composition of the out-year fleet of wells can be strongly influenced by external economic or market-driven factors.

Potential follow-on work could be beneficial in addressing possible limitations and imposed constrains in the research presented here. For instance, a potential area for improvement to the study in regards to the model development pertains to limited access to geologic data which could be used as inputs. Readily available geologic data at the well level in large volumes is uncommon. Nevertheless, the inclusion of additional geologic characteristics that are known controlling factors to unconventional oil and gas recovery [26] may provide added utility in data-driven ML modeling. Additionally, our study was without access to key time series data pertaining to how wells were operated (i.e., choke, bottom-hole pressure, lift type). The result of which presents a challenge in integrating the human element as part of forecasting component. In regards

to forecasting oil production, gradual or abrupt changes in the producing rate of a well due to reservoir depletion, fluctuation in bottom-hole producing pressure, and changes in conditions in or immediately adjacent to the wellbore are not directly considered when using the Arps models alone. Lastly, potential model performance improvement gains might be realized thorough the development of separate models for predicting monthly water and gas individually instead of in joint fashion.

5.0 Machine Learning Classification Approach for Formation Delineation at the Basin-Scale

5.1 Chapter Summary

Machine learning and artificial intelligence approaches have rapidly gained popularity for use in many subsurface energy applications. They are seen as novel methods that may enhance existing capabilities, providing for improved efficiency in exploration and production operations. Furthermore, their integration into reservoir management workflows may shape the future landscape of the energy industry. This study implements a framework that generates predictive models using multiple machine learning classification-based algorithms which can identify specific stratigraphic units (i.e., formations) as a function of total vertical depth and spatial positioning. The framework is applied in a case study to 13 specific formations of interest (Upper Spraberry through Cisco/Cline [Wolfcamp D] reservoirs) in the Midland Basin, West Texas, United States; a prominent hydrocarbon producing sub-basin of the larger Permian Basin. The study dataset consists of over 275,000 records and includes attributes like formation identifier, true vertical depth (in feet) of formations observed, and latitude and longitude coordinates (in decimal degrees). A subset of 134,374 data records were relevant to 13 distinct formations of interest and were extracted and used for machine learning model training, validation, and testing. Four supervised learning approaches including random forest (RF), gradient boosting (GB), support vector machine (SVC), and multilayer perceptron neural network (MLP) were evaluated and their prediction accuracy compared. The best performing model was ultimately built on the RF algorithm and is capable of an overall prediction accuracy of 93 percent on holdout data. The RF-based model demonstrated high prediction accuracy for major oil and gas producing zones

including the San Andres, Upper Spraberry, Lower Spraberry, Clearfork, and Wolfcamp at 98, 94, 89, 94, and 94 percent respectively. Overall, the resulting data-driven model provides a robust, cost-effective approach which can complement contemporary reservoir management approaches for multiple subsurface energy applications.

5.2 Introduction

Fit for purpose development and deployment of machine learning-based technology for subsurface resource applications (e.g., conventional and unconventional oil and gas [O&G], liquid waste disposal, geothermal energy, and geologic carbon dioxide [CO_2] storage) has the potential to provide accurate, efficient, and cost-effective analytical compliments to conventional reservoir management strategies. Such approaches may transform future subsurface energy resource exploration, utilization, and management into a much more data-driven science [35, 40]. Machine learning tools use statistical techniques that can "mine" through data and potentially uncover hidden patterns or relationships within large, complex, multivariate datasets [42, 43, 31] that may otherwise go undiscovered. The resulting insights gained from deploying machine learning as a novel compliment to reservoir management may therefore enable better understanding of engineered subsurface system performance: thereby offering reduced risk, improved safety, and increased effectiveness of developing said subsurface resources [32, 43].

A number of potential use cases for machine learning as part of subsurface energy systems management and decision making have been noted in literature, spanning topics like improving oil and gas production [16, 285], modeling CO_2 injection and potential leakage [286, 287], informing drilling practices [17, 18], uncertainty quantification for field site monitoring [288, 289], and geologic formation, stratigraphy, and lithology classification or inversion [290, 291]. The latter of which related to geologic formational classification is the focus of the research discussed in this study.

Delineation of the burial depth, thickness, and lateral extent of specific geologic strata is a crucial undertaking when exploring for and ultimately developing subsurface energy resources [292]. Typically, each geologic formation is in some way unique and distinct in terms of

lithostratigraphic characteristics from neighboring formations [293, 294]. The specific depth, thickness, and lithostratigraphic characteristics of a given formation(s) of interest, along with those associated with overburden and possibly underburden strata, strongly influence operational decision-making. For instance, planning well drilling operations and logging, well completion engineering and design choices, subsurface resource quantification estimation (i.e., oil and gas productivity potential, CO_2 storage or liquid waste disposal resource capacity and containment amenability, geothermal potential, etc.), and project cost estimation are all dependent on the geologic sequencing and their associated areal extents at new well sites.

Determination of the lateral and vertical distribution of geologic formations typically occurs during the exploration and delineation phase of subsurface resource development. Information gleaned from this step can help determine where potentially viable sites exist. Technical information considered during this step may come from a variety of existing data sources, including production or injection data from existing wells, reviewing data from existing core samples, assessing available seismic surveys, analyzing well logs, reviewing records and sample descriptions from existing or plugged and abandoned wells, and reviewing other geologic data available in literature [295]. Once new a well site(s) is determined, the vertical formational extent of geologic formations present are confirmed from various methods, including rate of penetration (ROP) charts during drilling, well log analysis, or drill cutting and mud logging [290]. This information aids in establishing proper well casing and perforation placement to ensure proper zonal isolation as needed, as well as informing geologic model development.

The methods described above have their own advantages and limitations regarding their associated costs, accuracy attained, manpower resources, potential waste products generated, and time needed to implement [290]. Additionally, geologic formational sequencing and associated

lithostratigraphic characteristics are highly spatially dependent. Over larger spatial domains (e.g.; field to basin scale) interpolation approaches are known to be highly uncertain [77]. The currently deflated market prices in the O&G sector encourages operators pursuing any subsurface energy resource type to become more operational efficient and cost-effective [98]. Consequently, formational delineation is one aspect of subsurface resource exploration and development that may benefit from new approaches that provide efficient and cost-effective alternatives to more expensive exploration strategies. As mentioned, machine learning has emerged as promising option for operators and researchers alike to consider in this regard [38, 99]. Particularly, the recent growth in unconventional O&G development has prompted a simultaneous escalation in the types and volumes of data generated amenable for machine learning analyses for subsurface energy applications [53].

In this study we develop predictive models using multiple machine learning classification-based algorithms which can identify specific stratigraphic units (i.e., formations) as a function of total vertical depth and spatial positioning. The models rely on straightforward data types that would be common to basins with relatively mature O&G development. The basic framework could be directly applied to any basins where similar data exists. Machine learning model development in this study focuses on the Permian Basin region of the U.S; primarily in the Midland Basin portion of the Permian. The region holds enormous consequence regarding domestic oil and gas production. A report by the Texas Independent Producers & Royalty Owners Association indicates that yearly crude oil production in the Permian Basin has grown by 1.2 billion barrels since 2009, resulting in a 371% increase in oil output over the last ten years [184]. This overall growth was facilitated by the application of horizontal drilling and hydraulic fracturing technologies and has enabled the Permian to become the world's top-producing oil field [185]. While the region itself

major producer of both oil and gas, the basin still faces several challenges regarding 1) improvements needed in production efficiency and 2) associated natural gas and water production management [186, 187, 183].

The resulting data-driven model helps inform future decision-making by learning from previous development. The model provides a robust and cost-effective approach which can supplement contemporary reservoir management best-practices, spanning multiple subsurface applications, including:

- Delineation of over and underburden expected at new drill sites which can aid in well drilling planning, like efficient drill bit management and fluid selection;
- Delineation of over and underburden expected at new drill sites which can aid in well completion and design considerations, like optimizing casing and cement placement;
- Ability to develop geologic formation cross-sections for a given basin(s) using a consistent stratigraphic framework without extensive manual interpretations from well logs;
- Facilitate the development of structural mapping like isopachs, depth to formation top, and depth to base of formation contours – data that can complement static geomodel development;
- Real-time synchronization with geo-steering techniques to help ensure optimal wellbore placement; and
- Integration into subsurface visualization applications [296, 297, 298, 299, 300, 301] to depict stratigraphy and other geologic properties in three dimensions.

5.3 Materials and Methods

The machine learning-based prediction models (called "Formation Labeler" from this point) were developed using a Formation Tops dataset acquired from oil and gas data vendor DrillingInfo [302]. The workflow used to develop the Formation Labeler model is presented in Figure 43. Ultimately, several machine learning models were systematically developed using various algorithms and training data configurations in order to achieve a model with the highest classification prediction accuracy as possible.

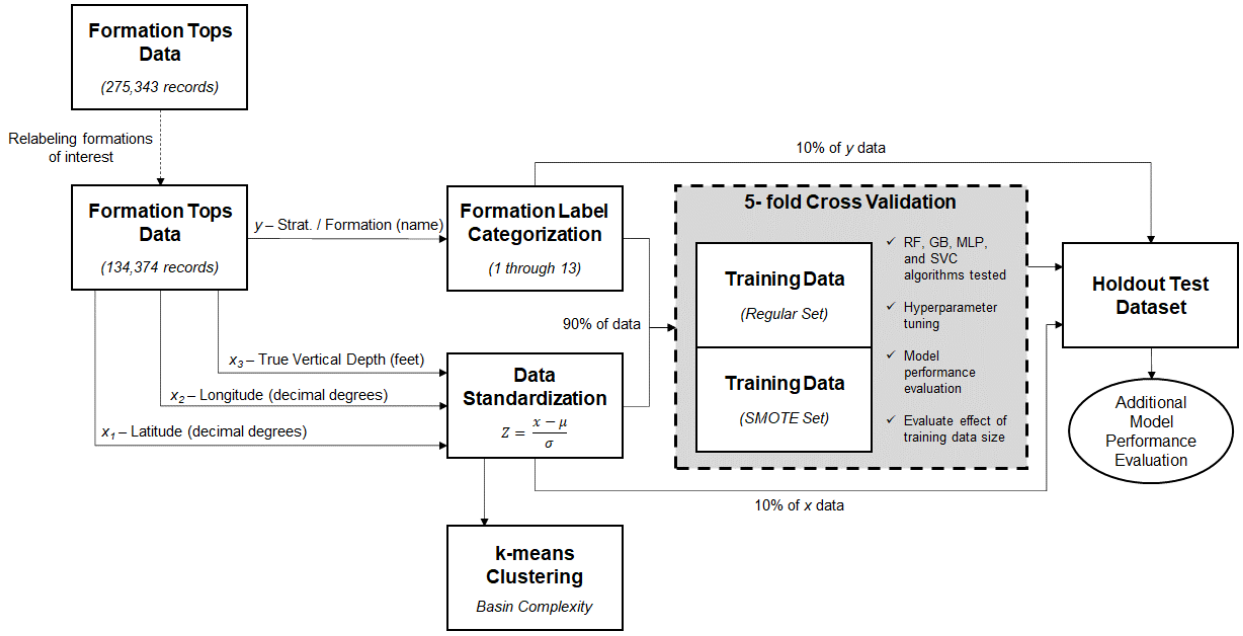

Figure 43. Workflow implemented to develop the Formation Labeler Model. Random forest = RF; GB = gradient boosting; MLP = multi-layer perceptron neural network; SVC = support vector machine classification; SMOTE = Synthetic Minority Oversampling Technique.

5.3.1 Study Data

The Formation Tops dataset contains over 275,000 records specific to the Midland Basin. The dataset is primary a large list of formation identifiers at given well locations and the associate depth the formations were noted. Specifically, the Formation Tops dataset includes data attributes like formation identifier name (string), associated well API numbers where formations were noted (integer), the true vertical depth (in feet) (integer) of formations observed, and latitude and longitude coordinates (in decimal degrees) (float). Information within the Formation Tops dataset could be generated from field data derived from multiple source types, including from well log interpretation, evaluating drill cuttings and mud logging, and simply from interpretations published in open literature. The Formation Tops dataset required extensive relabeling into distinct and consistent formational nomenclature that aligned with the 13 Stratigraphic / Formation Name groups of interest within the Midland Basin per Figure 44. These stratigraphic units include a combination of both oil and gas producing and non-producing intervals (the prominence of each as a hydrocarbon producer can vary by location within the basin), and include the "Wolfberry" pay zones (highlighted in Upper Spraberry through Cisco/Cline [Wolfcamp D] reservoirs in Figure 44). The resulting dataset of the distinct formations were used to fit and test multiple classification-based machine learning models using different algorithms to generate the Formation Labeler model.

Era	Period	Epoch	Local Series	Stratigraphic / Formation Name	Operational Name	Prominent Hydrocarbon Producing Formation [198, 303]
Paleozoic	Permian	Guadalupe	Ward	San Andreas	San Andreas	●
				San Angelo / Glorieta	San Angelo / Glorieta	
		Leonard	Clearfork	Clearfork	Upper Leonard	
				Tubb	Tubb	
				Wichita Albany	Wichita Albany / Lower Clearfork	
			Wichita	Upper Spraberry	Spraberry	●
				Lower Spraberry		●
				Dean		●
			Lower Leonard	Wolfcamp	Wolfcamp A	●
		Wolfcamp	Wolfcamp		Wolfcamp B	●
					Wolfcamp C	
	Pennsylvanian	Virgil	Cisco / Cline		Wolfcamp D	●
		Missouri	Canyon		Canyon	●
		Des Moinesian	Strawn		Strawn	
		Atokan	Atoka / Morrow		Atoka / Bend	●
	Mississippian	Morrow			Morrow / Bend	●

Figure 44. Stratigraphic description for a subset of Midland Basin, Texas relevant to the stratigraphic / formation names of interest to this study. The figure was amalgamated from lithostratigraphic interpretations from several literature sources **[194, 195, 196, 197, 198, 192]**.

The areal extent of the relabeled Formation Tops dataset in the Midland Basin, and broader Permian Basin, is plotted in Figure 45. The Permian Basin is a complex sedimentary system that covers an area of more than 75,000 square miles in portions of West Texas and Southeast New Mexico. The basin consists of a number of sub-basins and platforms, including the Delaware Basin, Central Basin Platform, and the Midland Basin [192]. The regional extent of the Central Platform and Midland sub-basins is shown in Figure 45. The Midland Basin (the study area for this research) is constrained to the east by formational shallowing towards the Eastern shelf and to the west by folding and faulting on the eastern portion of the neighboring Central Platform. In its

southern portion, the Midland Basin formations start to thin in an extension of Central Basin Platform called the Ozona Arch [193]. The Permian's Northern shelf limits the Midland Basin's extent to the north. The basin is also deepest on its western flank and shallower towards the east; in many portions stratigraphic units experience several thousand feet of relief across the basin.

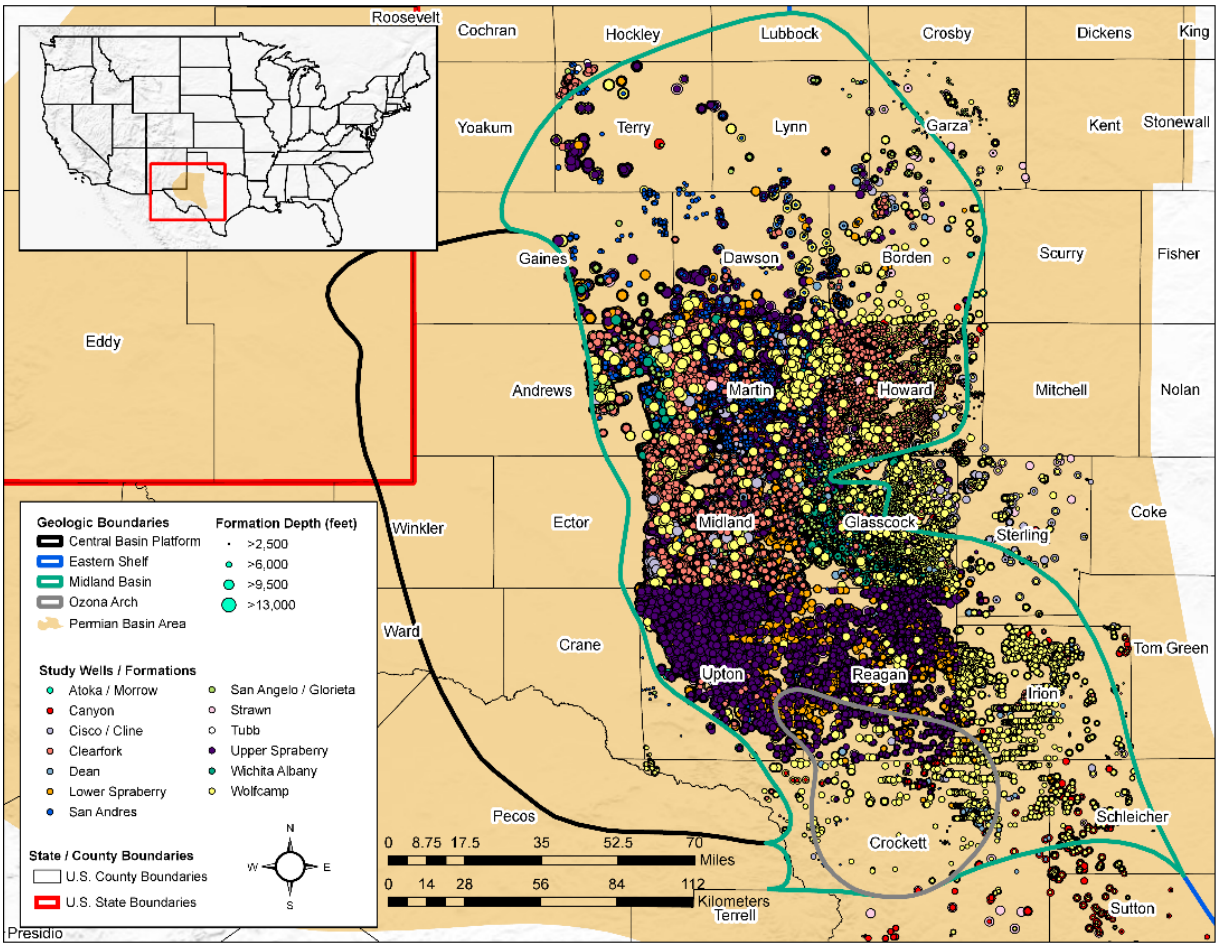

Figure 45. Map of the study area in the Midland Basin, Texas. Well data evaluated as part of the Formation Tops dataset where noted observations for each formation of interest had occurred are presented. Geographic information system (GIS) layers used to create this figure were acquired from the University of Texas at Austin **[216]** and United States Geological Survey **[217]**.

5.3.2 Machine Learning Approaches Applied

The Formation Labeler model uses latitude, longitude, and true vertical depth as input attributes (*x* variables) and predicts the specific formation (*y* or response variable – limited to the Stratigraphic / Formation Name in Figure 44) given those input conditions. The workflow that includes the application of multiclass classification machine learning is intertwined with data preprocessing steps (following formation relabeling) that includes data standardization, sub-division, and augmentation as shown in Figure 43. Input attributes are standardized to z-values (*Z*) per Equation 5-1.[11]

$$Z = \frac{x - \mu}{\sigma} \qquad \text{Equation 5-1}$$

Where:

x = input attribute value
μ = input attribute mean value
σ = input attribute standard deviation

The input dataset then undergoes a preliminary data evaluation step using *k*-means clustering [251]. This step aims to determine the optimal number of clusters for the input dataset as a means to evaluate dataset complexity spanning across the basin (influenced mostly by the dimensions of the spatial extent and burial depth of formational observations in the study area).

[11] Two of the four machine learning algorithms applied are decision tree-based models; Random Forest and Gradient Boosting. Since decision trees are invariant to different input variable scales, training data was not standardized for Random Forest and Gradient Boosting.

The optimal number of clusters occurs when adding additional clusters results in no substantial improvement to the within-cluster sum of squared error. Two heuristic algorithms are applied to determine the optimal number of clusters - the Elbow method [253] and Hartigan's Rule [254]. The optimal number of clusters is compared against the number of stratigraphic / formation groupings (13 total) as an independent data quality assurance step / comparative analysis for the formation relabeling effort. A large noted discrepancy between optimal clusters and the number of discrete stratigraphic / formation groups used may suggest that a more extensive formation relabeling effort that generates more expansive formation groupings is warranted prior to classification-based machine learning model development.

The input and response variable datasets are sub-divided into training and test datasets through a 90/10 percentage-based split. A data augmentation step was applied to the training dataset (not the holdout test dataset) prior to cross-validation, resulting in two separate training dataset options that were ultimately evaluated as part of model development. This step applied a Synthetic Minority Oversampling Technique (SMOTE) which involves oversampling data examples in minority classes as a way to reduce data imbalance across the 13 stratigraphic / formation groups. The augmented data examples do not add new information to the dataset, but their addition is intended to help classification models more effectively learn decision boundaries between groups, and therefore improve in accuracy [304]. The resulting SMOTE training dataset was expanded by a roughly 19 percent increase compared to the original training dataset (described as Regular Set in Figure 43). SMOTE manipulation was implemented using the Imbalanced-learn Python library [305].

A *k*-fold cross-validation approach [306] using 5-modeling folds was then implemented on the training datasets (Regular and SMOTE) as part of model training. The cross-validation

approach was implemented in two steps. The first step was used to evaluate different combinations of hyperparameters for each machine learning algorithm evaluated. An exhaustive grid search approach was used where different models are built on the training data for the distinctive hyperparameter combinations considered specific to each algorithm employed [89]. Model performance is then evaluated against the holdout data in each fold. For simplicity, the Regular training dataset was used as part of hyperparameter tuning step for each machine learning algorithm evaluated. The model formulation for each algorithm resulting from the hyperparameter combination generating the most accurate model was selected as the finalized model formulation for that algorithm.

The second cross-validation step involved evaluating final model formulation performance on each of the five validation folds. Separate models using each algorithm / optimal hyperparameter combination formulation underwent cross-validation on both the Regular and the SMOTE datasets and the performance of each model was noted against the holdout data for each of the five folds. As a result, a total of eight Formation Labeler predictive models (built on four different algorithms and two different training datasets) were generated as part of this study. Finalized model performance was conducted on the 10 percent subset holdout test data that was not used in cross-validation as a secondary evaluation of model performance accuracy.

