APPLICATION OF ARTIFICIAL INTELLIGENCE FOR CO2 STORAGE IN SALINE AQUIFER (SMART PROXY FOR SNAP-SHOT IN TIME)

Marwan Mohammed Alnuaimi

wvu, mma0042@mix.wvu.edu

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Part of the Artificial Intelligence and Robotics Commons, Data Science Commons, Environmental Engineering Commons, Environmental Health and Protection Commons, Environmental Monitoring Commons, Geology Commons, Oil, Gas, and Energy Commons, and the Petroleum Engineering Commons

Recommended Citation

Alnuaimi, Marwan Mohammed, "APPLICATION OF ARTIFICIAL INTELLIGENCE FOR CO2 STORAGE IN SALINE AQUIFER (SMART PROXY FOR SNAP-SHOT IN TIME)" (2022). Graduate Theses, Dissertations, and Problem Reports. 11278.

https://researchrepository.wvu.edu/etd/11278

This Dissertation is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself. This Dissertation has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.
APPLICATION OF ARTIFICIAL INTELLIGENCE FOR CO$_2$ STORAGE IN SALINE AQUIFER (SMART PROXY FOR SNAP-SHOT IN TIME)

Marwan Al Nuaimi, BSc, MSc

Dissertation submitted to the
Statler College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in
Petroleum and Natural Gas Engineering

Shahab D. Mohaghegh, Ph.D., Chair
Samuel Ameri, M.S., P.E.
Kashy Aminian, Ph.D.
Mehrdad Zamirian, Ph.D.
Hassan Amini, PhD.
Department of Petroleum and Natural Gas Engineering

Morgantown, West Virginia
2022

Keywords: Smart Proxy Modeling in Reservoir Engineering, Artificial Intelligence, Data-Driven, Neural Network, CO$_2$ sequestration

Copyright 2022 Marwan Al Nuaimi
APPLICATION OF ARTIFICIAL INTELLIGENCE FOR CO₂ STORAGE IN SALINE AQUIFER (SMART PROXY FOR SNAP-SHOT IN TIME)

Marwan Al Nuaimi

In recent years, artificial intelligence (AI) and machine learning (ML) technology have grown in popularity. Smart Proxy Models (SPM) are AI/ML based data-driven models which have proven to be quite crucial in petroleum engineering domain with abundant data, or operations in which large surface/subsurface volume of data is generated. Climate change mitigation is one application of such technology to simulate and monitor CO₂ injection into underground formations.

The goal of the SPM developed in this study is to replicate the results (in terms of pressure and saturation outputs) of the numerical reservoir simulation model (CMG) for CO₂ injection into saline aquifers. In so doing, the artificial intelligence model was used to particularly predict the pressure distribution as well as carbon dioxide plume at any time-step throughout the period of injection and post-injection. There are four injectors injecting approximately two million metric tons of CO₂ per year for a period of ten years. The project seeks to unravel what happens to CO₂ and pressure during and after the injection process, commonly referred to as injection and post-injection periods. This process was monitored for 10 years of injection and 190 years of post-injection.

There are 46 geologic realizations of the porosity and permeability distributions which along with some 300 static and dynamic data and features extracted from the model are used as the main input to the artificial neuron network for training, calibration and validation. The dataset produced is then distributed into three major parts; the training dataset, which is majorly aimed at training smart proxy model, the calibration dataset which is majorly a watchdog, and a blind validation which is used to perform the final evaluation on the model after it achieves the desired training accuracy. The results show that the developed SPM can successfully mimic the pressure and CO₂ behavior of the CMG outputs which are determining factors of the amount and safety of CO₂ sequestration. When implemented on a large scale, this technology has the potential to be very competitive with existing numerical reservoir simulators, providing an additional toolbox for petroleum engineers and CO₂ sequestration specialists to monitor the pressure and CO₂ plume, as well as perform uncertainty quantification and optimization.
ACKNOWLEDGEMENTS

Praise be to Allah, Lord of all creation, for all the blessings that contributed to the success of this research.

To begin with, I would like to sincerely from the bottom of my heart to thank my advisor Prof. Shahab Mohaghegh. For not only his continuous guidance and leadership but also his friendship throughout my stay at West Virginia State University, indeed it is a great honor. His unmatched mentorship skills have been indeed paramount in equipping me with the experiences and insights important to my long-term career goals. He continuously encouraged me to purpose to be an instructor and an independent thinker in my bid to become a successful engineer in life.