The accuracy metric [307] was used to evaluate model prediction performance in this study and is described in Equation 5-2.

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \qquad \text{Equation 5-2}$$

Where:

y_i = response attribute true value of the *i*-th observation
$\widehat{y_i}$ = response attribute predicted value of the *i*-th observation
$n_{samples}$ = total samples

The accuracy metric (Equation 5-2) was used as the performance metric in several components of the model development workflow, including the hyperparameter tuning step and to compare across the finalized model formulations generated from the different machine learning algorithms employed (both in cross-validation and against the test dataset).

The performance of various machine learning algorithms was compared in their prediction of stratigraphic / formation names. In total, four different supervised machine learning algorithms were evaluated, including random forest (RF), gradient boosting (GB), multilayer perceptron-based neural network, and support vector machine (SVC). Given the nature of this research, all algorithms were applied for solving multiclass classification. A basic description of each algorithm is provided in the bullets below; the foundational mathematics for each can be readily found elsewhere in literature:

- **Random Forest (RF)** – An ensemble-based learning method proposed by Breiman (2001) in which randomized decision trees are constructed in parallel on bootstrapped training samples and their individual predictions aggregated into a single response [308, 87]. In this study, the RF hyperparameters tuned as part of cross-validation included 1) the number of trees in the forest; 2) the minimum number of samples required to split an internal node, 3) the maximum tree depth (i.e., limits the number of nodes in each tree); and 4) with and without bootstrap sampling when building trees. Optimum values were found to be 1,250 trees, a

minimum of two samples to split an internal node, a maximum tree depth of 20, and the use of bootstrapping.

- **Gradient Boosting (GB)** – Gradient boosting algorithms produce a finalized prediction model in the form of an additive ensemble of weak prediction models, typically decision trees. The model is built in sequential fashion (i.e., boosting approach) where new decision trees are fit to prior stage model residuals in a greedy fashion. The newly added tree attempts to minimize loss given the previous ensemble of the model [68, 81, 80]. Ultimately, the final model is a linear combination ensemble of every weak prediction decision tree. A shrinkage parameter (v; where $0 \leq v \leq 1$) sets a learning rate. This parameter controls the contribution of each tree to minimize the loss function within the final model. Smaller values ($v \leq 0.1$) tend to result in improved model performance [81, 68] but at the cost of requiring a greater number of decision trees and potentially a larger computational expense to fit the final model. In this study, the GB hyperparameters tuned as part of cross-validation included 1) the number of trees as weak learners; 2) the minimum number of samples required to split an internal node, 3) the shrinkage parameter v; and 4) the maximum tree depth. Optimum values were found to be 1,000 trees, a minimum of two samples to split an internal node, v set to 0.1, and a maximum tree depth of 15.
- **Multilayer perceptron neural network (MLP)** – Neural networks are machine learning algorithms that have interconnected neurons used to reconstruct complex nonlinear input and output relationships—considered analogous to synapses in the human brain [309]. Each neuron consists of a network of nodes and weighted links

(i.e., synaptic connections) that connect predictor variables to response variables [310]. Neural networks are specified by network structure (like the number of hidden layers and neurons within each), activation functions, and training algorithms [311]. Different variants of neural networks exist, but multilayer perceptron (MLP) is one of the more widely used [312, 15]. MLP is a class of feedforward artificial neural network where at least three layers of nodes exist: (1) an input layer, (2) one or more hidden layers, and (3) an output layer. With the exception of the input nodes, every other node is considered a neuron and is related to other neurons in the preceding layer. Every neuron implements a nonlinear activation function that defines the output of that particular neuron, given an input or set of inputs. The neuron output is then used as input for the next set of neurons in the following layer. The signal passing through a given neuron via feed-forward propagation is adjusted by weights that alter the functions and nonlinear activation which alters the combined output signal that eventually reaches neurons in the following layer [309]. Modification to the weights occur via gradient decent through back-propagation of error between prediction and response. This process continues through multiple iterations (i.e., epochs) until a desired solution that results in a suitable error rate is achieved [313]. In this study, the MLP hyperparameters tuned as part of cross-validation included 1) alpha hyperparameter, which serves as an L2 weight regularization function [314]; 2) the number of iterations or epochs; 3) activation function types; 4) the number of hidden layers; and 5) number of neurons per hidden layer. Optimum values were found to be alpha at 0.0001, 1,000 epochs, the hyperbolic tan activation function

(tanh), four hidden layers, and 100 neurons per hidden layer. Additionally, the adam stochastic gradient-based optimizer was used for weight adjustment. Early stopping was applied using a validation fraction of 10 percent and a loss tolerance of 0.005.

- **Support vector machine (SVC)** – The concept of SVC was introduced by Cortes and Vapnik (1995) [315]. The concept is based on constructing a hyperplane or a set of hyperplanes in a multi-dimension feature space. The hyperplane(s) enable(s) categorization of new data depending on that data's given position in relation to the hyperplanes(s) within the multi-dimensional space. Good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (those data are called support vectors) [316]. Kernel function options such as linear, polynomial, and radial basis function (rbf) are used to map the input vectors into high-dimension feature spaces [317, 318]. The Gaussian rbf kernel was used here. The gamma (γ) parameter controlling the width of the Gaussian function [319] was set to scale based on the number of input features and variance of those features in the input dataset [320]. In this study, the SVC hyperparameters tuned as part of cross-validation included 1) the soft margin cost function parameter (C). The optimum value was found to be 10,000 for C.

All models have been developed by leveraging Python's scikit-learn library [79] on Python 3.

5.4 Results and Discussion

This section summarizes key results related to the Formation Tops relabeling effort, the findings from the optimal cluster count determination through *k*-means, and the Formation Labeler

model performance. The results discussed throughout this section indicate that the Formation Labeler model for the Midland Basin performs well at predicting each of the 13 stratigraphic / formation groups given the spatial positioning within the basin as well as true vertical depth below ground surface.

5.4.1 Formation Label Categorization

The formation relabeling effort resulted in a substantial reduction in the overall data from the original Formation Tops dataset that was ultimately used going forward. Out of the full 275,343 records, 134,374 had formation identifiers in the original Formation Tops dataset that were sufficiently distinctive and could be confidently relabeled into one of the 13 stratigraphic / formation groupings. However, the nomenclature used for the original formation identifiers was highly disparate. Out of the 134,374 records relabeled, a total of 902 unique label identifiers existed from which the 13 stratigraphic / formation groupings were derived.[12]

Table 20 provides a statistical overview of the relabeled dataset grouped by distinct stratigraphic / formation name and sorted by average depth across the basin. For each stratigraphic / formation name, Table 20 summarizes the standard deviation for all observations (influenced by formation thickness extent and burial depth across the basin), the observation count, and the

[12] To provide additional context, a few examples of the formation identifiers in the original Formation Tops dataset that were mapped to the relabeled distinctive stratigraphic / formation names in Figure 44 are provided here:

- Upper Spraberry (Original identifier examples: Top Upper Spraberry, T. Spraberry, Upper Sprayberry, Top of Upper Spraberry)
- Dean (Original identifier examples: Dean, Top of Dean, Base of Dean, Top Dean, Dean Sand)

percent of total observations. The distribution of observations indicates that the dataset suffers from imbalance.

Table 20. Summary statistics of the Formation Tops dataset used for the study following relabeling. A total of 134,375 formation observations from over 32,800 specific wells were utilized as part of the study.

Stratigraphic / Formation Name	Average Formation Depth (feet)	Depth Standard Deviation (feet)	Observation Count	Observation Count Percent of Total
San Andres	3,801	1,071	16,590	12.3%
San Angelo / Glorieta	4,525	1,290	5,984	4.5%
Clearfork	5,557	1,433	13,749	10.2%
Tubb	5,707	1,392	351	0.3%
Wichita Albany	6,636	1,343	1,298	1.0%
Upper Spraberry	7,049	1,078	26,891	20.0%
Lower Spraberry	7,273	1,272	4,781	3.6%
Dean	8,144	1,119	16,849	12.5%
Wolfcamp	8,196	1,330	22,797	17.0%
Canyon	8,224	1,547	2,545	1.9%
Cisco / Cline	9,022	1,351	3,546	2.6%
Strawn	10,130	722	13,229	9.8%
Atoka / Morrow	10,386	682	5,764	4.3%

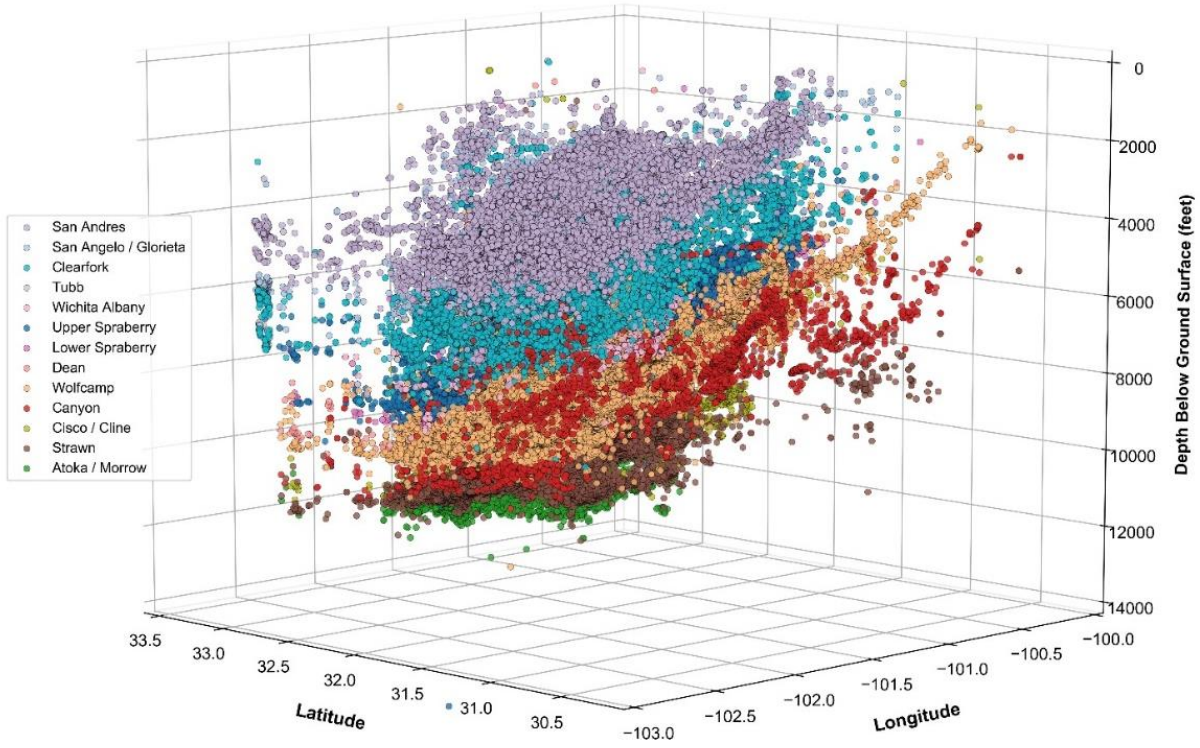

Figure 46. Three-dimensional visualization of the observations from the relabeled stratigraphic / formations of interest.

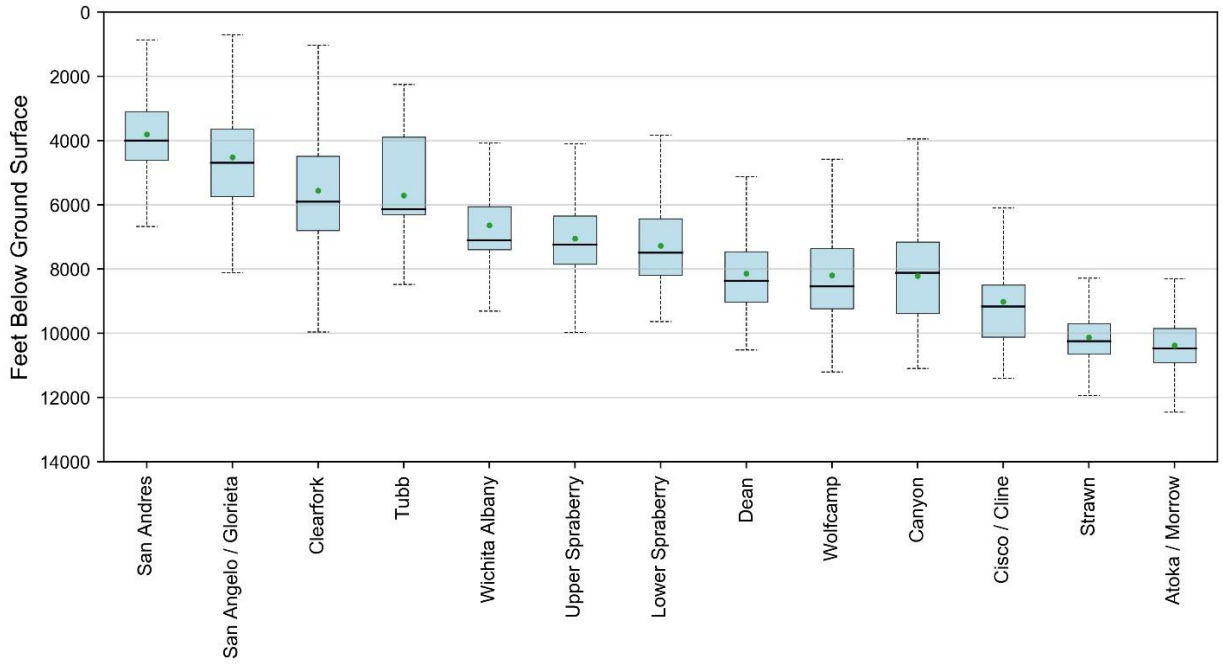

Figure 47. Box and whisker plots of the dataset well observations for each formation of interest as a function of depth below ground surface. The box extends from the 25th to 75th quartile values of the data, with a line at the median (50th quartile). The circle is at the data mean. Whiskers extend to the minimum and maximum values of the data absent outliers.

Figure 46 provides a visual representation of the distribution of the relabeled 134,374 datapoints for each of the stratigraphic / formation groupings based on their spatial coordinates and true vertical depth. The box-and-whisker plots in Figure 47 depict the distribution of the resulting aggregation of observations for each stratigraphic / formation group. These figures indicate that the burial sequencing for the stratigraphic / formation groups as they proceed from shallow to deep follows the stratigraphic description order presented in Figure 44. The one exception is the Canyon formation, which has a shallower average depth than the Cisco / Cline based on dataset observations. However, the observed depths for the Canyon formation across the basin was shown to be more variable on average (based on depth standard deviation in

Table 20 and resulting box and whisker extent in Figure 47); which could attribute to the average depth value. As a result, the relabeling efforts seemed to enable appropriate mapping of data from formation identifiers to stratigraphic / formation groupings.

5.4.2 *k*-means Clustering Analysis

The results from the k-means clustering analysis to determine the optimal number of clusters from the Formation Tops input dataset are presented in Figure 48 using the Elbow method and Hartigan's rule.

The Elbow method evaluates the total within-cluster sum of squared errors (SSE) as a function of the number of clusters. The optimal number of clusters occurs at the point in which adding another cluster does not result in a substantial improvement to the total within-cluster SSE. Hartigan's Rule is based on comparing the Hartigan's Index, which is a ratio between the within-cluster sum of squared error based on k number of clusters to that based on $k + 1$ clusters. The rule utilizes the intuition that when clusters are well separated, the ratio becomes less than 10 (dashed red line in Figure 48 B) and is taken as k to be the optimal number of clusters. The visual heuristic results for the Elbow method suggest between approximately 17 and 23 clusters (i.e., stratigraphic / formation groups) seem appropriate (Figure 48 A). The Hartigan Solution in Figure 48 B however indicates that 21 clusters are ideal, and that adding the 22nd cluster (where 22 is the $k + 1$ where the Hartigan Index ratio between k and $k + 1$ is < 10) results in negligible reductions to SSE.

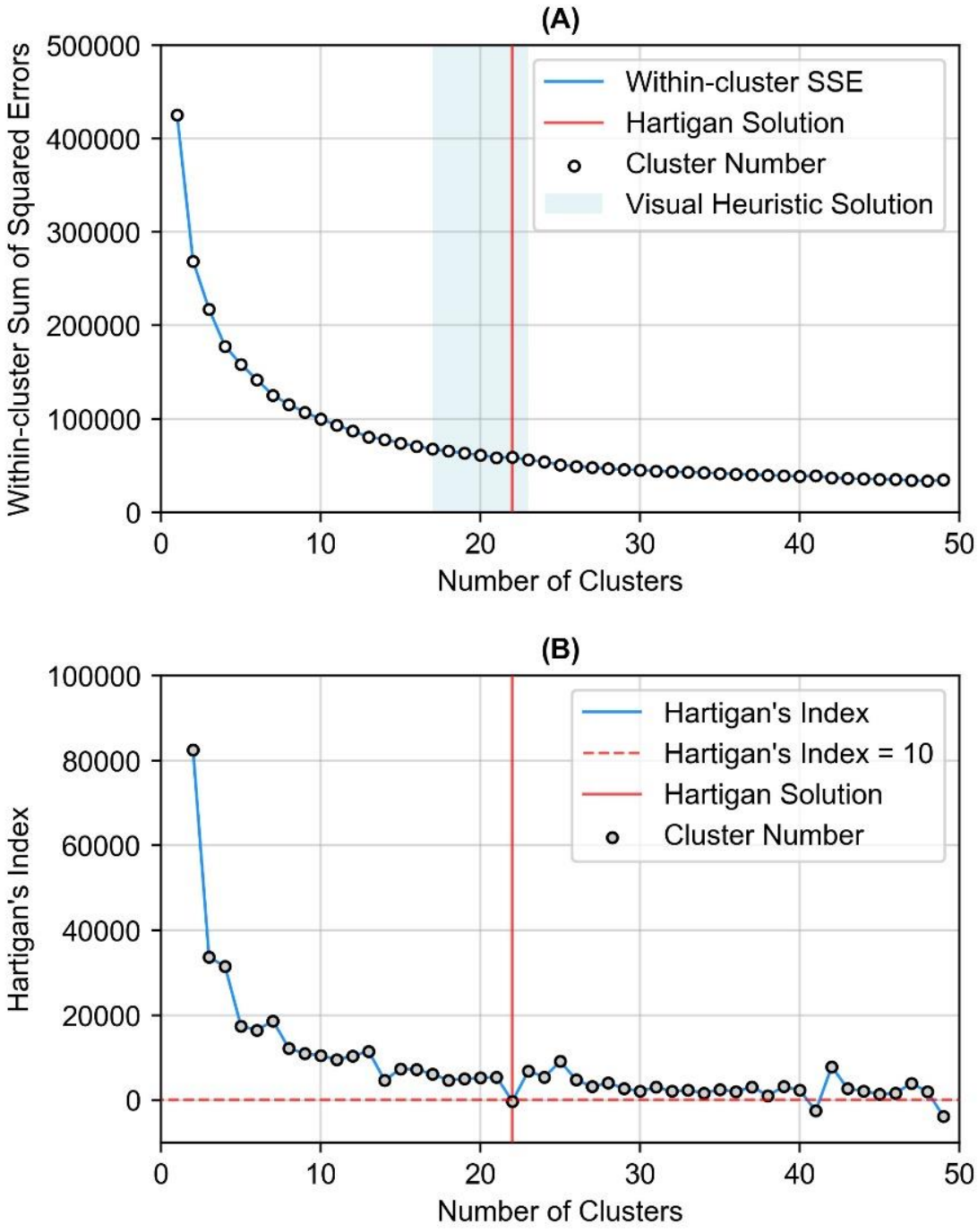

Figure 48. Elbow diagrams from k-means clustering results. The top figure (A) represents the total within-cluster sum of squared errors based on the number of clusters evaluated. The lower figure (B) represents the resulting Hartigan's Index based on the numbers of clusters evaluated.

The optimal number of clusters determined through this exercise via Hartigan's rule (21) is not extensively different from the number of stratigraphic / formation groupings (13 total) used for distinctive reservoir labeling. However, this finding suggests that the finalized Formation Tops dataset relabeling effort could possibly benefit from a more granular labeling schema to separate the stratigraphic / formation groupings into a higher number of groups prior to the machine learning model development step.

5.4.3 Formation Labeler Model Performance

During model cross-validation, final model formulation performance was evaluated on each of the five validation folds once the optimal hyperparmeter settings were determined for each algorithm. Finalized models were trained on both the Regular and the SMOTE datasets and the performance of each model was noted against the holdout data for each of the five folds; totaling eight different Formation Labeler predictive models being generated. The accuracy results are presented in Figure 49, in which the box and whisker plots show the results from the 5-fold cross-validation using the best hyperparameter combinations for each model. The prediction accuracy against the 10 percent holdout test data is also presented as an "X" in the figure. The results demonstrate that all models are capable of greater than an overall 85 percent prediction accuracy against holdout data (for either the 5-fold cross-validation performance assessment or against the test dataset). Additionally, the augmented SMOTE training dataset afforded models greater prediction accuracy (limited to the 5-fold cross-validation). However, the overall best performing model was determined to be the RF model developed on the SMOTE dataset, which was capable of an overall prediction accuracy of 93 percent on holdout data.

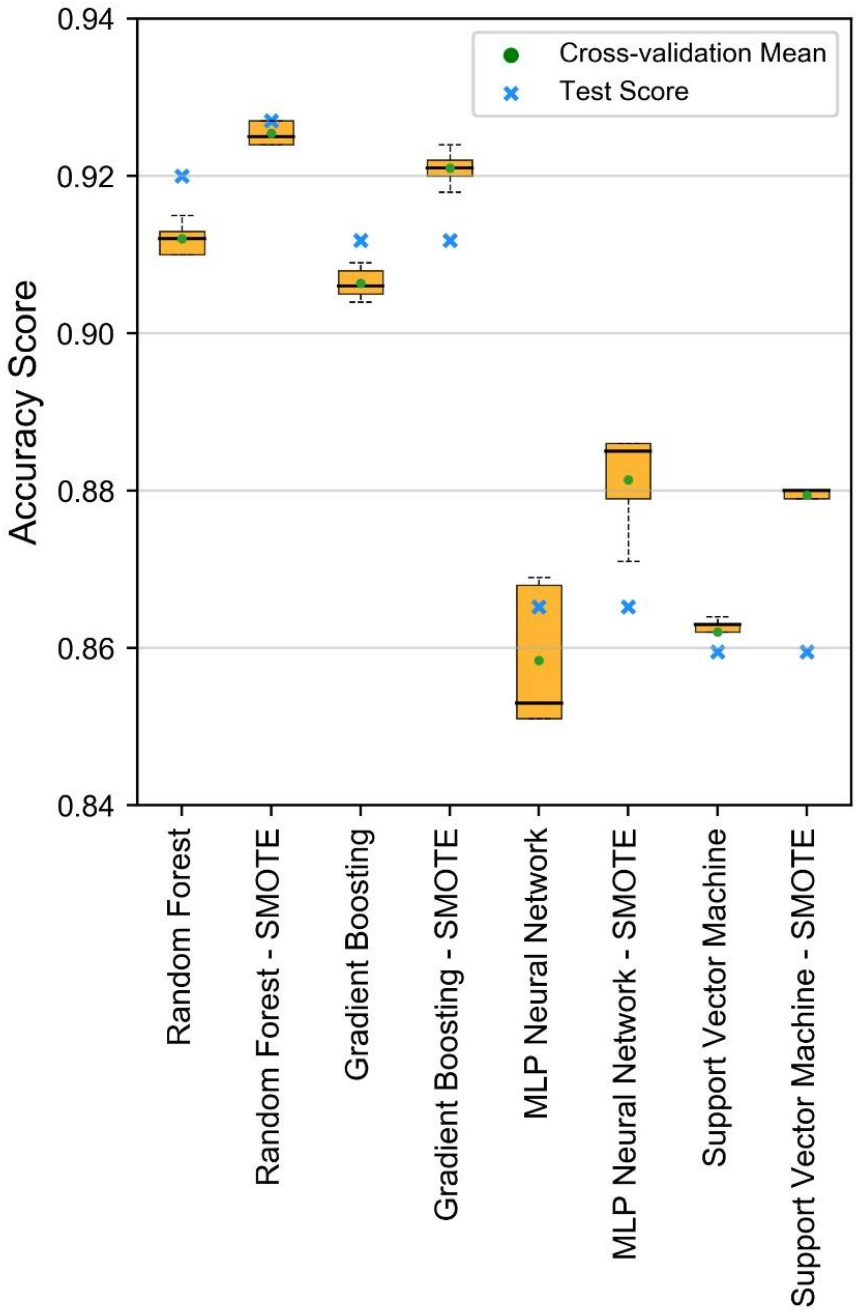

Figure 49. Box and whisker plot of the Formation Labeler classification model performance under various algorithms and training datasets. The box extends from the 25th to 75th quartile values for prediction accuracy across each of the five folds from the cross-validation step. The line is at the median value (50th quartile) and green circles are at the mean. Whiskers extend to the minimum and maximum values of the data absent outliers. The blue "X" represents the prediction accuracy on the holdout test data.

Since the RF model trained on the augmented SMOTE training dataset was the best performing model, it was further explored and evaluated. The RF model's prediction results against the test dataset are further analyzed in Figure 50, which presents a confusion matrix of the predicted formations by the model against the test dataset compared to the actual formation groupings. Figure 50 highlights how well the RF model was at predicting each of the 13 stratigraphic / formation groups by quantifying the predicted fraction between the predicted versus actual formation within the test dataset. Additionally, Figure 50 shows the fraction of misclassification predictions for each associated true formation. The model was most accurate in predicting the San Andres Formation (average depth of 3,801 feet with a standard deviation 1,071 feet per Table 20) and least accurate predicting the Wichita Albany Formation (average depth of 6,636 feet with a standard deviation of 1,343 feet per Table 20). The fraction of misclassifications when Wichita Albany was the true formation included predicted fractions of six percent for the Clearfork, eight percent for the Dean, two percent for the Lower Spraberry, and five percent for the Wolfcamp. Major oil and gas producing zones in the Midland Basin including the San Andres, Upper Spraberry, Lower Spraberry, and Wolfcamp were accurately predicted at fractions of 98, 94, 89, 94 and 94 percent respectively.

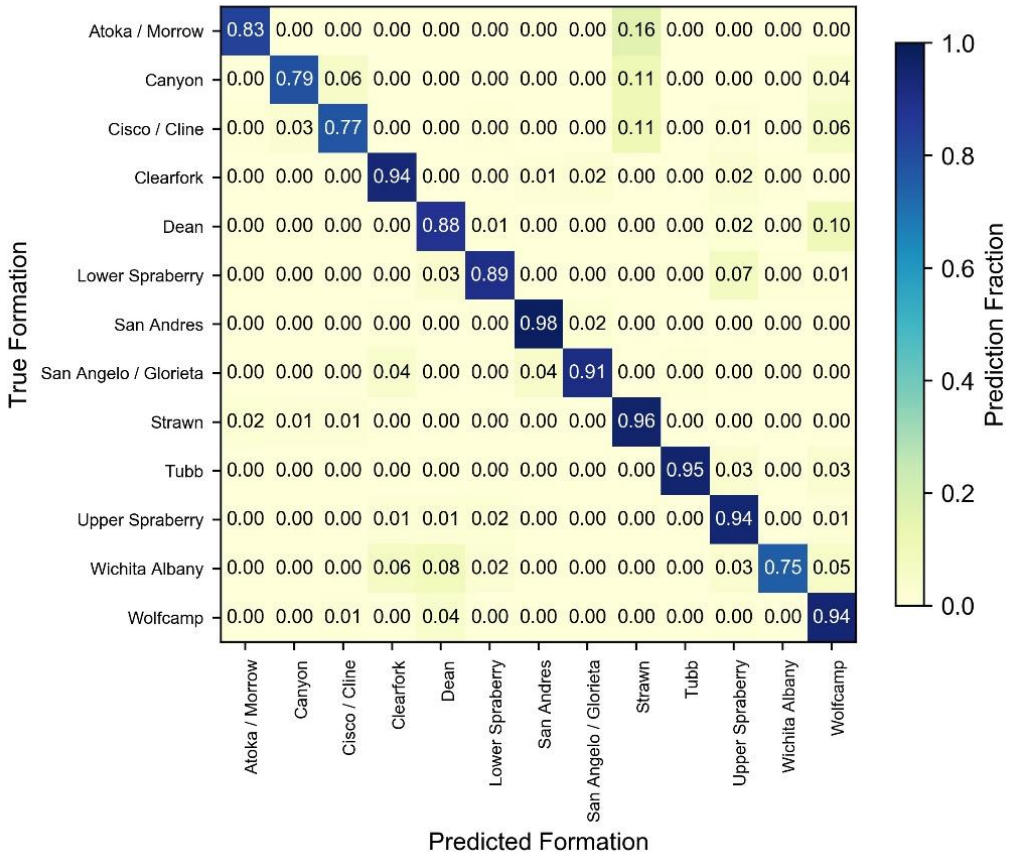

Figure 50. Confusion matrix of prediction accuracy against the test dataset using the random forest model trained on the SMOTE dataset.

5.4.4 Case Study Evaluation

The results presented in Section 5.4.3 indicate that the finalized Formation Labeler model is capable of accurate prediction on holdout test data. In this section, the model was applied in a case study that compares the Formation Labeler's prediction of stratigraphic / formation groups at known in-field well sites to those where the stratigraphy has been interpreted by oil and gas professionals and made available in literature.

For this example, two in-field wells within the Midland Basin study area were selected for evaluation. Both wells are located in Reagan County, Texas. In each case, the literature sources leveraged provide interpretations of stratigraphic columns via wireline logging analysis. Additionally, the spatial coordinates can be inferred from well API numbers and a reference is provided to the corresponding depth at which stratigraphic / formation groups have been interpreted – all necessary inputs for the Formation Labeler model. The first in-field example came Baumgardner, Hamlin, and Rowe (2014) for the O.L. Greer 2 well and includes interpretation from the Wolfcamp through Strawn Formations [321]. The second is provided by Flumerfelt (2014) from a stratigraphic interpretation of the Saxon Oil Branch well which includes the Lower Spraberry through Atoka / Morrow Formations [279]. Neither of these wells had data that were part of the training or test datasets. The Formation Labeler model was used to predict the stratigraphic / formation group as a function of depth at both the O.L. Greer 2 and Branch well sites (predictions occurred on 20-foot interval resolution).

Table 21 provides a comparison of the Formation Labeler's predicted stratigraphic / formation groups at each of the known in-field well sites to those determined by interpretation. For the O.L. Greer 2 well, the predicted depth for the Wolfcamp identically matched the interpreted depth via logging analysis. However, some discrepancy exists between predicted and interpreted for the deeper Cisco / Cline and Strawn Formations. For the Branch well, the Wolfcamp and Strawn Formations have less discrepancy between predicted and interpreted depths compared to other known interpreted formations at that well site. The top of the Lower Spraberry between interpreted and predicted also closely align.

Table 21. Comparison of the Formation Labeler predicted top depth for various Stratigraphic / Formation Groupings versus those interpreted from well logs at two specific in-field well locations. The "--" indicates the grouping was not included in the specific well interpretation or was not indicated to be present based on the Formation Labeler prediction.

Stratigraphic / Formation Grouping	O.L. Greer 2 Well		Branch Well	
	Interpreted Top Depth (feet)	Predicted Top Depth (feet)	Interpreted Top Depth (feet)	Predicted Top Depth (feet)
San Angelo / Glorieta	--	--	--	4,000 to 4,260**
Clearfork	--	4,140	--	4,780
Tubb	--	--	--	--
Wichita Albany	--	--	--	--
Upper Spraberry	--	5,760	--	6,080
Lower Spraberry	--	6,580	7,050	7,040
Dean	--	7,040	7,850	7,520
Wolfcamp	7,780	7,780	8,050	7,980
Canyon	--	--	--	--
Cisco / Cline	9,620	9,000 – 9,280*	9,950	8,820
Strawn	9,900	9,300	10,200	10,140
Atoka / Morrow	--	--	10,350	10,640

*The prediction for the top of the Cisco / Cline spans a depth of 9,000 – 9,280 feet and is interbedded with the Wolfcamp Formation

**The prediction for the San Angelo / Glorieta spans a depth of 4,020 and 4,280 feet and is interbedded with the San Andres Formation

For both well instances, predicted results diverged most substantially from the interpreted depths for the Cisco / Cline Formation (i.e., Wolfcamp D). Per results in the confusion matrix in Figure 50, the Formation Labeler model most commonly misclassifies the Cisco / Cline as either the Wolfcamp (six percent of prediction fraction) or Strawn (11 percent of prediction fraction); two formations that commonly bound the Cisco / Cline at shallower and deeper burial depths respectively. Additionally, potential discrepancies are likely to originate from operator-specific variability related to both 1) nomenclature choice for geologic horizons within the Midland Basin and 2) formation top selection, determination, and bench-marking [322]; both of which may have contributed to inconsistencies with data used in the relabeled Formation Tops dataset used to train predictive models.

5.5 Conclusion

In this chapter, we have introduced a straightforward framework for the development of a machine learning-based predictive model that can determine the specific stratigraphic / formation group in the Midland Basin given spatial positioning and true vertical depth. The Formation Labeler model developed can be effectively trained and applied to as a resource delineation tool in domains spanning subsurface energy and environmental applications. The final model formulation, which was built on the random forest algorithm, was capable of an overall prediction accuracy of 93 percent on holdout data.

Despite its noted high predictive accuracy on holdout test data, the model may benefit from modifications that would improve its utility and performance precision. For instance, results from the *k*-means clustering analysis suggests the existing dataset may contain inherent complexity better aligned to 21 distinct groupings opposed to the 13 which were used in this study. The use of additional stratigraphic / formational groupings may be one approach to further improve model prediction utility (i.e., may occur though inclusion us of middle Sprayberry intervals, the Jo Mill, or sub-benches of the Wolfcamp). However, the added level of complexity would likely require the use of more data as part of model training. Additionally, the findings presented in Table 21 beg for ensured consistency in formation nomenclature and top bench-marking across all domains to improve framework utility.

Another viable next step for this work would be to incorporate the framework described with a lithology classification approach from well log data as described by the likes of Wang and Carr (2012), Xie et al., (2019), and Ren et al., (2019) [291, 323, 324]. The coupling of the two concepts may offer a basin-wide formation classification and geologic property approximation approach. Given more granularity in datasets (for example, well log data at foot or half-foot

resolution), the framework discussed here could be applied to enable higher resolution and more precise prediction for specific geologic features and property characteristic of interest at new drilling sites. Specific examples in this regard include the identification of specific landing targets for oil and gas applications, identifying the depth and thickness of injection and / or confining layers for carbon dioxide geologic storage cases, or even reservoir temperature regimes for geothermal energy exploration. While this approach was implemented across the Midland Basin producing region, it could be easily applied to other mature basins; either in another Permian sub-basin or elsewhere.

6.0 Conclusions and Future Work

The use of ML-based techniques is gaining substantial interest to complement existing decision-making strategies in unconventional O&G exploration and production operations. Researchers are taking advantage of the emergence of digital O&G datasets by exploring the use of ML, the result of which provides an opportunity for the various facets of O&G to become much more data-informed than ever before. However, a research need remains regarding the application of ML modeling beyond data-driven replication of O&G exploration and operational tasks, but aimed towards informing improvements in future O&G development and operations over current industry baselines. The objectives of this dissertation were targeted, in part, at this specific research need. The approaches taken included an evaluation of region-specific industry performance data through time for identifying opportunities through the use of ML conducive to: 1) Garnering insight associated with the interaction of specific well designs and spatially-distinctive geology in studied areas through feature importance analyses; 2) improving the recovery of hydrocarbons in unconventional reservoirs; and 3) estimation of the types and volumes of fluids produced at the well level – each of which require specific management strategies and hold potential environmental implications.

In general, the resulting ML-based models developed through this dissertation work were specific to the targeted plays chosen for evaluation. However, the methodological framework implemented could transfer to other O&G reservoirs not evaluated in this dissertation relatively seamlessly. Furthermore, the data parameters utilized are relatively common for plays across the U.S. and may be readily acquired from public sources. The combination of these topics suggests that there is ample opportunity to continue or expand upon the research conducted within this

dissertation. Section 6.1 below highlights key findings and implications from the studies conducted and compiled in Chapters 2 through 5, as well as reiterate suggested potential follow on or continuation research under each. Section 6.2 below suggests broader topics which are worthy of consideration for future work related to ML in subsurface applications. Even the most recent accounts of exploration and production from O&G plays are creating an abundance of data from which ML based techniques can be implemented and evaluated. The result of which affords certainty towards a bright future for applying research across various facets of the O&G domain using data-driven approaches.

6.1 Summary of Conclusions and Potential Next Steps from Dissertation Research

Work conducted in Chapter 2 involved application of ML to a large dataset encompassing the producing extent of the Marcellus Shale. Models were developed that were capable of accurate prediction of two different productivity indicators at the well level that strongly correlate to EUR. Specifically, GBRT was applied; an algorithm that has been narrowly investigated for O&G applications but enables straightforward parametric importance and influence evaluation, as well as assessment of parameter interaction effects. The models developed provide a capability beneficial to reservoir management in the Marcellus Shale that enables fast and effective evaluation of the impact of various well placement and design choices. The models developed performed well (based on R^2 and RMSE scores against hold out datasets) for predicting the productivity indicators evaluated over a large, geologically diverse study domain. The noted performance accuracy is attributed to two main factors: 1) the use of GBRT that handles the complexity associated with unconventional O&G systems quite well, and its inherent sequential

construction of weak learners that compensates for shortcomings of those previously developed; and 2) the Top 12-months productivity indicator developed through this research, which provides for added utility for estimating the production potential for a given well based on its design and placement in the reservoir.

Study results indicate that the importance of the geologic proxy parameters of well surface longitude and latitude on well productivity. A valuable expansion to this approach would be to consider inclusion of explicit geologic data as part of modeling inputs to further explore the specific geologic conditions influencing productivity. However, a noted challenge in such a pursuit is that the availability of adequate levels of geological data at the well level is hard to come by [51]. Evaluation over expansive spatial domains with the inclusion of large well counts within the dataset (e.g., over 7,000 wells used in Chapter 2 work) can make this type of approach even more challenging. One solution is to evaluate geology post-hoc to ML model development and analysis. Geologic data, which may be limited based on availability, can be more effectively evaluated through targeted analysis based on the findings and implications from ML modeling results and output. This post-hoc conceptual approach was considered and implemented as the basis behind Chapter 3 work.

Chapter 2 study results also identified gross perforated interval length and water and proppant per foot as critical well completion and design choices influencing productivity. These results suggest consistency with others studies that have also modeled and evaluated/ranked well design and geologic parameters on productivity performance in other unconventional plays [55, 64, 15, 58]; a finding which also supports the efficacy of the GBRT models in this case. When employing GBRT models to evaluate specific well case studies case by case, relative improvements in well productivity were found tied to upscaling of water and proppant volumes

per foot as part of hydraulic fracturing. However, the relative magnitude of noted productivity improvements based on water and proppant levels is spatially dependent across the Marcellus, influenced by the prevailing geologic conditions. Other well completion design considerations like additive per foot and spacing to other wells were shown to have noticeable, but more understated effects on productivity compared to gross perforated interval, water and proppant per foot, and geology. The effect of wellbore azimuth was found to be marginal at best despite being widely considered a critical well design consideration in unconventional reservoirs. A potential hypothesis worth exploration is to evaluate if operators have historically aligned wellbore trajectories to near optimal azimuth orientations, hence the suppressed marginal effect noted in Chapter 2.

Additionally, Chapter 2 study results indicate that Marcellus well performance improves most with upscaling perforated interval lengths and water and proppant volumes per foot; but relative well-level productivity improvements are spatially dependent across the play. Additionally, optimal combinations of water and proppant on well performance were found to vary depending on well location, emphasizing the utility of data-driven models capable of broad application across a play of interest for informing tailored well design approaches prior to their field deployment.

The intent of the research conducted in Chapter 3 is to further build off of the accomplishments achieved via Chapter 2 work, but do so in a way that attempts to offer insights that may prompt improvements over industry well performance benchmarks or best/common practices. Chapter 3 work introduces the ensembled framework which combines the GBRT predictive model developed and discussed in Chapter 2 with a well design optimization approach that maximizes productivity. The ensembled framework is applied across the producing extent of the Marcellus under various sampling strategies; the result of which enables spatial segregation

and regional ranking of the Marcellus based on productivity potential. This analysis strategy enables systematic evaluation of the distribution of the major drivers noted as highly influential to natural gas productivity by ranked region with common production potential (i.e., called "bins" in Chapter 3), including: 1) key geological properties as well as 2) the resulting optimized well design and completion attributes from LHC sampling. The production drivers were evaluated statistically to comprehend controlling factors on shale well production, and to identify if commonality or disparity exists in the prominent features. The outcome of this analysis can help informing the tailoring of designs of future well given their placement in the Marcellus and the associated geologic conditions they might encounter. The simulation results from the LHC sampling and brute force well optimization approach show the expansion of increased potential productivity across the Marcellus when wells were tailored based on placement as compared to standard well designs. Additionally, the LHC-derived optimized well deigns estimated higher gas productivity when compared to actual in-field designs at randomly-selected locations within the extent of the study area (to ensure fair comparison, in-field wells had similar gross perforated interval lengths [within +/- 75 feet from LHC designs] and were roughly proximal to the locations of the LHC-based simulated wells).

The statistical analyses conducted in Chapter 3 indicates that regions higher in productivity ranking show a significant difference for certain (but not all) geologic features favorable to gas production potential relative to lower productivity regions. Net reservoir thickness and porosity were notable geologic parameters that were significantly different between higher productivity and lower productivity groupings. Optimized well design parameter settings were shown to vary relative to their placement across the study area and subsequent productivity ranking region. Results indicate that wells in the highest productivity region/bin were generally 1) thicker and

more porous than lower productivity bins and 2) on average, tailored designs that maximize productivity are more intensive designs (more proppant, water, and additive, and reduced spacing) relative to wells in the other bins. Lower productivity regions / bins were relatively much thinner on average. To maximize productivity in these lower productive regions, optimal wells were found to be less proppant, water, and additive intense, and wells required spacing farther apart. The outcome of this ensembled ML deployment strategy can be insightful for future well planning and design exercises given their specific intended placement in the Marcellus.

The resulting tailored well designs discussed in Chapter 3 help to provide further comprehension towards achieving higher production potential in Marcellus wells given the prevailing geologic conditions. However, it remains unknown if the tailored well designs would be the most economically-viable well design considerations given their placement. A viable next step would be one that explores coupling the ensembled framework described in Chapter 3 (or similarly to deploy the reduced order model in Equation 3-6) with a cash flow / economic model to evaluate profitability potential of tailored well designs. For instance, a specific well that theoretically produces larger volumes of oil or gas and was tailored for the specific geologic conditions for which it was placed may not necessarily result in the most favorable economics compared to other designs that produce lower volumes of hydrocarbons. Wells that produce less could be favored by operators over wells that produce more, but do so as result of over capitalization of the higher-producing well. A study by Shahkarami and Wang (2017) is one example that explored this concept, but in the context of drilling spacing decisions on resulting production estimates and overall operational net present value [325]. Such a follow-on study could derive well designs that achieve either the most economic and/or most hydrocarbon production; and evaluate the differences between each.

Research discussed in Chapter 4 (and subsequently Chapter 5) shifts focus away from the Appalachian Basin (major natural gas producing region) and towards the Permian Basin region of the U.S. The Permian Basin holds enormous consequence regarding both domestic oil and gas production and has become the world's top-producing oil field. Despite its prominence as a vital oil and gas producing asset, the basin currently faces several development challenges, including: 1) noticeably steeper well decline rates and IP's as development shifts to non-core regions; 2) associated natural gas production is exceeding pipeline takeaway capacity, resulting in flaring or even venting excess natural gas; and 3) produced water volumes are substantially large, and volumes are typically managed via disposal through deep well underground injection. Collectively, these circumstances may threaten the Permian's overall production efficacy while consequently increasing the environmental burden associated with O&G operations. Chapter 4 research targets these specific challenges.