Many thanks to the committee members for the continuous corrections and comments that saw me come up with successful ideas throughout my career journey indeed it was a pleasure working with you. The committee members comprise of Prof. Shahab Mohaghegh, Prof. Samuel Ameri Prof. Kashy Aminian, Dr. Mehrdad Zamirian, and Dr. Hassan Amiri. Their valuable remarks were true turning points in my education journey at West Virginia State University Department of Petroleum and Natural Gas Engineering. I don’t take their sacrifices for granted.

In a special way, I would like to thank the Computer Modeling Group, Inc, and Intelligent Solutions for being by my side to provide crucial items for this report such as CMG, Petrel, and Data-Driven Modeling Software from Intelligent Solutions Inc. Moreover, they provided technical and useful comments towards my study that enabled it to be a success.

I cannot miss out to thank my adorable friends and teammates that provided a conducive environment for me to undertake my studies without disturbances. It is in the small things in life that people get to appreciate the presence of individuals in their lives. I also appreciate the continuous encouragement that you gave me throughout my academic period in West Virginia State University. It is a pleasure and an honor to have such adorable personalities around me.

Finally, I would like to give a special shout of its kind to my family members especially my dad and mom for their deep love, sacrifices, and encouragement throughout my journey of education. Despite the fact that they are far away, they never let the distance interfere with the unconditional love and belief they bestow on me. From deep down my heart, I say ‘thank you so much’.
Table of Contents

CHAPTER 1: INTRODUCTION .................................................................................................................. 1
   1.1 Problem Statement ....................................................................................................................... 1
   1.2 Research Aim and Objectives .................................................................................................... 2
   1.3 Practical and Theoretical Significance/Implication of The Study .............................................. 3
   1.4 Dissertation Outline .................................................................................................................. 4

CHAPTER 2: WHY PROXY MODELS? ................................................................................................. 6
   2.1 Review of Proxy Models .............................................................................................................. 7
      2.1.1. Reduced Physics Models: ................................................................................................... 7
      2.1.2 Response Surface Methodology ......................................................................................... 8
      2.1.3. Data-Driven Modeling ........................................................................................................ 11
      2.1.4. Reduced Order Models ..................................................................................................... 16
   2.2 Proxy Modeling in Literature ....................................................................................................... 18

CHAPTER 3: RESERVOIR SIMULATION MODELING (CMG) ................................................................. 25
   3.1 Overview ................................................................................................................................... 25
   3.2 Reservoir Geological Structures Description: ............................................................................. 25
   3.3 Geological Reservoir Realizations Development ........................................................................ 27
      3.3.1 Porosity: .............................................................................................................................. 27
      3.3.2 Permeability: ........................................................................................................................ 28
      3.3.3 Relative Permeability Curves: ............................................................................................ 30
   3.4 Reservoir Simulation ................................................................................................................... 35
      3.4.1 CO2 Injection Design (Number of Injector, Injection Time) .............................................. 36
   3.5 Summary: .................................................................................................................................. 38

CHAPTER 4: METHODOLOGY (SMART PROXY MODELING) .............................................................. 39
   4.1 Review: ..................................................................................................................................... 39
   4.2 Design of Artificial Neural Network and Smart Proxy Model ...................................................... 39
      4.2.1 Network structure: .............................................................................................................. 39
      4.2.2 Input/output parameters: .................................................................................................. 42
      4.1.3.1 Data Preparation: ......................................................................................................... 43
      4.1.3.2 Data Processing: ............................................................................................................ 44
      4.1.3.3 Features Generation and Engineering: ......................................................................... 44