Chapter 4 work leverages a combination of supervised and unsupervised ML approaches as part of a framework to enable joint prediction of both produced water and natural gas volumes associated with oil production from unconventional reservoirs in a time series fashion; focusing primarily in the pay zones within the Spraberry and Wolfcamp Formations of the Midland Basin in the U.S. The supervised learning component included a deep learning-based model (based on LSTM) with the capability to generate reliable estimates of produced water and natural gas in a time series manner based on well completion and placement factors. The unsupervised learning component (based on *k*-means clustering) establishes groupings of related wells based on a multitude of factors, including spatial placement within the basin, well design component features (i.e., proppant, water, additive volumes, perforated interval length, etc.), and historic productivity.

The clustering method provides a structure from which to extract attribute properties related to Arps Decline, well completion, and reservoir and spatial attributes characteristic of each cluster group. The ensemble of the supervised and unsupervised elements of this work facilitates a means to forecast oil, water, and natural gas production at the well level as influenced by specific development considerations. The ML model functionality developed enables well level three-stream production volume estimates which can be used generate production outlooks at various scales. Fluid volume outlooks can be used to help support the formulation of management and/or remedial strategies based on the volumes of fluids expected from O&G development. The example outlooks demonstrated in Chapter 4 emphasize the variability in water and gas volumes produced depending on wellbore design and placement considerations – a finding which indicates the high degree of complexity which exists and can confound forecasting tasks.

Time series-based predictive forecasting of hydrocarbon production using deep learning ML strategies has garnered substantial research interest as of late. Chapter 4 work intends to expand upon the foundation established from prior research that focuses heavily on time series oil production, to other aspects critical to O&G development also known to share 1) dynamic, temporal dependencies in data and 2) aspects with implications to production forecasting. Continued research as it relates to time series, associated fluid production worthy of consideration may include topics that address the imposed constrains in the research presented in Chapter 4. For instance, the inclusion of additional geologic characteristics (also relevant to Chapters 2 and 3) that are known controlling factors to unconventional oil and gas recovery [26] may provide added utility in the data-driven ML modeling (some of which may be temporally-dependent, like oil, water, and gas saturations, formation temperature, and formation pressure). Additionally, the work in Chapter 4 lacked access to temporal attributes pertaining to how wells were operated and

managed (i.e., choke, bottom-hole pressure, lift type). Essentially, oil production volumes were used as a proxy to capture those temporal components (i.e., they served as inputs to the joint prediction model). Without such insight, there are challenges to integrating the human element as part of the forecasting component. As a result, any gradual or abrupt changes in the producing rate of wells due to reservoir depletion, fluctuations in bottom-hole pressure, or changes in conditions in or immediately adjacent to the wellbore cannot be directly considered when only using the Arps decline models in isolation. This level of temporal data is often held proprietary by operators that oversee in-field wells. However, the additional insight this type of data may provide when applied to ML could be vast – a concept that is discussed in more detail in Section 6.2 of this conclusion section. Furthermore, the potential inclusion of this level of data enables forecasting (separate from history matching replication) to include dynamic aspects more representative of the physical interactions expected in-field.

Chapter 5 introduces a framework for creating classification-based predictive models to identify specific formations given input data consisting of total vertical burial depth, latitude, and longitude coordinates. The example used in this dissertation was designed to identify 13 different reservoirs in the Midland Basin. Study results indicated that a random forest classifier algorithm performed best out of the four different algorithms evaluated. The final model formulation is capable of an overall prediction accuracy of 93 percent on holdout data. This type of ML model can be employed to provide rapid delineation of specific geologic strata. Essentially, one can easily infer burial depth, thickness, and lateral extent of formations of interest – a critical undertaking relatively common when exploring for, visualizing, and ultimately developing or appraising subsurface resources [292]. Specific examples in this regard include, planning well drilling operations and logging, well completion engineering and design choices, subsurface resource

quantification estimation (i.e., oil and gas productivity potential, CO_2 storage or liquid waste disposal resource capacity and containment amenability, geothermal potential, etc.), and project cost estimation; all of which are dependent on the geologic sequencing and their associated areal extents at new well sites.

This modeling described in Chapter 5 provides a straightforward and cost-effective approach applicable to multiple subsurface applications where identification or delineation of resources is needed. As discussed in Chapter 6, viable options for follow on research that would add the utility of the 3D mapping potential, by infusing into the framework, the ability to map specific subsurface properties of interest; either via ML or more physically-rooted approaches. For instance, the lithology classification framework from well log data implemented by the likes of Wang and Carr (2012), Xie et al., (2019), and Ren et al., (2019) may provide ML-specific examples to couple with the formation labeler framework, potentially offering a basin-wide formation classification and geologic property approximation approach. The requirement here would be higher granularity in the datasets needed (for example, point data like well log data at foot or half-foot resolution, or more spatial sources like seismic data in 2D or 3D formats). The basis of the framework discussed in Chapter 6 could be applied (with some tailoring expected) to integrate higher resolution data, which could enable more precise prediction for specific geologic features and property characteristic of interest. The specific uses of the output from this type of modeling could be vast (specific examples listed in bullets in the introduction in Chapter 5) and attract interest that spans multiple domains (i.e., O&G, environmental, geothermal, energy storage).

6.2 Broader Research Concepts

Several ML-based analyses, including those discussed in this dissertation, offer new ways to implement O&G field exploration or development strategies. However, there are limited examples of actual deployment of in-field exploration or development informed by ML. A novel research endeavor could involve the implementation of an in-field exploration, well design and production strategy, or forecasting exercise that is largely ML-informed. Such a research approach can be used to 1) evaluate the efficacy of ML-derived field strategies, as well as 2) provide ground-truth data from which ML-based processes can be refined. Researches at West Virginia University and Los Alamos National Laboratory are exploring this concept at the Marcellus Shale Energy and Environment Laboratory under the U.S. DOE-supported SMART Initiative as an example [326]. Additionally, physics-based reservoir modeling and simulation can create synthetic data streams which can be used to validate ML models as an alternative to field test confirmation (a potential earlier mover topic and predecessor to full field-scale testing). The successful demonstration of notable in-field improvements using ML may prompt wider and more rapid adoption from industry moving forward.

A second concept topic relates to broader access to expanded datasets not typically available in the public domain. Publicly-available datasets, as well as those curated by vendors (i.e., DrillingInfo and IHS Markit), have been critical in enabling ML-based analyses for O&G applications. DrillingInfo, in particular, has been extensively leveraged throughout this dissertation. However, publicly available datasets alone may be considered insufficient to achieve certain targeted solutions for specific O&G problems. Site-specific field data collected by O&G operators often exists and can help augment those available publicly [32]. Unfortunately, these types of data are often held in proprietary formats that are not easily accessible by others outside

that particular organization. Access to proprietary datasets may enable the development of more complex ML-based models by increasing the magnitude of available input parameters that may be influencing system response. As a result, there is an opportunity to potentially explain more of the modeling variability by including input parameters that influence the response variable(s). Such an exercise may provide further insight into the controlling factors (and their magnitude) on productivity response. Additionally, higher complexity models that account for greater variability in the specific O&G application evaluated will be more accurate and provide greater insight into improving industry best practices. Resulting models, for instance, could be integrated into similar ensemble frameworks described in Chapter 3 for process optimization.

In this regard, there has been an interest, particularly by the research community, towards data sharing to gain larger insight for improving subsurface engineered systems. Access to large datasets, often owned by more than one operator, may be necessary to perform more meaningful ML analyses. For instance, site characterization data in the O&G industry are often held stringently and not shared with others because of the competitive value that they are perceived to possess. However, in certain cases, a single operator might not have a sufficient volume of data assets in a single field or play to justify the use of ML. Recent studies have shown that, for some applications, data from several hundred to a thousand wells may be needed before data analytics and machine learning can inform operators in meaningful ways [327].

Data sharing or data pooling should be encouraged to create larger datasets that can provide significant new insights to field operators when circumstances to do so make practical sense. Even when data pooling among operators is not necessary, researchers may generate breakthroughs via access to significant volumes of data from operators. For such an endeavor to be successful, operators must be willing to share resources, but, in return, should also expect to receive some type

of ancillary benefit as a result of the research performed. Secure agreements that can allow researchers access to data while ensuring that proprietary data remain secure will be necessary as new techniques are applied and developed for specific use cases [32]. Research endeavors using ML, or through non-data-driven approaches, that can demonstrate the impact of proprietary data in attaining insightful findings and results could be of substantial benefit for operators who provided the data, as well as the research community as a whole. The hope would then facilitate continued data sharing from operators moving forward and prompt new research that can generate findings which can improve the way we explore and use our vast subsurface resources.

Lastly, the final concept mentioned here is in regards to potential inclusion of additional physics-informed insight into ML modeling for subsurface applications. In general, ML (and deep learning) serve as universal interpolators that attempt to find correlations within large datasets in multidimensional spaces. While they do offer the benefits and advantages of speed in prediction, accuracy, and the potential to gain new insights from large datasets, ML, by nature, they may only perform well when models are used for interpolating under circumstances when cause and effect are fairly well known and constrained. For instance, these circumstances would include when ML-derived models are used in a parameter space in which the controlling factors remain constant and the empirical observations from which the data are derived are made with the same biases. However, ML approaches can struggle and often fail when links between cause and effect are more uncertain. Vasudevan et al. (2021), as one example, has argued for the infusion of physical statistics (e.g., Bayesian methods with prior knowledge, loss function augmentation to account for physics-relevant regularization, constrained model response, and pre-trained models with synthetic data from physical models [328]) into ML modeling workflows in order to make ML and deep learning more robust [329]. This concept would be applicable in the subsurface energy /

environmental application research domain where ML methods are generating substantial interest and value in their use. From an O&G perspective, a blend of physics-based modeling with ML may enable new and improved reservoir management strategies than from the use of either in isolation. Two use-case examples have been proposed by Bettin et al. (2019) related to subsurface applications [32]; these include: 1) Coupling physical modeling with ML to enable attainment of key information within an emerging play / basin where limited exploration and data collection has previously occurred and 2) the use of physics-based modeling with ML approaches may enable model transferability to enable extrapolation from a mature and well-understood formation (i.e., like the Midland Basin or Appalachian [Marcellus] Basin) and into a new plays, either within the basin or elsewhere.

Appendix A

University of Pittsburgh Day in Harrisburg Research Showcase Poster Presentation

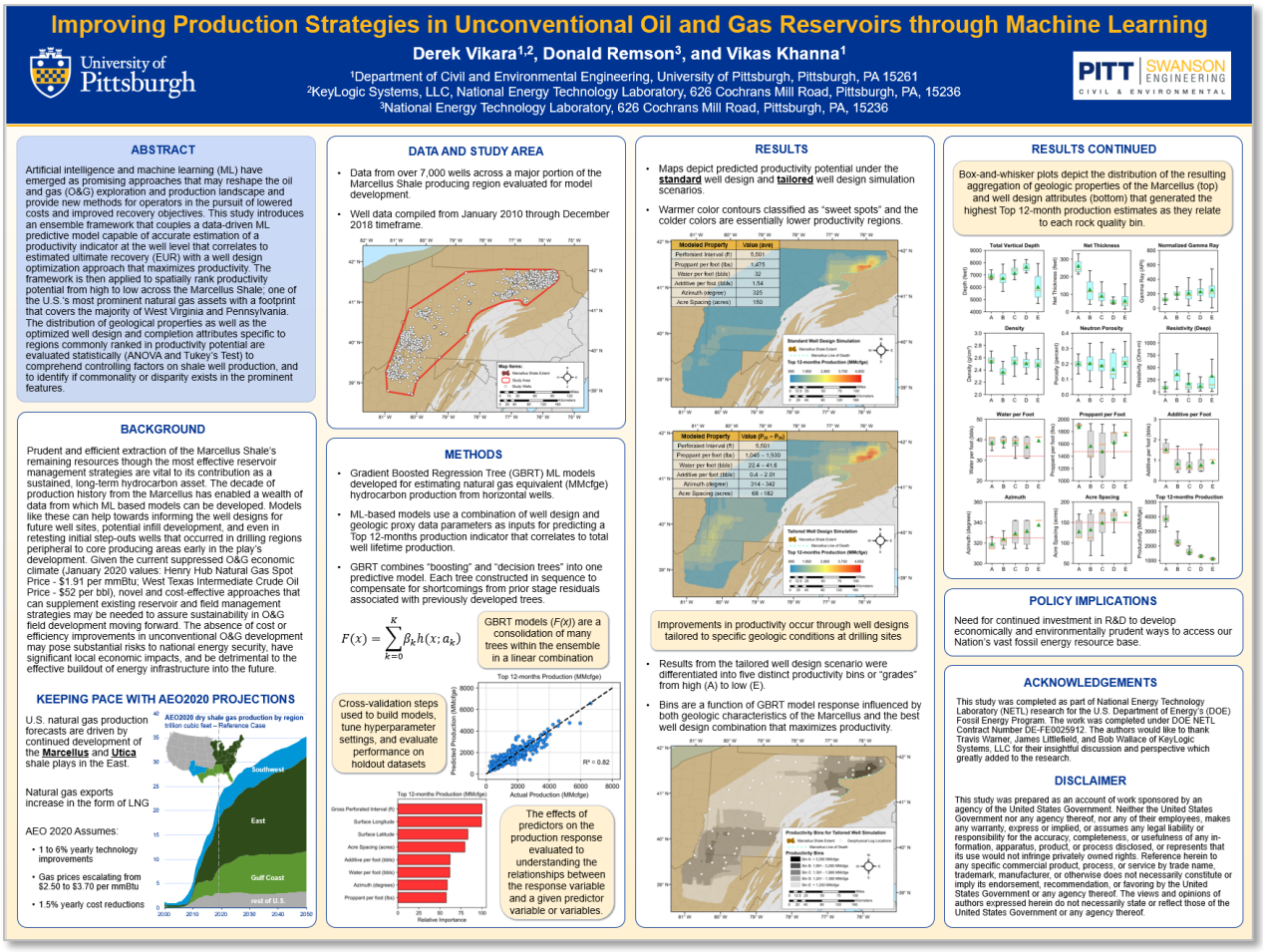

Figure 51. Screenshot of research poster for presentation as part of University of Pittsburgh's 2020 Pitt Day in Harrisburg.

Appendix B

Supporting Material for "Gaining Perspective on Unconventional Well Design Choices through Play-level Application of Machine Learning Modeling"

Table 22. Summary statistics of the well dataset used for the study. A total of 4,257 wells were utilized. Every well included had data available for each of the parameters listed below.

Parameter	Mean	Std. Dev	Minimum	25%	50%	75%	Maximum
First 12-months (MMcfge)	1,503	1,030	0	774	1,284	1,975	8,295
Top 12-months (MMcfge)	1,637	1,084	49	863	1,373	2,132	8,313
Gross Perforated Interval (ft)	5,501	2,088	507	4,051	5,180	6,651	15,530
Proppant per foot (lbs)	1,475	866	11.8	1,112	1,371	1,788	30,091
Water per foot (bbls)	32	19.1	0.6	23.6	31.6	40	895
Additive per foot (bbls)	1.54	3.81	0.001	0.47	0.89	1.69	102
Azimuth (degrees)*	325	29.3	180	319	331	340	360
Nearest Neighbor (ft)	1,197	944	4	699	921	1,140	5,241
Acre Spacing (acres)	150	126	0	79	113	166	1,094
Surface Latitude (decimal degrees)	40.643055	0.97	38.787357	39.818957	40.614137	41.658483	41.996642
Surface Longitude (decimal degrees)	-78.721317	1.95	-81.026254	-80.529332	-79.844768	-76.834225	-75.555286

*Per similar approaches by Shih et al. (2018) and LaFollette et al. (2013), all well azimuth trajectory data was adjusted to fall between 180^{o} and 360^{o} to avoid a bi-modal distribution of well orientation.

Analysis of Model Residuals: The histograms in Figure 52 compare the residuals between final model formulations predicting either productivity indicator against the testing dataset. For both models, the bulk of the residuals occur below a 500 MMcfge difference between actual and predicted production (82 percent of residuals for Top 12-months and 78 percent of residuals for First 12-months). Figure 52 also emphasizes that the First 12-months prediction typically has a larger count of wells at most residual levels along the x-axis when compared to the Top 12-months production; providing an alternative perspective toward the error manifestation between models.

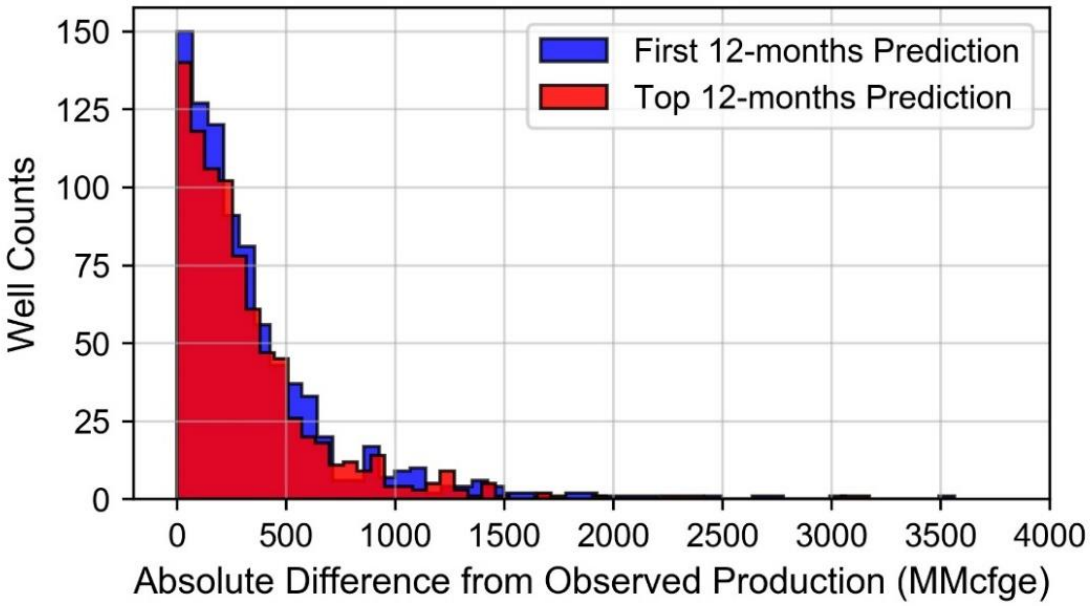

Figure 52. Histogram of the absolute values prediction residuals for each productivity indicator against the testing dataset using the final model formulations.

A mapping exercise was conducted for visual inspection of the distribution of model residuals for prediction against the testing dataset (Figure 53). This execrise is not intended to be a quantitative analysis of residual patterns or spatial autocorrelation, but rather to point out if noticable trends in error manifestation emerge. No obvious or unusual patterns seem to exist, suggesting that models are effectively handling spatial variabilities.

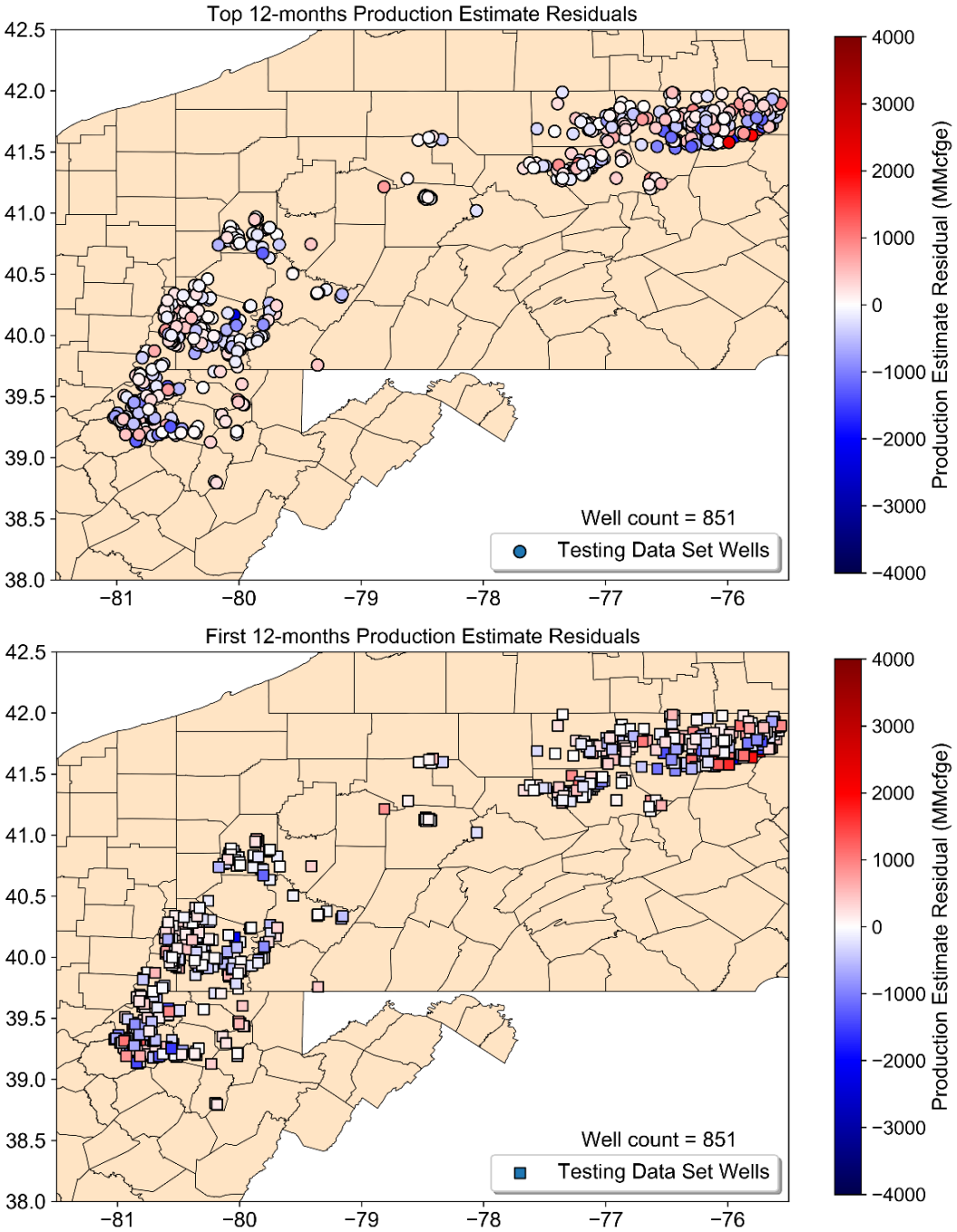

Figure 53. Maps depicting the final model formulations prediction residuals for testing dataset wells for the Top 12-months productivity indicator response (top) and First 12-months prediction indicator response (bottom). Positive residuals (red coloration) indicate models over-estimate production compared to observed values, and negative residuals (blue coloration) indicate models under-estimate production compared to observed values.

Under a few instances, larger residuals (darker blue or brighter red coloration in Figure 53) appear at various points across the play. The potential causes for large discrepancy between actual and predicted production values in many of these cases could occur for a number of reasons—several of which might not be reflected in parameters of the current project dataset. For instance, large residuals could simply be a result of the models slightly under or over predicting production on high performing wells and the actual percentage delta between actual and predicted production values is relatively small (this can be reconciled via review of the scatter plots in Figure 11 in the article). Additionally, there is no way to assess if other well design characteristics are influencing productivity that are not captured in the current dataset. As one example, well laterals that may have deviated substantially out of zone or have drastically dissimilar wellbore orientations (toe up vs. toe down vs. high wellbore tortuosity) to other proximal wells in comparable geologic conditions could be subject to unknown factors that influence model prediction discrepancy compared to actual observed production. There could also be cases where drastic but localized variation in geologic characteristics occur compared to nearby wells. The spatial assessment of residuals presented here enables identification of certain areas or of individual wells where more focused studies can be conducted to evaluate if unique circumstances or characteristics exist that are impacting productivity not captured in the dataset.

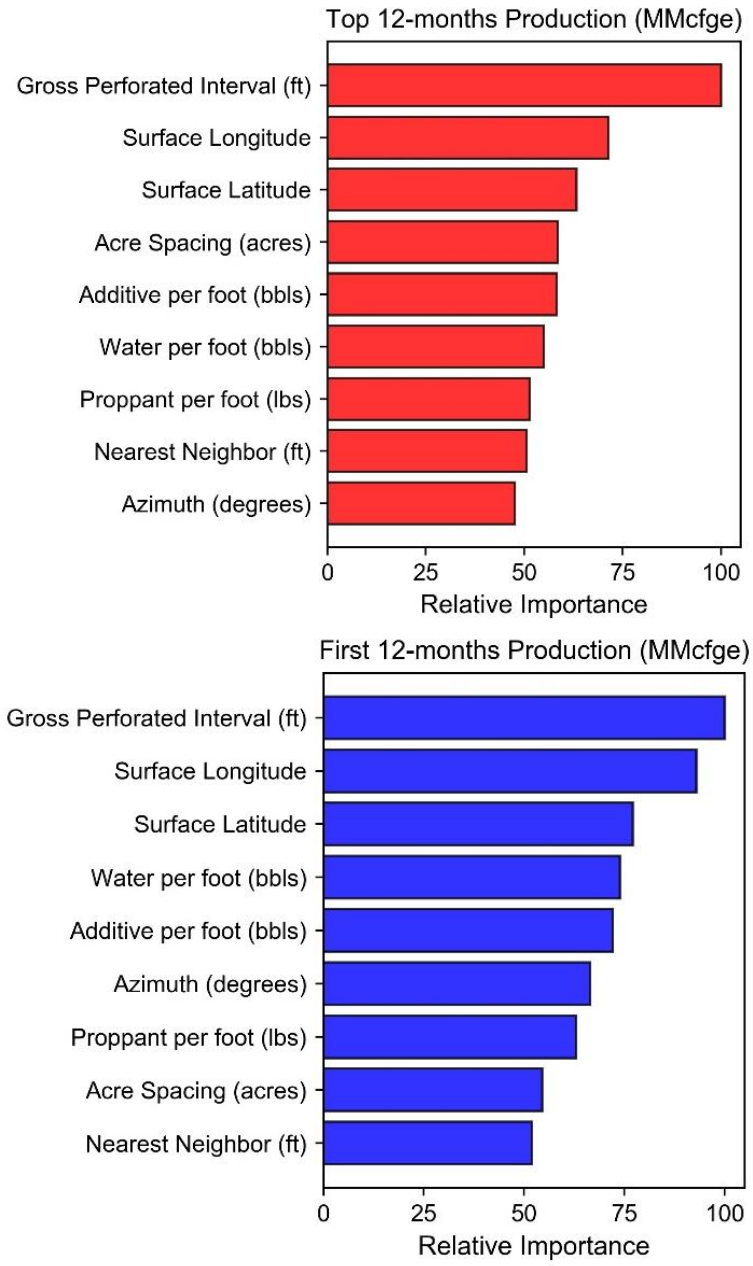

Figure 54. Summary of the relative importance of the predictor variables for the Full Model formulations (with nearest neighbor parameter included) for the Top 12-months response (left) and First 12-months response (right).

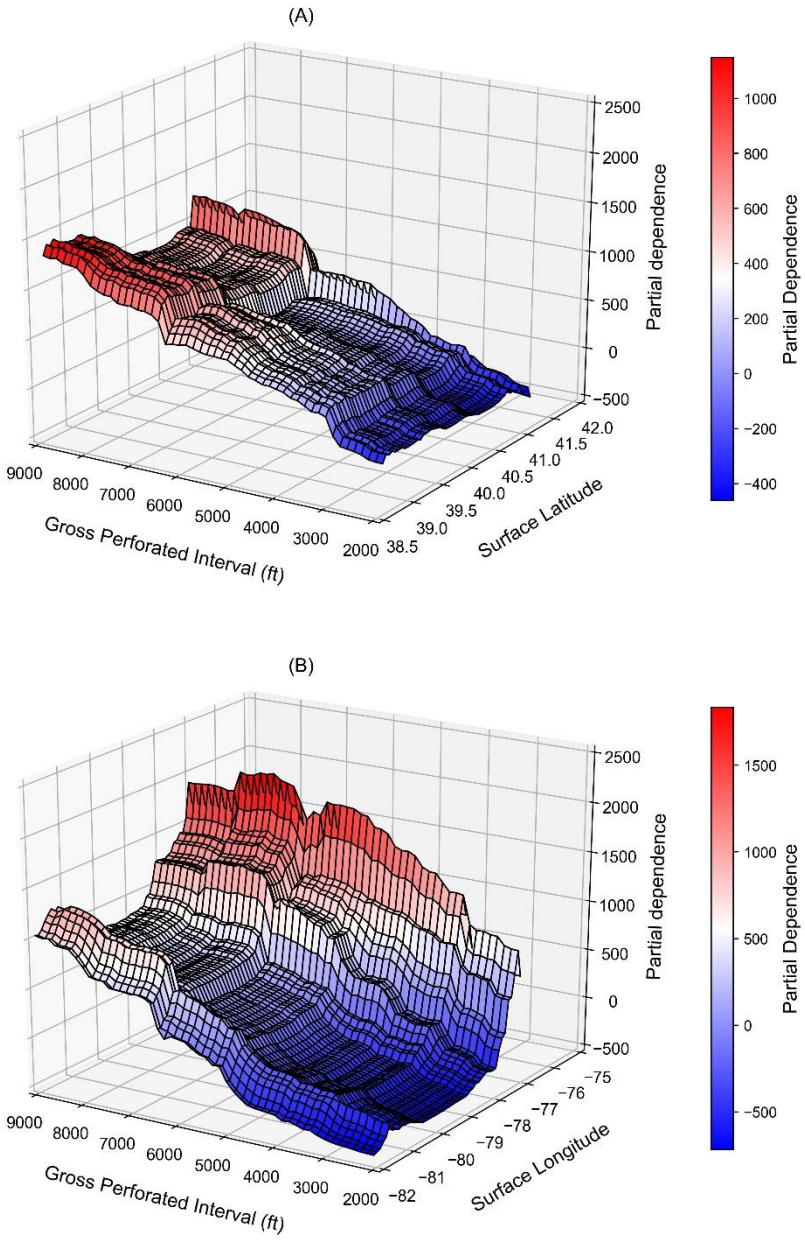

Figure 55. Three dimensional plots of partial dependence for predicting the First 12-months productivity indicator using the final model formulation. The top figure (A) evaluates the interaction of perforated interval length and surface latitude. The bottom figure (B) evaluates the interaction of perforated interval length and longitude.

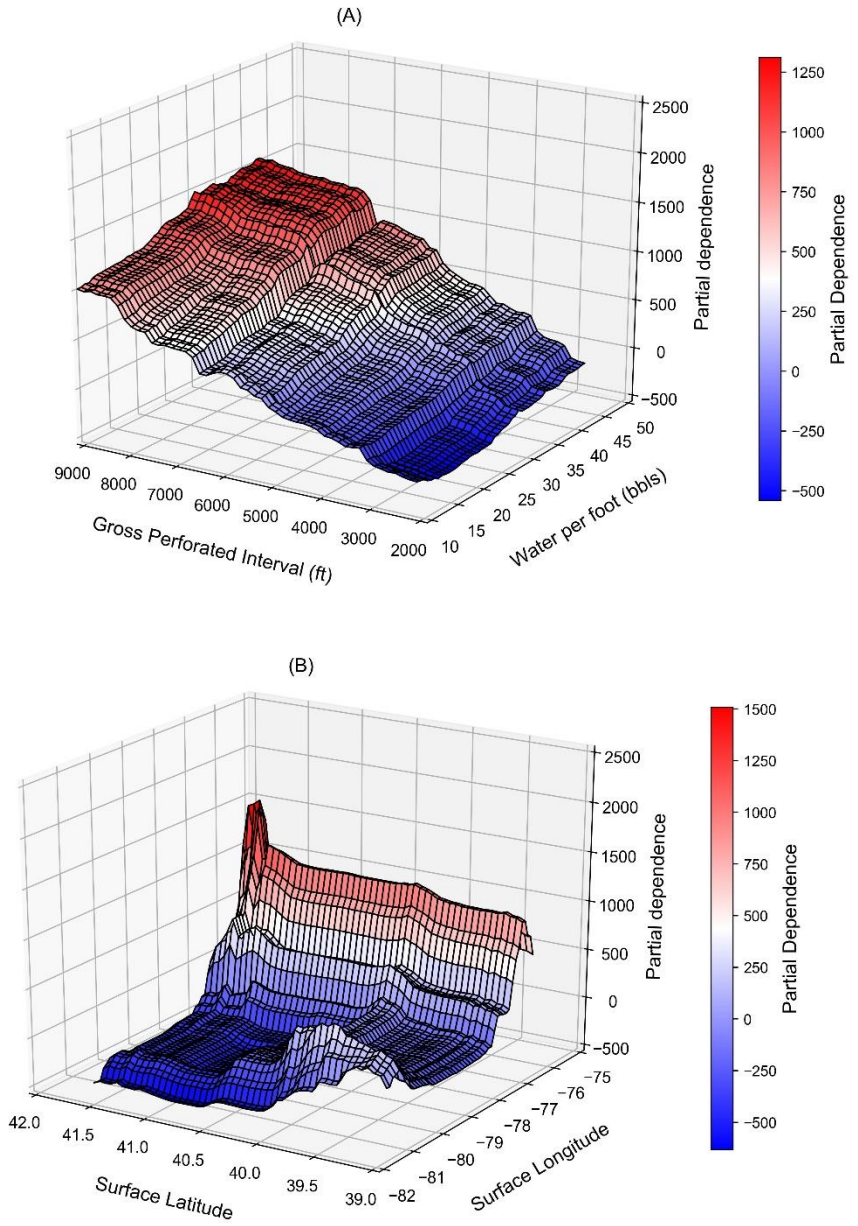

Figure 56. Three dimensional plots of partial dependence for predicting the First 12-months productivity indicator using the final model formulation. The top figure (A) evaluates the interaction of perforated interval length and water injected per foot. The bottom figure (B) evaluates the interaction of latitude and longitude.

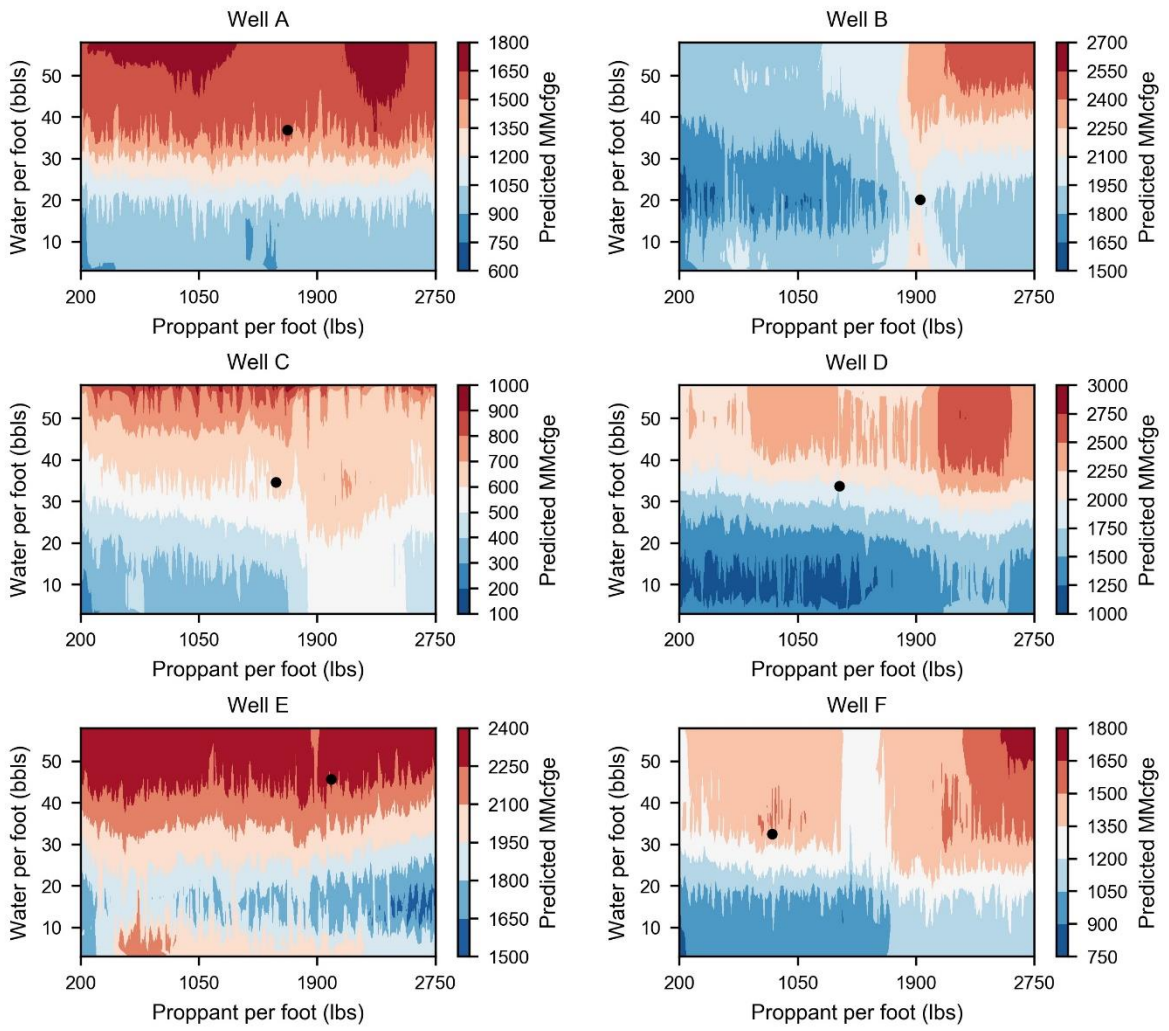

Figure 57. Contour diagrams for estimated First 12-months production for each well evaluated in the case study with varying water and proppant per foot input values. The black dots represent the implemented field designs for each corresponding well.

Appendix C

Supporting data for "Machine Learning-informed Ensemble Framework for Evaluating Shale Gas Production Potential: Case Study in the Marcellus Shale"

Supplementary material files and data were generated as part of this study and provided as appendices on Mendeley Data [144]. These data can be cited as:

Vikara, D., Remson, D., Khanna, V. 2020, "Supplementary Data for Machine Learning-informed Ensemble Framework for Evaluating Shale Gas Production Potential: Case Study in the Marcellus Shale." *Mendeley Data*, V1, doi: 10.17632/vbgywdcdjp.1

The first component includes the results from simulations at all pseudo well locations for the standard and tailored modeling scenarios. The second component is a compilation of the well log data used in this analysis. The third component includes the training dataset generated for the reduced order linear model, as well as an evaluation of model residuals. The datasets are extensively long and therefore, raw data is not appended in this thesis. Each dataset can be downloaded at the following Mendeley website:

https://data.mendeley.com/datasets/vbgywdcdjp/1

Bibliography

[1] Van Eck Global, "Unconventional Oil & Gas - Demystifying Fracking and Understanding Global Opportunities," Van Eck Global, 2005. [Online]. Available: https://www.vaneck.com/research-unconventional-oil-and-gas-pdf. [Accessed 9 February 2020].

[2] U.S. Department of Energy, "Ethane Storage and Distribution Hub in the United States," U.S. DOE, Washington, D.C., 2018.

[3] Pirog, R., and Ratner, M., "Natural Gas in the U.S. Economy: Opportunities for Growth," Congressional Research Service, 2012.

[4] U.S. Energy Information Administration, "Today in Energy: Both natural gas supply and demand have increased from year-ago levels," U.S. Department of Energy, 4 October 2018. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=37193. [Accessed 31 March 2019].

[5] Clemente, J., "U.S. Natural Gas Demand for Electricity Can Only Grow," Forbes, 15 January 2019. [Online]. Available: https://www.forbes.com/sites/judeclemente/2019/01/15/u-s-natural-gas-demand-for-electricity-can-only-grow/#27b0ba844c74. [Accessed 31 March 2019].

[6] National Energy Technology Laboratory, "Unconventional Resources," U.S. Department of Energy, Undated. [Online]. Available: https://netl.doe.gov/unconventional. [Accessed 7 February 2020].

[7] Baker Institute, "Natural Gas Markets Beyond COVID-19," Forbes, 1 April 2020. [Online]. Available: https://www.forbes.com/sites/thebakersinstitute/2020/04/01/natural-gas-markets-beyond-covid-19/#1b410a6354c4. [Accessed 1 May 2020].

[8] U.S. Energy Information Administration, "Cushing, OK WRI Spot Price FOB," U.S. Department of Energy, 22 January 2021. [Online]. Available: https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RWTC&f=M. [Accessed 2021 24 January].

[9] U.S. Energy Information Administration, "Henry Hub Natural Gas Spot Price," U.S. Department of Energy, 22 January 2021. [Online]. Available: https://www.eia.gov/dnav/ng/hist/rngwhhdm.htm. [Accessed 24 January 2021].

[10] U.S. Energy Information Administration, "Annual Energy Outlook 2020," U.S. Department of Energy, Washington, D.C., 2020.

[11] U.S. Energy Information Administration, "Assumptions to AEO2020," U.S. Department of Energy, 29 January 2020. [Online]. Available: https://www.eia.gov/outlooks/aeo/assumptions/. [Accessed 27 December 2020].

[12] R. Barree, S. Cox, J. Miskimins, J. Gilbert and M. Conway, "Economic optimization of horizontal well completions in unconventional reservoirs," *SPE Production & Operations,* vol. 30, no. 4, pp. 293-311, 2015.

[13] M. Vincent, "The next opportunity to improve hydraulic-fracture stimulation," *Journal of Petroleum Technology,* vol. 64, no. 3, pp. 118-127, 2012.

[14] D. Alfarge, M. Wei and B. Bai, "Evaluating the performance of hydraulic-fractures in unconventional reservoirs using production data: Comprehensive review," *Journal of Natural Gas Science and Engineering,* vol. 61, pp. 133-141, 2019.

[15] S. Wang and S. Chen, "Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling," *Journal of Petroleum Science and Engineering,* vol. 174, pp. 682-695, 2019.

[16] G. Luo, Y. Tian, M. Bychina and C. Ehlig-Economides, "Production Optimization Using Machine Learning in Bakken Shale," in *Unconventional Resources Technology Conference*, Houston, Texas, 2018.

[17] Hegde, C.; Gray, K., "Use of machine learning and data analytics to increase drilling efficiency for nearby wells," *Journal of Natural Gas Science and Engineering,* vol. 40, pp. 327-335, 2017.

[18] C. Noshi and J. Schubert, "The Role of Machine Learning in Drilling Operations; A Review," in *Society of Petroleum Engineers - SPE/AAPG Eastern Regional Meeting*, Pittsburgh, Pennsylvania, 2018.

[19] T. Zhao, V. Jayaram, K. Marfurt and H. Zhou, "Lithofacies Classification in Barnett Shale Using Proximal Support Vector Machines," in *Society of Exploration Geophysicists 2014 Annual Meeting*, Denver, Colorado, 2014.

[20] T. Zhao, S. Verma, D. Devegowda and J. Jayaram, "TOC Estimation in the Barnett Shale From Triple Combo Logs and Time Series Analysis," in *Society*

of Exploration Geophysicists International Exposition and 85th Annual Meeting, New Orleans, Louisiana, 2015.

[21] S. Bhattacharya, T. Carr and M. Pal, "Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA," *Journal of Natural Gas Science and Engineering*, vol. 33, pp. 1119-1133, 2016.

[22] Z. Zheng, P. Kavousi and D. Haibin, "Multi-Attributes and Neural Network-Based Fault Detection in 3D Seismic Interpretation," *Advanced Materials Research*, Vols. 838-841, pp. 1497-1502, 2014.

[23] B. Cline., R. Niculescu, D. Huffman and B. Deckel, "Predictive maintenance applications for machine learning," in *Annual Reliability and Maintainability Symposium (RAMS)*, Orlando, Florida, 2017.

[24] D. Pandya, A. Srivastava, A. Doherty, S. Sundareshwar, C. Needham, A. Chaudry and S. Krishnalyer, "Increasing Production Efficiency via Compressor Failure Predictive Analytics Using Machine Learning," in *Offshore Technology Conference*, Houston, Texas, 2018.