CHAPTER 5: RESULTS AND DISCUSSIONS ......................................................................................... 56
   5.1 Review: ..................................................................................................................................... 56
Figure 3. Porosity Distribution for the 46 realizations .................................................................28
Figure 4. Permeability Distribution for the 46 realizations .............................................................29
Figure 5. Zoomed in Permeability Distribution for the 46 realizations ........................................30
Figure 6. Relative permeability curves for rock type 1 .................................................................31
Figure 7. Oil & Gas Relative permeability curves for rock type 1 ..................................................31
Figure 8. Gas & Water Relative permeability curves for rock type 1 ...............................................32
Figure 9. Oil & Water Relative permeability curves for rock type 2 ...............................................32
Figure 10. Oil & Gas Relative permeability curves for rock type 2 ..................................................33
Figure 11. Water & Gas Relative permeability curves for rock type 2 ............................................33
Figure 12. Oil & Water Relative permeability curves for rock type 3 ...............................................34
Figure 13. Oil & Gas Relative permeability curves for rock type 3 ..................................................34
Figure 14. Water & Gas Relative permeability curves for rock type 3 ............................................35
Figure 15. Pressure Distribution for the 46 realizations ..............................................................37
Figure 16. CO₂ Saturation Distribution for the 46 realizations .......................................................37
Figure 17. Zoomed in CO₂ Saturation Distribution for the 46 realizations ....................................38
Figure 18. A simple schematics of a single-layer fully connected neural network structure with one output ..................................................................................................................41
Figure 19. Schematics showing the generation of new features .......................................................45
Figure 20. Schematics showing the implementation of tier cells .....................................................46
Figure 21. Schematics showing the implementation of tier cells and inactive cells .........................47
Figure 22. Relative permeability as input to the model .................................................................48
Figure 23. Violin plots of permeability distribution ........................................................................49
Figure 24. Zoomed in Violin plots of permeability distribution .....................................................49
Figure 25. Raw permeability data (non-transformed) ....................................................................50
Figure 26. Power-transformed permeability data ..........................................................................50
Figure 27. The general workflow utilized for SPM in this study .....................................................55
Figure 28. A list of selected features as input to the neural network a particular SPM mode .............58
Figure 29. Training pressure model P25-3 results for timestep 01-01-2021 .....................................60
Figure 30. Training pressure model P25-3 results for timestep 01-01-2021 .....................................61
Figure 31. Training pressure model P25-3 results for timestep 01-01-2021 .....................................62
Figure 32. Training pressure model P25-3 results for timestep 01-01-2021 .....................................63
Figure 33. Training pressure model P25-3 results for timestep 01-01-2021 .....................................64
Figure 34. Training pressure model P25-3 results for timestep 01-01-2021 .....................................65
Figure 35. Training pressure model P25-3 results for timestep 01-01-202 .......................................66
Figure 36. Training pressure model P25-3 results for timestep 01-01-202 .......................................68
Figure 37. Training pressure model P25-3 results for timestep 01-01-202 .......................................69
Figure 38. Training pressure model P25-3 results for timestep 01-01-202 .......................................70
Figure 39. Training pressure model P25-3 results for timestep 01-01-202 .......................................71
Figure 40. Training pressure model P25-3 results for timestep 01-01-202 .......................................72
Figure 41. Training pressure model P25-3 results for timestep 01-01-202 .......................................73
Figure 42. Training pressure model P25-3 results for timestep 01-01-202 .......................................74
Figure 43. Blind pressure model P25-5 results for timestep 01-01-202 ..........................................76
Figure 44. Blind pressure model P25-5 results for timestep 01-01-202 ..........................................77
Figure 45. Blind pressure model P25-5 results for timestep 01-01-202 ..........................................78
Figure 46. Blind pressure model P25-5 results for timestep 01-01-202 ..........................................79
Figure 47. Blind pressure model P25-5 results for timestep 01-01-202 ..........................................80
Figure 89. Scenario 1: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 131
Figure 90. Scenario 1: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 132
Figure 91. Scenario 1: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 133
Figure 92. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 136
Figure 93. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 137
Figure 94. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 138
Figure 95. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 139
Figure 96. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 140
Figure 97. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 141
Figure 98. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030 .......................................................... 142
Figure 99. Scenario 3: Layer 5: Train CO$_2$ model P25-3 results .......................................................... 145
Figure 100. Scenario 3: Layer 10: Train CO$_2$ model P25-3 results .......................................................... 146
Figure 101. Scenario 3: Layer 15: Train CO$_2$ model P25-3 results .......................................................... 147
Figure 102. Scenario 3: Layer 22: Train CO$_2$ model P25-3 results .......................................................... 148
Figure 103. Scenario 3: Layer 30: Train CO$_2$ model P25-3 results .......................................................... 149
Figure 104. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030 ............... 152
Figure 105. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030 ............... 153
Figure 106. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030 ............... 154
Figure 107. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030 ............... 155
Figure 108. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030 ............... 156
Figure 109. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030 ............... 157
Figure 110. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030 ............... 158
Figure 111. Scenario 1: Blind CO$_2$ model P50-6 results for timestep end of injection 01-01-2030 ............... 160
Figure 112. Scenario 1: Blind CO$_2$ model P50-6 results for timestep end of injection 01-01-2030 ............... 161
Figure 113. Scenario 1: Blind CO$_2$ model P50-6 results for timestep end of injection 01-01-2030 ............... 162
LIST OF TABLES

Table 1. Typical hyperparameters used in this study for one of the SPMs ......................... 42
Table 2. A list of the presented training and blind realizations in this Chapter...................... 57
CHAPTER 1: INTRODUCTION

This section of the study is aimed at giving the problem statement of the research, the hypotheses of the study, objectives, study deliverables, significance of the study, and the outline of the dissertation.