[25] P. Tahmasebi, F. Javadpour and M. Sahimi, "Data mining and machine learning for identifying sweet spots in shale reservoirs," *Expert Systems with Applications*, vol. 88, pp. 435-447, 2017.

[26] K. Qian, Z. He, X. Liu and Y. Chen, "Intelligent prediction and integral analysis of shale oil and gas sweet spots," *Petroleum Science*, vol. 15, pp. 744-755, 2018.

[27] Zeiss, "machine-learning," 28 November 2016. [Online]. Available: https://blogs.zeiss.com/digital/the-relation-between-computer-vision-and-machine-learning/machine-learning/. [Accessed 24 January 2020].

[28] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *International Journal of Advanced Research in Artifical Intelligence*, vol. 2, pp. 34-38, 2013.

[29] A. Tarca, V. Carey, X. Chen, R. Romero and S. Drăghici, "Machine Learning and Its Applications to Biology," *PLOS Computational Biology*, vol. 3, no. 6, pp. 0953-0963, 2007.

[30] J. Brownlee, "Supervised and Unsupervised Machine Learning Algorithms," Machine Learning Mastery, 16 March 2016. [Online]. Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/. [Accessed 6 May 2019].

[31] S. Mirsha and A. Datta-Gupta, Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences, Amsterdam, Netherlands: Elsevier, 2018.

[32] G. Bettin, G. Bromhal, M. Brudzinski, A. Cohen, G. Guthrie, P. Johnson, L. Matthew, S. Mishra and D. Vikara, "Real-time Decision Making for the Subsurface Report," Carnegie Mellon University Wilson E. Scott Institute for Energy Innovation, Pittsburgh, Pennsylvania, 2019.

[33] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, no. 5500, pp. 2323-2326, 2000.

[34] L. Buşoniu, R. Babuška and B. De Schutter, "Multi-agent Reinforcement Learning: An Overview. In: Srinivasan D., Jain L.C. (eds)," *Innovations in Multi-Agent Systems and Applications - 1. Studies in Computational Intelligence,* vol. 310, pp. 183-221, 2010.

[35] C. Sapp, "Preparing and Architecting for Machine Learning," Gartner, 17 January 2017. [Online]. Available: https://www.gartner.com/binaries/content/assets/events/keywords/catalyst/catus8/preparing_and_architecting_for_machine_learning.pdf. [Accessed 10 September 2018].

[36] R. Thallam and M. Dominguez, "Build end-to-end machine learning workflows with Amazon SageMaker and Apache Airflow," Amazon, 8 May 2019. [Online]. Available: Build end-to-end machine learning workflows with Amazon SageMaker and Apache Airflow. [Accessed 30 March 2021].

[37] G. King, "Maximizing Recovery Factors: Improving Recovery Factors In Liquids-Rich Resource Plays Requires New Approaches," The American Oil & Gas Reporter, March 2014. [Online]. Available: https://www.aogr.com/magazine/editors-choice/improving-recovery-factors-in-liquids-rich-resource-plays-requires-new-appr. [Accessed 30 March 2021].

[38] S. Mishra and L. Lin, "Application of Data Analytics for Production Optimization in Unconventional Reservoirs: A Critical Review," in *Unconventional Resources Technology Conference*, Austin, Texas, 2017.

[39] J. Feblowitz, "Analytics in Oil and Gas: The Big Deal About Big Data," in *SPE Digital Energy Conference and Exhibition*, The Woodlands, Texas, 2013.

[40] A. Baaziz and L. Quoniam, "How to use Big Data technologies to optimize operations in Upstream Petroleum Industry," in *21st World Petroleum Congress*, Moscow, Russia, 2014.

[41] A. Abubakar, *Potential and challenges of applying artificial intelligence and machine-learning methods for geoscience,* Houston, Texas: Society of Exploration Geophysicists, 2020.

[42] Holdaway, K., Harnessing Oil and Gas Big Data with Analytics, Wiley, 2014.

[43] A. Shahkarami, S. Mohaghegh and Y. Hajizadeh, "Assisted History Matching Using Pattern Recognition Technology," in *Digital Energy Conference and and Exhibition*, The Woodlands, Texas, 2015.

[44] National Energy Technology Laboratory, "Data Analytics and Machine Learning Panel," in *Mastering the Subsurface Through Technology Innovation, Partnerships, and Collaboration: Carbon Storage and Oil and Natural Gas Technologies Review Meeting*, Pittsburgh, Pennsylvania, 2018.

[45] International Energy Agency, "Digitalisation and Energy," November 2017. [Online]. Available: https://www.iea.org/reports/digitalisation-and-energy. [Accessed 10 May 2020].

[46] M. Ratner and M. Tiemann, "An Overview of Unconventional Oil and Natural Gas: Resources and Federal Actions," Congressional Research Service, Washington, D.C., 2015.

[47] S. Esmaili and S. Mohaghegh, "Full field reservoir modeling of shale assets using advanced data-driven analytics," *Geoscience Frontiers,* vol. 7, pp. 11-20, 2016.

[48] J. Arthur, B. Bohm, B. Coughlin and M. Layne, "Evaluating Implications of Hydraulic Fracturing in Shale Gas Reservoirs," in *2009 SPE Americas E&P Environmental & Safety Conference*, San Antonio, Texas, 2009.

[49] N. Solano, C. Clarkson, F. Krause, S. Aquino and A. Wiseman, "On the Characterization of Unconventional Oil Reservoirs," *Canadian Society of Exploration Geophysicists Recorder,* vol. 38, no. 4, pp. 43-47, 2013.

[50] U.S. Energy Information Administration, "Assumptions to the Annual Energy Outlook 2019: Oil and Gas Supply Module," U.S. Department of Energy, Washington, D.C., 2019.

[51] C. McGlade, J. Speirs and S. Sorrell, "Methods of estimating shale gas resources - Comparison, evaluation and implications," *Energy,* vol. 59, no. 15, pp. 116-125, 2013.

[52] S. Mohaghegh, "A Critical Review of Current State of Reservoir Modeling of Shale Assets," in *Society of Petroleum Engineers Eastern Regional Conference and Exhibition*, Pittsburgh, Pennsylvania, 2013.

[53] M. Mohammadpoor and F. Torabi, "Big Data analytics in oil and gas industry: An emerging trend," *Petroleum,* pp. In Press, Corrected Proof, 2018.

[54] A. Bahga and V. Madisetti, Big Data Science & Analytics: A Hands-On Approach, VPT, 2016.

[55] R. LaFollette, G. Izadi and M. Zhong, "Application of Multivariate Analysis and Geographic Information Systems Pattern-Recognition Analysis to Produce Results in the Bakken Light Oil Play," The Woodlands, Texas, 2013.

[56] C. Shih, D. Vikara, A. Venkatesh, A. Wendt, S. Lin and D. Remson, "Evaluation of Shale Gas Production Drivers by Predictive Modeling on Well Completion, Production, and Geologic Data," National Energy Technology Laboratory, Pittsburgh, Pennsylvania, 2018.

[57] S. Mohaghegh, R. Gaskari and M. Maysami, "Shale Analytics: Making Production and Operational Decisions Based on Facts: A Case Study in the Marcellus Shale," The Woodlands, Texas, 2017.

[58] R. Smith, T. Mukerji and T. Lupo, "Correlating geologic and seismic data with unconventional resource production curves using machine learning," *Geophysics,* vol. 84, no. 2, pp. O39-O47, 2019.

[59] J. Montgomery and F. O'Sullivan, "Spatial variability of tight oil well productivity and the impact of technology," *Applied Energy,* pp. 334-355, 2017.

[60] J. Browning, S. Tinker, S. Ikonnikova, G. Gulen, E. Potter, Q. Fu, S. Horvath, T. Patzek, F. Male, W. Fisher, F. Roberts and K. Medlock III, "Barnett study determines full-field reserves, production forecast," Oil & Gas Journal, 2013.

[61] S. Ikonnikova, J. Browning, G. Gulen, K. Smye and S. Tinker, "Factors influencing shale gas production forecasting: Empirical studies of Barnett, Fayetteville, Haynesville, and Marcellus Shale plays," *Economics of Energy & Environmental Policy,* vol. 4, no. 1, pp. 19-35, 2015.

[62] P. Ghahfarokhi, T. Carr, S. Bhattacharya, J. Elliott, A. Shahkarami and K. Martin, "A Fiber-optic Assisted Multilayer Preceptron Reservoir Production Modeling: A Machine Learning Approach in Prediction of Gas Production from the Marcellus Shale," in *Unconventional Resources Technology Conference*, Houston, Texas, 2018.

[63] Q. Zhou, R. Dilmore, A. Kleit and J. Wang, "Evaluating gas production performances in Marcellus using data mining technologies," *Journal of Natural Gas Science Engineering,* vol. 20, pp. 109-120, 2014.

[64] J. Schuetter, S. Mishra, M. Zhong and R. LaFollette, "Data Analytics for Production Optimization in Unconventional Reservoirs," in *Unconventional Resources Technology Conference*, San Antonio, Texas, 2015.

[65] R. LaFollette, W. Holcomb and J. Aragon, "Impact of completion system, staging, and hydraulic fracturing trends in the Bakken formation of the eastern Williston Basin," in *Society of Petroleum Engineers Hydraulic Fracturing Technology Conference*, The Woodlands, Texas, 2012.

[66] O. Awoleke and R. Lane, "Analysis of data from the Barnett shale using conventional statistical and virtual intelligence techniques," *SPE Reservoir Evaluation & Engineering,* vol. 14, no. 5, pp. 544-556, 2011.

[67] Leathers-Miller, H., "Procedure for Calculating Estimated Ultimate Recoveries of Wells in the Missippian Barnett Shale, Bend Arch-Fort Worth Basin Province of North-Central Texas," United States Geological Survey Investigations Report 2017–5102, Reston, Virginia, 2017.

[68] J. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics,* vol. 29, pp. 1189-1232, 2001.

[69] Enverus - DrillingInfo, "Enverus - DrillingInfo," 2019. [Online]. Available: https://info.drillinginfo.com. [Accessed 21 January 2019].

[70] U.S. Energy Information Administration, "Marcellus Shale Play - Geology Overview," U.S. Department of Energy, Washington, D.C., 2017.

[71] W. Zagorski, M. Emery and J. Ventura, "The Marcellus Shale Play: Its Discovery and Emergence as a Major Global Hydrocarbon Accumulation," *in R.K. Merrill and C.A. Sternbach, eds, Giant fields of the decade 2000-2010: AAPG Memoir,* vol. 113, pp. 55-90, 2017.

[72] U.S. Energy Information Administration, "Annual Energy Outlook 2016 with projections to 2040," U.S. Department of Energy, Washington, D.C., 2016.

[73] T. Inks, T. Engelder, E. Jenner, B. Golob, J. Hocum and D. O'Brein, "Marcellus fracture characterizaion using P-wave azimuthal velocity attributes: Comparison with production and outcrop data," *Interpretation,* vol. 3, no. 3, pp. SU1 - SU15, 2015.

[74] W. Zagorski, D. Bowman, M. Emery and G. Wrightstone, "An overview of Some Key Factors Controlling Well Productivity in Core Areas of the

Appalachian Basin Marcellus Shale Play," in *American Association of Petroleum Geologists Search and Discovery Article #110147*, Houston, Texas, USA, 2011.

[75] K. Carter, J. Harper, K. Schmid and J. Kostelnik, "Unconventional natural gas resources in Pennsylvania: The backstory of the modern Marcellus Shale play," *Environmental Geosciences,* vol. 18, no. 4, pp. 217-257, 2011.

[76] G. Gullickson, K. Fiscus and P. Cook, "Completion Influence on Production Decline in the Bakken/Three Forks Play," in *SPE Western North American and Rocky Mountain Joint Regional Meeting*, Denver, Colorado, 2014.

[77] C. Randle, C. Bond, R. Lark and A. Monaghan, "Uncertainty in geological interpretations: Effectiveness of expert elicitations," *Geosphere,* vol. 15, no. 1, pp. 108-118, 2019.

[78] DrillingInfo, "Pre-calculated, Proprietary EUR Database from DrillingInfo White Paper," DrillingInfo, 2016.

[79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825--2830, 2011.

[80] J. Friedman, "Stochastic Gradient Boosting," 26 March 1999. [Online]. Available: https://statweb.stanford.edu/~jhf/ftp/stobst.pdf. [Accessed 24 March 2019].

[81] J. Elith, R. Leathwick and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology,* vol. 77, pp. 802-813, 2008.

[82] V. Smolyakov, "Ensemble Learning to Improve Machine Learning Results: How ensemble methods work: bagging, boosting and stacking," 2017. [Online]. Available: https://blog.statsbot.co/ensemble-learning-d1dcd548e936. [Accessed 24 March 2019].

[83] S. Kotsiantis, "Decision trees: a recent overview," *Artificial Intelligence Review,* vol. 39, no. 4, pp. 261-283, 2013.

[84] L. Rokach and O. Maimon, Data Mining With Decision Trees: Theory and Applications, 2nd ed., River Edge, New Jersey, United States: World Scientific Publishing Company, Inc., 2014.

[85] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Predication, New York, New York, United States: Springer-Verlag, 2001.

[86] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees, 1 ed., Montery, California, United States: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[87] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[88] S. Deswal and M. Pal, "Artificial neural network based modeling of evaporation losses in reservoirs," *International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering,* vol. 2, no. 3, pp. 18-22, 2008.

[89] F. Hutter, H. Hoos and K. Leyton-Brown, "An Efficient Approach for Assessing Hyperparameter Importance," in *31st International Conference on Machine Learning,* Beijing, China, 2014.

[90] J. Friedman and J. Meulman, "Multiple additive regression trees with application in epidemiology," *Statistics in Medicine,* vol. 22, pp. 1365-1381, 2003.

[91] C. Molnar, Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Lean Publishing, 2019.

[92] J. Friedman and B. Popescu, "Predictive Learning via Rule Ensembles," 5 October 2005. [Online]. Available: http://statweb.stanford.edu/~jhf/ftp/RuleFit.pdf. [Accessed 29 June 2019].

[93] Q. Zhao and T. Hastie, "Casual Interpretations of Black-Box Models," 2017. [Online]. Available: https://web.stanford.edu/~hastie/Papers/pdp_zhao.pdf. [Accessed 8 April 2019].

[94] W. Zagorski, G. Wrightstone and D. Bowman, "The Appalachian Basin Marcellus gas play: Its history of development, geologic controls on production, and future potential as a world-class reservoir," *in J.A. Breyer, ed., Shale reservoirs -Giant resources for the 21st century: AAPG Memoir,* vol. 97, pp. 172-200, 2012.

[95] Marcellus Center for Outreach and Research, "Thickness of Marcellus," 2019. [Online]. Available: http://www.marcellus.psu.edu/resources-maps-graphics-videos.html. [Accessed 3 August 2019].

[96] J. Arthur, "The Marcellus and Utica Shales: Geologic Considerations," in *Marcellus Shale Summit*, Harrisburg, Pennsylvania, United States, 2011.

[97] J. Perrin, "Horizontally drilled wells dominate U.S. tight formation production," U.S. Energy Information Administration, 6 June 2019. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=39752. [Accessed 11 June 2020].

[98] U.S. Energy Information Administration, "Trends in U.S. Oil and Gas Upstream Costs," U.S. Department of Energy, Washington, D.C., 2016.

[99] S. Randolph and J. McBride, "AI & Machine Learning: The Next Transformation for Oil & Gas," Opportune, January 2019. [Online]. Available: https://opportune.com/Energy-Sector-Insights-Events/Insights/AI-Machine-Learning-The-Next-Transformation-for-Oil-Gas/. [Accessed 28 December 2019].

[100] N. Tamimi, S. Samani, M. Minaei and F. Harirchi, "An Artificial Intelligence Decision Support System for Unconventional Field Development Design," in *Unconventional Resources Technology Conference (URTeC)*, Denver, Colorado, 2019.

[101] S. Chen, W. Zhao, Y. Ouyang, Q. Zeng, Q. Yang, H. Hou, S. Gai, S. Bao and X. Li, "Prediction of sweet spots in shale reservoir based on geophysical well logging and 3D seismic data: A case study of Lower Silurian Longmaxi Formation in W4 block, Sichuan Basin, China," *Energy Exploration & Exploitation*, vol. 35, no. 2, pp. 147-171, 2017.

[102] W. Liu, G. Zhang, J. Cao, J. Zhang and G. Yu, "Combined petrophysics and 3D seismic attributes to predict shale reservoirs favorable areas," *Journal of Geophysics and Engineering*, vol. 16, pp. 974-991, 2019.

[103] X. Zhao, X. Pu, F. Jin, W. Han, Z. Shi, A. Cai, A. Wang, Q. Guan, W. Jiang and W. Zhang, "Geological characteristics and key exploration technologies of continental shale oil sweet spots: A case study of Member 2 of Kongdian Formation in the Cangdong sag in Huanghua depression, Bohai Bay Basin," *Petroleum Research,* vol. 4, pp. 97-112, 2019.

[104] M. Chapman, S. Maultzsch, E. Liu and X. Li, "The effect of fluid saturation in an anisotropic multi-scale equant porosity model," *Journal of Applied Geopyhsics*, vol. 54, no. 3-4, pp. 191-202, 2003.

[105] S. Mo, Y. Zhu, N. Zabaras, X. Shi and J. Wu, "Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media," *Water Rresources Research,* vol. 55, pp. 703-728, 2019.

[106] P. Bestagini, V. Lipari and S. Tubaro, "A machine learning approach to facies classification using well logs," in *Society of Exploration Geophysicists - International Exposition and Annual Meeting*, Houston, Texas, 2017.

[107] C. Kurison, H. Sadi Kuleli and M. Mubarak, "Unlocking well productivity drivers in Eagle Ford and Utica unconventional resources through data analytics," *Journal of Natural Gas Science and Engineering,* vol. 71, pp. 1-24, 2019.

[108] B. Willigers, S. Begg and R. Bratvold, "Combining geostatistics with bayesian updating to continually optimize drilling strategy in shale-gas plays," *SPE Reservoir Evaluation & Engineering,* vol. 17, no. 04, pp. 507-519, 2014.

[109] B. Willigers, S. Begg and R. Bratvold, "Hot spot hunting: Optimising the staged development of shale plays," *Journal of Petroleum Science and Engineering,* vol. 146, pp. 553-563, 2017.

[110] Z. Chen, C. Yang, C. Jiang, D. Kohlruss, K. Hu, X. Liu and M. Yurkowski, "Production characteristics and sweet-spots mapping of the Upper Devonian-Lower Missippian Bakken Formation tight oil in southeastern Saskatchewan, Canada," *Petroleum Exploration and Development,* vol. 45, no. 4, pp. 662-672, 2018.

[111] H. Wang, "What Factors Control Shale-Gas Production and Production-Decline Trend in Fractured Systems: A Comprehensive Analysis and Investigation (SPE-179967-PA)," *SPE Journal,* vol. 22, no. 2, pp. 562-581, 2017.

[112] M. Zobak and D. Arent, "Shale Gas: Development Opportunities.," *The Bridge on Emerging Issues in Earth Resources Engineering,* vol. 44, no. 1, pp. 16-23, 2014.

[113] A. Dayal and D. Mani, Shale Gas Exploration and Environmental and Economic Impacts, Hyderabad, India: Elsevier, 2017.

[114] Z. Jiang, W. Zhang, C. Liang, Y. Wang, H. Liu and X. Chen, "Basic characteristics and evaluation of shale oil reservoirs," *Petroleum Research,* vol. 2, pp. 149-163, 2016.

[115] W. Fertl and G. Chillngar, "Total Organic Carbon Content Determined from Well Logs," *SPE Formation Evaluation,* pp. 407-419, 1988.

[116] K. Heslop, "Generalized Method for the Estimation of TOC from GR and Rt," in *American Association of Petroleum Geologists Search and Discovery Article #80117*, New Orleans, Louisiana, 2010.

[117] M. Kamel and W. Mabrouk, "Estimation of shale volume using a combination of the three porosity logs," *Journal of Petroleum Science and Engineering,* vol. 40, no. 3-4, pp. 145-157, 2003.

[118] O. Serra, "Developments in Petroleum Science - Chapter 3: The Measurement of Resistivity," in *Fundamentals of well-log Interpretation*, Amsterdam, Netherlands, Elsevier, 1984, pp. 51-76.

[119] G. Karthikeyan, A. Kumar, A. Shrivastava and M. Srivastava, "Overpressure estimation and productivity analysis for a Marcellus Shale gas reservoir, southwest Pennsylvania: A case study," *The Leading Edge,* vol. 37, no. 5, pp. 344-349, 2018.

[120] Y. Tang, Y. Xing, L. Li, B. Zhang and S. Jiang, "Influence factors and evaluation methods of the gas shale fracability," *Earth Science Frontiers,* vol. 26, no. 4, pp. 356-363, 2012.

[121] J. Richardson and W. Yu, "Calculation of Estimated Ultimate Recovery and Recovery Factors of Shale-Gas Wells Using a Probabilistic Model of Original Gas in Place," *SPE Reservoir Evaluation & Engineering,* pp. 638-653, 2018.

[122] D. Vikara, D. Remson and V. Khanna, "Gaining Perspective on Unconventional Well Design Choices through Play-level Application of Machine Learning Modeling," *Upstream Oil and Gas Technology*, vol. 4, pp. 1-18, 2020.

[123] DrillingInfo, "DrillingInfo," 2019. [Online]. Available: https://info.drillinginfo.com. [Accessed 21 January 2019].

[124] D. Kargbo, R. Wilhelm and D. Campbell, "Natural gas plays in the Marcellus Shale: Challenges and Potential Opportunities," *Environmental Science Technology,* vol. 44, no. 15, pp. 5679-5684, 2010.

[125] H. King, "Marcellus Shale - Appalachian Basin Natural Gas Play," Geology.com, Undated. [Online]. Available: https://geology.com/articles/marcellus-shale.shtml. [Accessed 13 June 2020].

[126] T. Engelder, "Marcellus 2008: Report card on breakout year for gas production in the Appalachian Basin," *Fort Worth Basin Oil and Gas Magazine,* pp. 19-22, 2009.

[127] S. Ikonnikova, K. Smye, J. Browning, R. Dommisse, G. Gülen, S. Hamlin, S. Tinker, F. Male, G. McDaid and E. Vankov, "Final Report on Update and Enhancement of Shale Gas Outlooks," University of Texas at Austin Bureau of Economic Geology for the U.S. Department of Energy, Austin, Texas, 2018.