1.1 Problem Statement

Ideally, injection of CO$_2$ into the subsurface has risen to be one of the most viable technologies for decreasing as much as possible the anthropogenic CO$_2$ emission into the atmosphere. In order to realize this particular target, deep saline formation is arguably the best answer. Deep saline storage, according to a data released from North America, has of a storage capacity of approximately 3400 billion tons of CO$_2$ or the equivalent of emissions from centuries. Over the previous two decades or so, the development as well as enhancement of the various geologic sequestration technologies in order to mitigate the greenhouse emissions has been on the rise. In the United States, for example, the U.S Department of Energy’s Carbon Storage Program has in the recent past has given tremendous support for the R & D activities within the country to not only develop but also provide advanced technology that is bound to boost in a big way the efficacy of the geologic carbon storage technology, cut the cost of implementation, and be ready for commercial deployment in the near future.

When CO$_2$ is injected into the underground saline aquifer, various factors are usually involved in order to make the process smooth and a success at the same time. Two of the most crucial factors are the pressure distribution and the carbon dioxide (CO$_2$) saturation. To begin with, pressure is of concern with special regards to the cap rock integrity as well as potential migration of brine or carbon dioxide outside of the injection zone. On the other hand, carbon dioxide saturation is of importance because of the displacement efficiency. It is in the public domain that when huge amounts of carbon dioxide are sequestrated underground, huge amounts of pressure readily build up within the storage formations and may result into formation of fault lines within the cap rock which might in turn result into displacement of the carbon dioxide from the underground. To avoid such scenarios, there ought to be well organized and established technological models that seek to monitor both the pressure distribution and carbon dioxide plume at any time step.

Numerical reservoir simulation has been of great value in the recent past in the bid to predict the pressure distribution and carbon dioxide plume within the petroleum and oil industry to avoid any errors and leakages associated with cap rock faults. It is, however, quite important to note that the numerical reservoir simulation models are computationally slow in complex projects involving uncertainty assessment. It takes quite a huge amount of time (up to approximately one week
duration) to undertake its predictive role on carbon dioxide plume and pressure distribution within the carbon dioxide sequestration process. It is in this view that the project seeks to come up with more technological model that will replicate the roles of the numerical reservoir simulation model within a shorter duration of time while maintaining the same accuracy previously offered by the numerical reservoir simulation model. The technological model of choice for this particular project is the artificial intelligent model also known as the smart proxy model.

Artificial neural networks have proven to be powerful data processing tools and can be applied to a wide variety of problems in different areas such as medical, science, financial, business etc. It has been about three decades that neural networks became a point of interest in petroleum engineering and geoscience. Smart Proxy Technology was first designed by Prof. Shahab Mohaghegh at the West Virginia University in 2006. Since then, there has been several studies focused on application of such technology by the Laboratory for Engineering Application of Data Science (LEADS) at the West Virginia University Department of Petroleum and Natural Gas Engineering (PNGE) (Jalali, Mohaghegh, 2009; Kalantari-Dahaghi, Mohaghegh, et al. 2011; Mohaghegh, Amini et al. 2012; Amini, Mohaghegh, et al. 2014; Shahkarami, Mohaghegh, et al, 2014; Gholami, Mohaghegh, et al, 2014; Mohaghegh, Gaskari, et al. 2014; Haghighat, Mohaghegh, 2015; Alenezi, Mohaghegh, 2017; Alaboodi, Mohaghegh, 2021). In this particular project, the smart proxy is aimed at carrying out two crucial models in the process of carbon dioxide sequestration: prediction of the pressure distribution and carbon dioxide plume after continuous injection of two million metric tons of carbon dioxide per year for a significant period of ten years. Moreover, the model seeks to predict the pressure and carbon dioxide plume after the injection of carbon dioxide injection is stopped (post-injection period) for a duration of 190 years.

It is worth noting that a single run of basic numerical simulation models like the numerical reservoir simulation model takes hours to days to undertake their role while smart proxy model runs in a matter of seconds thereby saving 98.9% of calculating time. The result of the project clearly demonstrates the advantage of the proposed workflow for solving computational time, high run-time, and computational cost of the CMG models used in this study. Additionally, the smart proxy model predicts the pressure distribution and carbon dioxide plume with the same accuracy as that of the CMG model.