[128] M. Boyce and T. Carr, "Lithostratigraphy and Petrophysics of the Devonian Marcellus Interval in West Virginia and Southwestern Pennsylvania," 18 October 2009. [Online]. Available: http://www.unconventionalenergyresources.com/marcellusLithoAndPetroPaper.pdf. [Accessed 29 November 2019].

[129] R. Milici and C. Swezey, "Assessment of Appalachian Basin Oil and Gas Resources: Devonian Shale - Middle and Upper Paleozoic Total Petroleum System," U.S. Geological Survey Open-File Report 2006-1237, Reston, Virginia, 2006.

[130] T. Engelder and G. Lash, "Marcellus Shale Play's Vast Resource Potential Creating a Stir in Appalachia," *The American Oil & Gas Reporter,* vol. 51, no. 6, pp. 76-87, 2008.

[131] C. Zou, Q. Zhao, D. Dong, Z. Yang, Z. Qiu, F. Liang, N. Wang, Y. Huang, A. Duan, Q. Zhang and Z. Hu, "Geologic characteristics, main challenges and future prospect of shale gas," *Journal of Natural Gas Geoscience,* vol. 2, pp. 273 - 288, 2017.

[132] D. Soeder, P. Randolph and R. Matthews, "Porosity and Permeability of Eastern Devonian Gas Shale," Institute of Gas Technology - prepared for U.S. Department of Energy, Morgantown Energy Technology Center, Morgantown, West Virginia, 1986.

[133] W. Zhang, C. Wu, H. Zhong, Y. Li and L. Wang, "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," *Geoscience Frontiers,* vol. Article in Press, pp. 1-9, 2020.

[134] J. Ogutu, P. Hans-Peter and T. Schulz-Streek, "A comparison of random forests, boosting and support vector machines for genomic selection," *BMC Proceedings,* vol. 5, no. 3:S11, pp. 1-5, 2011.

[135] S. Amir Naghibi, H. Hashemi, R. Berndtsson and S. Lee, "Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors," *Journal of Hydrology,* vol. 589, no. 125197, 2020.

[136] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys,* vol. 4, pp. 40-79, 2010.

[137] L. Devroye and T. Wagner, "Distribution-free performance bounds for potential function rules," *IEEE Transaction in Information Theory,* vol. 25, no. 5, p. 601–604, 1979.

[138] M. de Berg, O. Cheong, M. van Kreveld and M. Overmars, Computational Geometry: Algorithms and Applications - Third Edition, Berlin, Germany: Springer-Verlag, 2008.

[139] N. Miller, "Characterization the Productive Limit of the Northeast Pennsylvania Marcellus Dry Gas Window: An Investigation of Low Resistivity Along the Line of Death," Texas A&M University, College Station, Texas, United States, 2017.

[140] C. Laughrey, "Black Shale Diagenesis: Insights from Integrated High-Definition Analyses of Post-Mature Marcellus Formation Rocks, Northeastern Pennsylvania," *American Association of Petroleum Geology Search and Discovery Article #110150,* 2011.

[141] A. Abramov, "Optimization of well pad design and drilling - well clustering," *Petroleum Exploration and Development,* vol. 46, no. 3, pp. 614-620, 2019.

[142] M. McKay, R. Beckman and W. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics,* vol. 21, no. 2, pp. 239-245, 1979.

[143] B. Ayyub and K. Lai, "Selective sampling in simulation-based reliability assessment," *International Journal of Pressure Vessels and Piping,* vol. 46, no. 2, pp. 229-249, 1991.

[144] D. Vikara, D. Remson and V. Khanna, "Supplementary Data for "Machine Learning-informed Ensemble Framework for Evaluating Shale Gas Production Potential: Case Study in the Marcellus Shale"," *Mendeley Data,* vol. V1, no. http://dx.doi.org/10.17632/vbgywdcdjp.1, 2020.

[145] S. Mlada, "Permian Midland Review: Acerage high grading and breakeven prices," March 2017. [Online]. Available: https://www.rystadenergy.com/newsevents/news/newsletters/UsArchive/shale-newsletter-march-2017/. [Accessed 2 November 2019].

[146] L. Chorn, J. Serice and S. Rosario-Davis, "Using High-Grading and Portfolio Tools to Allocate Resources Among Shale Play Opportunities," in *SPE/CSUR Unconventional Resources Conference – Canada,* Calgary, Alberta, Canada, 2014.

[147] L. Chorn, "Where Should I Put The Next Well In The Shale Play?," 13 November 2013. [Online]. Available: https://halliburtonblog.com/where-should-i-put-the-next-well-in-the-shale-play/. [Accessed 2 November 2019].

[148] W. Kruskal and W. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association,* vol. 47, no. 260, pp. 583-621, 1952.

[149] O. Dunn, "Multiple Comparisons Using Rank Sums," *Technometrics,* vol. 6, pp. 241-252, 1964.

[150] O. Dunn, "Multiple comparisons among means," *Journal of American Statistical Association,* vol. 56, pp. 52-64, 1961.

[151] A. Dinno, "Nonparametric pairwise multiple comparisons in independent groups using Dunn's test," *The Strata Journal,* vol. 15, no. 1, pp. 292-300, 2015.

[152] S. Orlich, " Kruskal - Wallis Multiple Comparison with a MINITAB Macro Dunn's Test," 2010. [Online]. Available: https://support.minitab.com/en-us/minitab/18/macro-library/macro-files/nonparametrics-macros/krusmc/. [Accessed 13 May 2020].

[153] J. Zhang, L. Lin, Y. Li, X. Tang, L. Zhu, Y. Xing, S. Jiang, T. Jing and S. Yang, "Classification and evaluation of shale oil," *Earth Science Frontiers,* vol. 19, no. 5, pp. 322-331, 2012.

[154] J. Arthur, B. Langhus and D. Alleman, "An overview of modern shale gas development in the United States," 2008. [Online]. Available: http://www.all-llc.com/publicdownloads/ALLShaleOverviewFINAL.pdf. [Accessed 20 October 2019].

[155] G. Lash and T. Engelder, "Tracking the burial and tectonic history of Devonian shale of the Appalachian Basin by analysis of joint intersection style," *Geological Society of America Bulletin,* vol. 121, pp. 265-277, 2009.

[156] T. Engelder and A. Whitaker, "Early jointing in coal and black shale: Evidence for an Appalachian-wide stress field as a prelude to the Alleghanian orogeny," *Geology,* vol. 34, pp. 581-584, 2009.

[157] R. Zeits, "Cabot's Macellus Wells Getting Bigger: 27 Bcf On Average in 2016," Seeking Alpha, 28 March 2016. [Online]. Available: https://seekingalpha.com/article/3961412-cabots-marcellus-wells-getting-bigger-27-bcf-on-average-in-2016. [Accessed 27 December 2019].

[158] Range Resources, "Company Presentation," Range Resources, 26 July 2016. [Online]. Available: https://www.slideshare.net/MarcellusDN/range-resources-company-presentation-july-2016. [Accessed 27 December 2019].

[159] C. Bishop, Pattern Recognition and Machine Learning, New York, New York: Springer, 2006.

[160] A. Jahandideh and B. Jafarpour, "Optimization of hydraulic fracturing design under spatially variable shale fracability," *Journal of Petroleum Science & Engineering,* vol. 174, pp. 174-188, 2016.

[161] B. Bonnell and C. Hurich, "Characterization of Reservoir Heterogeneity: An Investigation of the Role of Cross-Well Reflection Data," *CSEG Recorder,* vol. 33, no. 2, pp. 31-37, 2008.

[162] Q. Li, H. Zing, J. Liu and X. Liu, "A review on hydraulic fracturing of unconventional reservoir," *Petroleum*, vol. 1, pp. 8-15, 2015.

[163] B. Aadnøy and R. Looyeh, Petroleum Rock Mechanics - 2nd Edition, Gulf Professional Publishing, 2019.

[164] United States Geological Survey, "What is hydraulic fracturing?," U.S. Department of the Interior, Undated. [Online]. Available: https://www.usgs.gov/faqs/what-hydraulic-fracturing?qt-news_science_products=0#qt-news_science_products. [Accessed 21 November 2020].

[165] J. Hyman, J. Jiménez-Martínez, H. Viswanathan, J. P. M. Carey, E. Rougier, S. Karra, Q. Kang, L. Frash, L. Chen, Z. Lei, D. O'Malley and N. Makedonska, "Understanding hydraulic fracturing: amulti-scale problem," *Philosophical transactions. Series A, Mathematical, Physical, and Engineering Sciences,* vol. 374, no. 20150426, pp. 1-16, 2016.

[166] F. Aminzadeh, "Hydraulic Fracturing, An Overview," *Journal of Sustainable Energy Engineering,* vol. 6, no. 3, pp. 204-228, 2019.

[167] U.S. Environmental Protection Agency, "The Process of Unconventional Natural Gas Production," U.S. EPA, 22 January 2020. [Online]. Available: https://www.epa.gov/uog/process-unconventional-natural-gas-production. [Accessed 21 November 2020].

[168] D. Van Wagener and F. Aloulou, "Tight oil development will continue to drive future U.S. crude oil production," U.S. Energy Information Administration, 28 March 2019. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=38852. [Accessed 12 December 2020].

[169] D. Vikara, D. Remson and V. Khanna, "Machine learning-informed ensemble framework for evaluating shale gas production potential: Case study in the

Marcellus Shale," *Journal of Natural Gas Science and Engineering,* vol. 84, no. 103679, pp. 1-21, 2020.

[170] U.S. Department of Energy, "Quadrennial Technology Review 2015 - Chapter 7: Advancing Systems and Technologies to Produce Clearner Fuels," Washington, D.C., 2015.

[171] R. Mehrotra and R. Gopalan, "Factors Influencing Strategic Decision-Making Process for the Oil/Gas Industris of UAE - A study," *International Journal of Marketing & Financial Management,* vol. 5, no. 1, pp. 62-69, 2017.

[172] S. Wang, Z. Chen and S. Chen, "Applicability of deep neural networks on production forecasting in Bakken shale reservoirs," *Journal of Petroleum Science and Engineering,* vol. 179, pp. 112-125, 2019.

[173] L. Jie, C. Junxing and Y. Jiachun, "Prediction on daily gas production of single well based on LSTM," in *SEG 2019 Workshop: Mathematical Geophysics: Traditional vs Learning,* Beijing, China, 2019.

[174] A. Sagheer and M. Kotb, "Time series forecasting of petroleum production using deep LSTM recurrent networks," *Neurocomputing,* vol. 323, pp. 203-213, 2019.

[175] W. Liu, W. Liu and J. Gu, "Forecasting oil production using ensemble empirical model decomposition based Long Short-Term Memory neural network," *Journal of Petroleum Science and Engineering,* vol. 189, p. 107013, 2020.

[176] U.S. Department of Energy, "Natural Gas Flaring and Venting: State and Federal Regulatory Overview, Trends, and Impacts," Office of Fossil Energy - Office of Oil and Natural Gas, Washington, D.C., 2019.

[177] G. Myhre, D. Shindell, F. Bréon, W. F. J. Collins, J. Huang, D. Koch, J. Lamarque, D. Lee and B. Mendoza, "Anthropogenic and Natural Radiative Forcing. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change," Cambridge University Press, Cambridge, United Kingdom and New York, New York, United States, 2013.

[178] U.S. Energy Information Administration, "Natural Gas Annual," U.S. Department of Energy, 30 September 2020. [Online]. Available: https://www.eia.gov/naturalgas/annual/. [Accessed 10 December 2020].

[179] United States Geological Survey, "ANSS Comprehensive Earthquake Catalog (ComCat) Documentation," U.S. Department of the Interior, Undated. [Online].

Available: https://earthquake.usgs.gov/data/comcat/. [Accessed 11 December 2020].

[180] B. Scanlon, R. Reedy, P. Xu, M. Engle, J. Nicot, D. Yoxtheimer, Q. Yang and S. Ikonnikova, "Can we beneficially reuse produced water from oil and gas extraction in the U.S.?," *Science of the Total Environment,* vol. 717, p. 137085, 2020.

[181] M. Kah, "Columbia Global Energy Dialogue: Natural Gas Flaring Workshop Summary," Columbia Center on Global Energy Policy, 30 April 2020. [Online]. Available: https://www.energypolicy.columbia.edu/research/global-energy-dialogue/columbia-global-energy-dialogue-natural-gas-flaring-workshop-summary. [Accessed 12 December 2020].

[182] L. van Bedolla, W. Cai, Z. Martin and F. Yu, "Technology and Policy Solutions to Reduce Harmful Natural Gas Flaring," Columbia University School of International and Public Affairs , New York, New York, 2020.

[183] Oil & Gas Journal, "Permian gas flaring, venting reaches record high," 4 June 2019. [Online]. Available: https://www.ogj.com/general-interest/hse/article/17279037/permian-gas-flaring-venting-reaches-record-high. [Accessed 31 July 2020].

[184] Texas Independent Producers and Royalty Owners Association, "A Decade of the Permian Basin," Austin, Texas, 2020.

[185] The American Oil & Gas Reporter, "Importance of Permian Basin is Delineated in TIPRO Report," February 2020. [Online]. Available: https://www.aogr.com/magazine/markets-analytics/importance-of-permian-basin-is-delineated-in-tipro-report. [Accessed 26 July 2020].

[186] M. McEwen, "Wood Mackenzie analysts: Permian faces multiple challenges," MRT.com, 28 July 2019. [Online]. Available: https://www.mrt.com/business/oil/article/Wood-Mackenzie-analysts-Permian-faces-multiple-14149600.php#photo-17926034. [Accessed 31 July 2020].

[187] D. Vaucher, "No Free Lunch – The Water Challenges Facing Operating Companies in the Permian Basin," IHS Markit, 4 November 2019. [Online]. Available: https://ihsmarkit.com/research-analysis/no-free-lunch-the-water-challenges-facing-companies-permian.html. [Accessed 31 July 2020].

[188] S. Rassenfoss, "Rising Tide of Produced Water Could Pinch Permian Growth," Journal of Petroleum Technology, 12 June 2018. [Online]. Available: https://pubs.spe.org/en/jpt/jpt-article-detail/?art=4273. [Accessed 29 November 2020].

[189] Railroad Commission of Texas, "Permian Basin Information," Railroad Commission of Texas, 11 November 2020. [Online]. Available: https://www.rrc.state.tx.us/oil-gas/major-oil-and-gas-formations/permian-basin-information/. [Accessed 25 November 2020].

[190] U.S. Energy Information Administration, "Permian Basin Part 2: Wolfcamp Shale Play of the Midland Basin - Geology Review," U.S. Department of Energy, Washington, D.C., 2020.

[191] U.S. Energy Information Administration, "U.S. Crude Oil and Natural Gas Proved Reserves, Year end-2018," U.S. Department of Energy, 13 December 2019. [Online]. Available: https://www.eia.gov/naturalgas/crudeoilreserves/. [Accessed 25 November 2020].

[192] U.S. Energy Information Administration, "Permian Basin Wolfcamp Shale Play: Geology Review," U.S. Department of Energy, Washington, D.C., 2018.

[193] T. Hoak, K. Sundberg and P. Oroleva, "Overview of the Structural Geology and Tectonics of the Central Basin Platform, Delaware Basin, and Midland Basin, West Texas and New Mexico," U.S. Department of Energy, Washington, D.C., 1998.

[194] K. Yang and S. Dorobeck, "The Permian Basin of West Texas and New Mexico: Tectonic History of a “Composite” Foreland Basin and its Effects on Stratigraphic Development," in *Stratigraphic Evolution of Foreland Basins, Volume 52,* SEPM Society for Sedimentary Geology, 1995.

[195] J. Roberts, "GDS Geological Column: Geological Data Service," Dallas, Texas, 1989.

[196] University of Texas at Austin, "Wolfberry and Spraberry Play Of The Midland Basin," Bureau of Economic Geology, Undated. [Online]. Available: http://www.beg.utexas.edu/research/programs/starr/unconventional-resources/wolfberry-spraberry. [Accessed 2 September 2020].

[197] G. Wilson, "Midland Basin Wolfcamp Horizontal Development," in *AAPG DPA Forum Midland Playmaker*, Midland, Texas, 2015.

[198] R. King & Co., "Permian Basin Strategraphic Charts & Province Map," Undated. [Online]. Available: https://rkingco.com/wp-content/uploads/2014/12/PermianBasinStratChart.jpg. [Accessed 2 September 2020].

[199] H. Hamlin and R. Baumgardner, Wolfberry (Wolfcampian-Leonardian) Deep-Water Depositional Systems in the Midland Basin: Stratigraphy, Lithofacies,

Reservoirs, and Source Rocks, Austin, Texas: Part Number RI0277, University of Texas Bureau of Economic Geology, 2012.

[200] G. Schmitt, "Genesis and Depositional History of Spraberry Formation, Midland Basin, Texas," *AAPG Bulletin,* vol. 38, no. 9, pp. 1957-1978, 1954.

[201] G. Hunter, B. Šegvić, G. Zanoni, S. Omodeo-Salé and T. Adatte, "Evaluation of Shale Source Rocks and Clay Mineral Diagenesis in the Permian Basin, USA: Inferences on Basin Thermal Maturity and Source Rock Potential," *geosciences,* vol. 10, no. 10, pp. 1-32, 2020.

[202] A. James, "Evaluating and Hy-Grading Wolfcamp Shale Opportunities in the Midland Basin," AAPG Search and Discovery Article #110213, Adapted from presentation at the AAPG DPA Forum Midland Playmaker, Midland, Texas, 2015.

[203] C. Handford, "Sedimentology and Genetic Stratigraphy of Dean and Spraberry Formations (Permian), Midland Basin, Texas," *AAPG Bulletin,* vol. 65, no. 9, pp. 1602-1616, 1981.

[204] L. Sutton, "Permian Basin Geology: The Midland Basin vs. the Delaware Basin Part 2," Enverus, 23 December 2014. [Online]. Available: https://www.enverus.com/blog/permian-basin-geology-midland-vs-delaware-basins/. [Accessed 11 November 2020].

[205] J. Lorenz, J. Sterling, D. Schechter, C. Whigham and J. Jensen, "Natural fractures in the Spraberry Formation, Midland basin, Texas: The effects of mechanical stratigraphy on fracture variability and reservoir behavior," *AAPG Bulletin,* vol. 86, no. 3, pp. 505-524, 2002.

[206] J. Marshall, "Spraberry Reservoir of West Texas1: GEOLOGICAL NOTES," *AAPG Bulletin,* vol. 36, no. 11, pp. 2189-2191, 1952.

[207] B. Shattuck, "Spraberry Fields Forever," Forbes, 8 September 2017. [Online]. Available: https://www.forbes.com/sites/woodmackenzie/2017/09/08/spraberry-fields-forever/?sh=245b4309655a. [Accessed 26 November 2020].

[208] R. Murphy, "Depositional Systems Interpretation of Early Permian mixed Siliciclastics and Carbonates, Midland Basin, Texas," Master's Thesis - University of Indiana, Bloomington, Indiana, 2015.

[209] S. Gaswirth, "Assessment of Undiscovered Continuous Oil and Gas Resources in the Wolfcamp Shale of the Midland Basin, West Texas," in *AAPG Annual Convention and Exhibition*, Houston, Texas, 2017.

[210] U.S. Energy Information Administration, "EIA updates geological maps of Midland Basin's Wolfcamp formation," U.S. Department of Energy, 24 November 2020. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=46016. [Accessed 25 November 2020].

[211] A. Saller, A. Dickson and S. Boyd, "Cycle Stratigraphy and Porosity in Pennsylvanian and Lower Permian Shelf Limestones, Easten Central Basin Platform, Texas," *AAPG Bulletin*, vol. 78, no. 12, pp. 1820-1842, 1994.

[212] J. Peng, K. Milliken, Q. Fu, X. Janson and S. Hamlin, "Grain assemblages and diagenesis in organic-rich mudrocks, Upper Pennsylvanian Cline shale (Wolfcamp D), Midland Basin, Texas," *AAPG Bulletin*, vol. 104, no. 7, pp. 1593-1624, 2020.

[213] P. Blomquist, "Wolfcamp Horizontal Play Midland Basin, West Texas," IHS Markit, IHS Geoscience Webinar Series, 2016.

[214] U.S. Energy Information Administration, "The Wolfcamp play has been key to Permian Basin oil and natural gas production growth," U.S. Department of Energy, 16 November 2018. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=37532. [Accessed 25 November 2020].

[215] Enverus, "DrillingInfo Web App," 2020. [Online]. Available: https://www.enverus.com/products/di-web-app/. [Accessed 1 November 2020].

[216] University of Texas at Austin - Bureau of Economic Geology, "Integrated Synthesis of the Permian Basin: Data and Models for Recovering Existing and Undiscovered Oil Resources from the Largest OIl-Bearing Basin in the U.S.," Jackson School of Geosciences, 2008. [Online]. Available: http://www.beg.utexas.edu/resprog/permianbasin/gis.htm. [Accessed 9 September 2020].

[217] United States Geological Survery, "How to Use the National Map Services - Large Scale Base Map Dynamic Services," Undated. [Online]. Available: https://viewer.nationalmap.gov/help/HowTo.htm. [Accessed 2 September 2020].

[218] A. Kondash, N. Lauer and A. Vengosh, "The intensification of the water footprint of hydraulic fracturing," *Science Advances*, vol. 4, no. 8, pp. 1-8, 2018.

[219] R. Bruant, "Permian Water Outlook," B3 Insight, 26 February 2019. [Online]. Available: http://www.gwpc.org/sites/default/files/event-

sessions/Produced%20Water%20-%20Rob%20Bruant_0.pdf. [Accessed 12 December 2020].

[220] C. Leyden, "Satellite data confirms Permian gas flaring is double what companies report," Environmental Defense Fund, 24 January 2019. [Online]. Available: http://blogs.edf.org/energyexchange/2019/01/24/satellite-data-confirms-permian-gas-flaring-is-double-what-companies-report/. [Accessed 13 December 2020].

[221] A. Abramov and M. Bertelsen, "Permian gas flaring reaches yet another high," Rystad Energy, 5 November 2019. [Online]. Available: https://www.rystadenergy.com/newsevents/news/press-releases/permian-gas-flaring-reaches-yet-another-high/. [Accessed 24 December 2020].

[222] M. Agerton, B. Gilbert and G. Upton, "The Economics of Natural Gas Flaring in U.S. Shale: An Agenda for Research and Policy," Rice University's Baker Institute for Public Policy, Houston, Texas, 2020.

[223] J. Arps, "Analysis of Decline Curves," *Transactions of the AIME*, vol. 160, no. 1, pp. 228-247, 1945.

[224] J. Miller, "Short Report: Reaction Time Analysis with Outlier Exclusion: Bias Varies with Sample Size," *The Quarterly Journal of Experimental Psychology Section A*, vol. 43, no. 4, pp. 907-912, 1991.

[225] I. Ilyas and X. Chu, Data Cleaning, New York, New York: Association for Computing Machinery, 2019.

[226] DrillingInfo, "Pre-calculated, Proprietary EUR Database from DrillingInfo - White Paper," May 2016. [Online]. Available: https://www.enverus.com/wp-content/uploads/2017/11/WP_EUR_Customer-print.pdf. [Accessed 2020 November 2020].

[227] M. Fetkovich, E. Fetkovich and M. Fetkovich, "Useful Concepts for Decline Curve Forecasting, Reserve Estimation, and Analysis," *SPE Reservoir Engineering*, vol. 11, no. 1, pp. 13-22, 1996.

[228] E. Martin, "Behaviour of Arps Equation in Shale Plays," LinkedIn, 29 March 2015. [Online]. Available: https://www.linkedin.com/pulse/behavior-arps-equation-shale-plays-emanuel-mart%C3%ADn/. [Accessed 22 November 2020].

[229] R. Jimenez, "Using Decline Curve Analysis, Volumetric Analysis, and Baysian Methodology to Quantify Uncertainty in Shale Gas Reserves Estimates," Masters Thesis - Texas A&M University, College Station, Texas, 2012.

[230] U.S. Environmental Protection Agency, "Analysis of Hydraulic Fracturing Fluid Data from the FracFocus Chemical Disclosure Registry 1.0," U.S. EPA Office of Research and Development, Washington, D.C., 2015.

[231] T. Saba, F. Mohsen, M. Garry, B. Murphy and B. Hilbert, "White Paper Methanol Use in Hydraulic Fracturing," Exponent, Maynard, Massachusetts, 2012.

[232] R. Manchanda, P. Bhardwaj, J. Hwang and M. Sharma, "Parent-Child Fracture Interference: Explanation and Mitigation of Child Well Underperformance," in *Society of Petroleum Engineering Hydraulic Fracturing Technology Conference and Exhibition*, The Woodlands, Texas, 2018.

[233] A. Kumar, K. Shrivastava, B. Elliott and M. Sharma, "Effect of Parent Well Production on Child Well Stimulation and Productivity," in *Society of Petroleum Engineers Hydraulic Fracturing Technology Conference and Exhibition*, The Woodlands, Texas, 2020.

[234] N. Chithra Chakra, K. Song, M. Gupta and D. Saraf, "An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs)," *Journal of Petroleum Science and Engineering*, vol. 106, pp. 18-33, 2013.

[235] U.S. Energy Information Administration, "Maps: Oil and Gas Exploration, Resources, and Production," U.S. Department of Energy, 23 April 2020. [Online]. Available: https://www.eia.gov/maps/maps.htm#permian. [Accessed 25 November 2020].

[236] M. Shanker, M. Hu and M. Hung, "Effect of data standardization on neural network training," *Omega*, vol. 24, no. 4, pp. 385-397, 1996.

[237] Y. Kumar, K. Bello, S. Sharma, D. Vikara, D. Remson, D. Morgan and L. Cunha, "Neural Network-Based Surrogate Models for Joint Prediction of Reservoir Pressure and CO2 Saturation," in *2020 SMART Annual Review Meeting – Virtual Poster Sessions*, Pittsburgh, Pennsylvania, 2020.

[238] D. Bacon, "Fast Forward Model Development Using Image-to-Image Translation," in *2020 SMART Annual Review Meeting – Virtual Poster Sessions*, Pittsburgh, Pennsylvania, 2020.

[239] X. Hang Cao., I. Stojkovic and Z. Obradovic, "A robust data scaling algorithm to improve classification accuracies in biomedical data," *BCM Bioinformatics*, vol. 17, no. 1, p. 359, 2016.

[240] J. Liu, "Potential for Evaluation of Interwell Connectivity under the Effect of Intraformational Bed in Reservoirs Utilizing Machine Learning Methods," *Geofluids,* vol. 2020, no. Article ID 1651549, pp. 1-10, 2020.

[241] R. Aggarwal. and P. Ranganathan., "Common pitfalls in statistical analysis: The use of correlation techniques.," *Perspect Clin Res,* vol. 7, no. 4, pp. 187-190, 2016.

[242] J. Brownlee, "Recursive Feature Elimination (RFE) for Feature Selection in Python," Machine Learning Mastery, 25 May 2020. [Online]. Available: https://machinelearningmastery.com/rfe-feature-selection-in-python/. [Accessed 9 October 2020].

[243] B. Darst, K. Malecki and C. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Genet,* vol. 19, no. 65, 2018.

[244] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning,* vol. 46, no. 1, pp. 389-422, 2002.

[245] M. Kuhn and K. Johnson, Feature Engineering and Selection: A Practical Approach for Predictive Models, Boca Raton, Florida: CRC Press, Taylor & Francis Group, 2020.

[246] scikit learn, "sklearn.feature_selection_RFE," Undated. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html. [Accessed 9 October 2020].

[247] V. Svetnik, A. Liaw, C. Tong, C. Culberson, R. Sheridan and B. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," *Journal of Chemical Information and Computer Sciences,* vol. 43, no. 6, pp. 1947-1958, 2003.

[248] J. Hur, S. Ihm and Y. Park, "A Variable Impacts Measurement in Random Forest for Mobile Cloud Computing," *Wireless Communications and Mobile Computing,* vol. Article ID 6817627, pp. 1-13, 2017.

[249] P. Refaeilzadeh, L. Tang. and H. Liu, "Cross-Validation," in *In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems*, Boston, Massachusetts, 2009.

[250] F. Chollet and a. others, "Keras," 2015. [Online]. Available: https://keras.io.

[251] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical*

Statistics and Probability, vol. Volume 1: Statistics, no. University of California Press, Berkeley, California, pp. 281-297, 2967.

[252] R. de Amorim and C. Henning, "Recovering the number of clusters in data sets with noise features using feature rescaling," *Information Sciences,* vol. 324, pp. 126-145, 2015.

[253] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," *International Journal of Computer Applications,* vol. 105, no. 9, pp. 17-24, 2014.

[254] J. Hartigan, Clustering Algorithms, New York, New York: J. Wiley & Sons, 1975.

[255] G. Dematos, M. Boyd, B. Kermanshahi, N. Kohzadi and I. Kaastra, "Feedforward versus recurrent neural networks for forecasting monthly japanese yen exchange rates," *Financial Engineering and the Japanese Markets,* vol. 3, pp. 59-75, 1996.

[256] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 6, no. 2, pp. 107-116, 1998.

[257] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[258] K. Greff, R. Srivastava, J. Koutnik, B. Steunebrink and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 28, no. 10, pp. 2222-2232, 2017.

[259] H. Kwak and P. Hui, "Deep Health: Deep Learning for Heath Informatics reviews, challenges, and opportunities on medical imaging, electronic health records, genomics, sensing, and online communication health," 2019.

[260] C. Olah, "Understanding LSTM Networks," colah's blog, 27 August 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 6 December 2020].

[261] S. Poornima and M. Pushpalatha, "Prediction of Rainfall Using Intensified LSTM Based Recurrent Neural Network with Weighted Linear Units," *Atmosphere,* vol. 10, no. 668, pp. 1-18, 2019.

[262] F. Gers, J. Schmidhuber and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation,* vol. 12, pp. 2451-2471, 1999.

[263] P. Utgoff and D. Stracuzzi, "Many-Layered Learning," *Neural computation,* vol. 14, no. 10, pp. 2497-2529, 2002.

[264] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference for Learning Representations*, San Diego, California, 2014.

[265] Y. Ji, J. Hao, N. Reyhani and A. Lendasse, "Direct and Recursive Prediction of Time Series Using Mutual Information Selection," *IWANN,* vol. LNCS 3512, pp. 1010 - 1017, 2005.

[266] J. Carney and P. Cunningham, "The Epoch Interpretation of Learning," *IEEE Transaction on Neural Networks,* vol. 8, pp. 111-116, 1998.

[267] P. Manda and D. Bacon Nkazi, "The Evaluation and Sensitivity of Decline Curve Modeling," *Energies,* vol. 13, no. 2765, pp. 1-16, 2020.

[268] M. Paryani, M. Ahmadi, O. Awoleke and L. Hanks, "Decline Curve Analysis: A Comparative Study of Proposed Models Using Improved Residual Functions," *Journal of Petroleum & Environmental Biotechnology,* vol. 9, pp. 1-8, 2018.

[269] V. Okouma and D. Symmons, "Practical Considerations for Decline Curve Analysis in Unconventional Reservoirs - Application of Recently Developed Time-Rate Relations," in *Society of Petroleum Engineers Hydrocarbon, Economics, and Evaluation Symposium*, Calgary, Alberta, Canada, 2012.

[270] D. Montgomery, Design and Analysis of Experiments: Ninth Edition, Hoboken, New Jersey: John Wiley & Sons, Inc., 2017.

[271] R. Armstrong, F. Eperjesi and B. Gilmartin, "The application of analysis of variance (ANOVA) to different experimental designs in optometry," *Ophathalmic & Physiological Optics,* vol. 22, pp. 248-256, 2002.

[272] S. Sawyer, "Analysis of Variance: The Fundamental Concepts," *Journal of Manual & Manipulative Therapy,* vol. 17, no. 2, pp. 27E - 38E, 2009.

[273] J. Tukey, The Collected Works of John W. Tukey VIII. Multiple Compairsons: 1948 - 1983, New York, New York: Chapman and Hall, 1983.

[274] R. Brown, "Exponential Smoothing for Predicting Demand," Arthur D. Little Inc., Cambridge, Massachusetts, 1956.

[275] S. Ben Taieb and G. Bontempi, "Recursive Multi-step Time Series Forecasting by Perturbing Data," in *11th IEEE International Conference on Data Mining*, Vancouver, British Columbia, Canada, 2011.

[276] I. Fox, L. Ang, M. Jaiswal, R. Pop-Busui and J. Wiens, "Deep Multi-Output Forecasting: Learning to Accurately Predict Blood Glucose Trajectories," *24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* pp. 1387-1395, 2018.

[277] B. Scanlon, R. Reedy, F. Male and M. Walsh, "Water Issues Related to Transitioning from Conventional to Unconventional Oil Production in the Permian Basin," *Environmental Science and Technology,* vol. 51, no. 18, pp. 10903-10912, 2017.

[278] Laurentian Research, "Understanding GOR In Unconventional Play: Permian And Beyond," Seeking Alpha, 9 August 2017. [Online]. Available: https://seekingalpha.com/article/4096835-understanding-gor-in-unconventional-play-permian-and-beyond. [Accessed 26 December 2020].

[279] R. Flumerfelt, "The Wolfcamp Shale: Technical Learnings to Date and Challenges Going Forward," in *10th Annual Ryder Scott Reserves Conference*, Houston, Texas, 2014.

[280] E. Kim, "Permian basin GOR: recalibrating US gas production models," Wood Mackenzie, 9 July 2019. [Online]. Available: https://www.woodmac.com/news/editorial/permian-basin-gor-recalibrating-us-gas-production-models/. [Accessed 26 December 2020].

[281] Shale Newsletter, "Is the Permian getting gassier? Not necessarily in 2020," Rystad Energy, February 2020. [Online]. Available: https://www.rystadenergy.com/newsevents/news/newsletters/UsArchive/shale-newsletter-feb-2020/. [Accessed 26 December 2020].

[282] J. Lee, "Death by Bubble Point: Fact or Fantasy?," in *2018 Ryder Scott Reserves Conference,* Calgary, Alberta, Canada, 2018.

[283] A. Jai Persaud and U. Kumar, "An eclectic approach in energy forecasting: a case of Natural Resources Canada's (NRCan's) oil and gas outlook," *Energy Policy,* vol. 29, pp. 303-313, 2001.

[284] J. Browning, S. Ikonnikova, F. Male, G. Gulen, K. Smye, S. Horvath, C. Grote, T. Patzek, E. Potter and S. Tinker, "Study forecasts gardual Haynesville production recovery before final decline," *Oil & Gas Journal,* pp. 1-7, 2015.

[285] J. You, W. Ampomah, Q. Sun, E. Kutsienyo, R. Balch, Z. Dai, M. Cather and X. Zhang, "Machine learning based co-optimization of carbon dioxide

sequestration and oil recovery in CO2-EOR project," *Journal of Cleaner Production,* vol. 260, no. 1, p. 120866, 2020.

[286] S. Haghighat, S. Mohaghegh, V. Gholami, A. Shakarami and D. Moreno, "Using Big Data and Smart Field Technology for Detecting Leakage in a CO2 Storage Project," in *Society of Petroleum Engineers Annual Technical Conference and Exhibition*, New Orleans, Louisiana, 2013.

[287] M. Chen, O. Adballa, A. Izady, M. Nikoo and A. Al-Maktoumi, "Development and surrogate-based calibration of a CO2 reservoir model," *Journal of Hydrology,* vol. 586, p. 124798, 2020.

[288] B. Chen, D. Harp, Y. Lin, E. Keating and R. Pawar, "Geologic CO2 sequestration monitoring design: A machine learning and uncertainty quantification based approach," *Applied Energy,* vol. 225, no. 1, pp. 332-345, 2018.

[289] M. Das and K. Rangarajan, "Performance Monitoring and Failure Prediction of Industrial Equipments using Artificial Intelligence and Machine Learning Methods: A Survey," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2020.

[290] A. Al-AbdulJabbar, S. Elkatatny, M. Mahmoud and A. Abdulraheem, "Predicting Formation Tops While Drilling Using Artificial Intelligence," in *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*, Dammam, Saudi Arabia, 2018.

[291] C. Wang and T. Carr, "Marcellus Shale Lithofacies Prediction by Multiclass Neural Network Classification in the Appalachian Basin," *Mathematical Geosciences,* vol. 44, pp. 975-1004, 2012.

[292] R. Slatt, Stratigraphic reservoir characterization. Handbook of petroleum exploration and production, Elsevier, 2006.

[293] M. Hubbert, "Entrapment of Petroleum Under Hydrodynamic Conditions," *American Association of Petroleum Geologists Bulletin,* vol. 37, no. 8, pp. 1954-2026, 1953.

[294] V. Nwaezeapu, A. Okoro, E. Akpunonu, N. Ajaegwu, K. Ezenwaka and C. Ahaneku, "Sequence stratigraphic approach to hydrocarbon exploration: a case study of Chiadu field at eastern onshore Niger Delta Basin, Nigeria," *Journal of Petroleum Exploration and Production Technology,* vol. 8, pp. 399-415, 2018.

[295] National Energy Technology Laboratory, "Best Practices: Site Screening, Site Selection, and Characterization for Geologic Storage Projects," U.S. Department of Energy, Pittsburgh, Pennsylvania, 2017.

[296] A. Graciano, A. Rueda and F. Feito, "Real-time visualization of 3D terrains and subsurface geological structures," *Advances in Engineering Software,* vol. 115, pp. 314-326, 2018.

[297] J. Saravanavel, S. Ramasamy, K. Palanivel and C. Kumanan, "GIS based 3D visualization of subsurface geology and mapping of probable hydrocarbon locales, part of Cauvery Basin, India," *Journal of Earth System Science,* vol. 129, no. 36, 2020.

[298] M. Natali, E. Lidal, J. Parulek, I. Viola and D. Patel, "Modeing Terrains and Subsurface Geology," *Eurographics,* pp. 155-173, 2013.

[299] A. Turner, "Chapter 8: The role of three-dimensional geographic information systems in subsurface characterization for hydrogeological applications," in *Three dimensional applications in Geographical Information Systems*, Bristol, Pennsylvania, Taylor & Francis, Ltd., 1993.

[300] G. Bromhal, "Science-Informed Machine Learning for FE Subsurface Applications," in *Subsurface Data and Machine Learning Meeting; National Academies Committee on Earth Resources*, Washington, D.C., 2019.

[301] National Energy Technology Laboratory, "SMART Initiative," U.S. Department of Energy, 2020. [Online]. Available: https://edx.netl.doe.gov/smart/. [Accessed 28 December 2020].

[302] Enverus, "DrillingInfo," 2020. [Online]. Available: https://www.enverus.com/. [Accessed 15 September 2020].

[303] Shale Experts, "Permian Basin," Undated. [Online]. Available: https://www.shaleexperts.com/plays/permian-basin/Overview. [Accessed 4 September 2020].

[304] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal Of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.

[305] G. Lemaitre, F. Nogueira and K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research,* vol. 18, no. 17, pp. 1-5, 2017.

[306] R. P., T. L. and L. H., "Cross-Validation. In: Liu L., Özsu M. (eds) Encyclopedia of Database Systems," Springer, Boston, Massachusetts, 2009.

[307] scikit learn, "3.3. Metrics and scoring: quantifying the quality of predictions," Undated. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score. [Accessed 4 September 2020].

[308] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, Springer: New York, New York, 2013.

[309] O. Dolling and E. Varas, "Artificial nerual networks for streamflow prediction," *Journal of Hydraulic Research,* vol. 40, no. 5, pp. 547-554, 2002.

[310] Shortridge, J., Guikema, S., and Zaitchik, B., "Machine learning methods for empirical streamflow simulation: a comparison of model acccuracy, interpretability, and uncertainty in seasonal watersheds," *Hydrology and Earth System Sciences,* vol. 20, pp. 2611-228, 2016.

[311] H. Tongal and M. Booij, "Simulation and forecasting of streamflows using machine learning models coupled with base flow separation," *Journal of Hydrology,* vol. 564, pp. 266-282, 2018.

[312] A. Lohani, N. Goel and K. Bhatta, "Comparative study of neural network, fuzzy logic and linear transfer function techniques in daily rainfall-runoff modeling under different input domains," *Hydrologica Processes,* vol. 25, no. 2, pp. 175-193, 2011.

[313] Rosenblatt, F., Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Washington, D.C.: Spartan Books, 1961.

[314] J. McCaffery, "Neural Network L2 Regularization Using Python," Visual Studio Magazine, 5 October 2017. [Online]. Available: https://visualstudiomagazine.com/articles/2017/09/01/neural-network-l2.aspx. [Accessed 5 December 2020].

[315] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[316] scikit learn, "1.4 Support Vector Machines," 2018. [Online]. Available: https://scikit-learn.org/stable/modules/svm.html#regression. [Accessed 9 December 2018].

[317] Vapnik, V., The Nature of Statistical Learning Theory, New York, New York: Springer, 1995.

[318] Vapnik, V., Statistical Learning Theory, New York, New York, USA: Wiley, 1998.

[319] K. Rasouli, W. Hsieh and A. Cannon, "Daily streamflow forecasting by machine learning methods with weather and climate inputs," *Journal of Hydrology,* Vols. 414-415, pp. 284-293, 2012.

[320] scikit learn, "sklearn.svm.SVC," Undated. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html. [Accessed 5 September 2020].

[321] R. Baumgardner, H. Hamlin and H. Rowe, "High-Resolution Core Studies of Wolfcamp/Leonard Basinal Facies, Southern Midland Basin, Texas," in *American Association of Petroleum Geologists Search and Discovery Article #10607,* Adapted from poster presentation given at AAPG 2014 Southwest Section Annual Convention, Midland, Texas, 2014.

[322] W. Drake, A. Bazzell, J. Curtis and J. Zumberge, "Variability in Oil Generation and Migration with Thermal Maturity: Wolfcamp and Spraberry Formations, Northern Midland Basin, Texas," in *Unconventional Resources Technology Conference (URTeC),* Denver, Colorado, 2019.

[323] Y. Xie, C. Zhu, W. Zhou, Z. Li, X. Liu and M. Tu, "Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances," *Journal of Petroleum Science and Engineering,* vol. 160, pp. 182-193, 2019.

[324] X. Ren, J. Hou, S. Song, Y. Liu, D. Chen, X. Wang and L. Dou, "Lithology identification using well logs: A method by integrating artificial neural networks and sedimentary patterns," *Journal of Petroleum Science and Engineering,* vol. 182, p. 106336, 2019.

[325] A. Shahkarami and G. Wang, "Horizontal Well Spacing and Hydraulic Fracturing Design Optimization: A Case Study on Utica-Point Pleasant Shale Play," *Journal of Sustainable Energy Engineering,* vol. 5, no. 2, pp. 148-162, 2017.

[326] H. Viswanathan and T. Carr, "Task Presentation #7: Real-Time Forecasting: MSEEL," in *2020 SMART Annual Review Meeting - Task Presentations,* Virtual Event. Online: https://edx.netl.doe.gov/smart/2020-annual-review-meeting-presentations/, 2020.

[327] T. Jacobs, "Pioneer's analytics project reveals the good and bad of machine learning," *JPT Digital Editor,* vol. 70, no. 9, 2018.

[328] A. Karpante, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova and V. Kumar, "Theory-guided data science: a new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 29, no. 10, pp. 2318-2331, 2017.

[329] R. Vasudevan, M. Ziatdinov, L. Vlcek and S. Kalinin, "Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality," *npj Computational Materials,* vol. 7, no. 16, pp. 1-6, 2021.

[330] DrillingInfo / Enverus, "DI Research Products Glossary," Enverus, Undated. [Online]. Available: http://help.drillinginfo.com/robohelp/robohelp/server/general/projects/DI%20Desktop%20Online%20Manual/DI_Analytics/Other_Resources/DI_Research_Products_Glossary.htm. [Accessed 15 November 2020].

[331] U.S. Department of Commerce, "2017 Cartographic Boundary File, Current County and Equivalent for United States, 1:500,000," 2017. [Online]. Available: https://www2.census.gov/geo/tiger/GENZ2017/shp/cb_2017_us_county_500k.zip. [Accessed 13 March 2019].