1.2 Research Aim and Objectives

The main objective of this particular research is to successfully come up with an artificial intelligence model that can effectively replicate the roles of numerical reservoir simulation model with the main aim of predicting the pressure distribution and CO₂ plume at any given point in time within the simulation period and at every single grid block in the reservoir. The pressure distribution and carbon dioxide plume prediction are to be carried out via processes of injection and post injection. Four injectors are to be used with the target of injecting approximately two million metric tons per year for a period of ten years.
1.3 Practical and Theoretical Significance/Implication of The Study

Smart proxy models are very important models in as far as accuracy and effectiveness of a particular industry or portfolio is concerned. Most of the industries that have employed the use of artificial models have ripped big in terms of production, environmental conservation, and profit levels. The petroleum and oil industry are no exception to this continuously and fast-moving world. For the industry to record top-notch levels of production, world-class mitigation strategies, and compete favorably within the market, it is just prudent for them to employ this technology in their day-to-day activities. This particular study has a very crucial aim that seeks to develop a smart proxy that will be very fundamental in determining or predicting the pressure distribution and the CO$_2$ plume. Previously, the CMG model has been in use and has contributed to the tremendous strides made by the petroleum and oil industry. It goes without saying, therefore, that the model has been used and tested in various grounds and proven to be partially effective. The effectiveness is in doubt as it is time consuming in nature and cannot guarantee proper and enhanced storage mechanisms for the carbon dioxide produced in the industry. It is on this view, that a model known as the artificial intelligent model, or the smart proxy found its way into the discussion. Instead of the model taking approximately one week discharging its role, the new smart proxy will take minutes or even seconds to do the same work while maintaining the required and similar accuracy as that of the CMG. This is a plus to the petroleum and oil industry as little time will be used in prediction process and the reservoir manager will be able to generate much more scenario predictions to maximize the effectiveness of reservoir management decision-making towards reservoir optimization.

This particular study is aimed at developing the intelligence smart proxy models and making some tests to it before releasing it to the fields to begin its intended duties. In order to attain those particular aims and objectives, there are crucial steps that ought to be followed in order to ensure success of the process. First, various realizations will be obtained from the saline CO$_2$ sequestration models. The realizations will be generated with special consideration to two major factors which comprise of the porosity and the permeability. The unrealistically low and unrealistically high realizations will be left out in a bid to get adequate and comprehensive findings within the study. For this particular study, a total of 61 realizations will be selected with the only changing aspects within the realizations being porosity and permeability as earlier mentioned. Of the 61 realizations, only 46 of them will be used in the process. CO$_2$ will then be effectively injected into the system reservoir through the help of four injectors for a period of ten years in order to help determine the pressure distribution changes and the CO$_2$ plume on the saline aquifer. A post-injection observation will effectively be carried to determine what happens to the pressure and the CO$_2$ after stopping the injection. This will be carried out for a period of 190 years. Three types of datasets will be employed in this study: the training dataset, the calibration dataset, and the blind validation dataset. There are thirty layers within the model with three rock types (type 1, 2, and 3). Of the thirty layers, the first two will act as seals as they are impervious and only contains rock type 1. The injectors will be fitted from layers 3 to 30 which contain mixtures of rock type 2 and 3. Porosity and
permeability are the two factors that define the various rock types. Simulation runs will be carried out to predict the pressure distribution and carbon dioxide plume.

The results of the study are bound to be positive with the smart proxy level being able to effectively carry out an accurate prediction for the two variables. The results of the prediction will be compared against the actual results and the errors calculated. A very low error value will indicate that the smart proxy is viable and accurate enough to be used in the field to determine the pressure distribution and CO₂ plume. This will be a major boost to the industry as the predictive rate will shoot thereby enabling the reservoir manager to generate more scenario prediction. Consequently, time factor will adequately be addressed as the smart proxy will be able to do the work in minutes or seconds as opposed to the weeks that was previously done by the CMG model. The CMG model is good but requires a model that may replicate its functions in a smart and candid manner and without derogating its previous roles but enhancing its time efficiency.

1.4 Dissertation Outline

The general proxy techniques that are adequately pertinent to this study are extensively discussed and outlined within the second chapter (Chapter 2). This is the chapter that is concerned with the introduction of the concept of smart proxy as well as various techniques that are used to develop well-functioning proxy model. The introduction and extensive discussion on the smart proxy models is then followed by the literature reviews on smart proxy or the artificial intelligence models within the petroleum and oil industry. Furthermore, this chapter also highlights the various applications of the various proxy models that are found within the petroleum and natural gas industry in a broad manner. The literature reviews in this chapter are important in providing the background to the smart proxy models that have been tried and tested over the years within the petroleum and oil industry. To enable one completely to comprehend the background information, the various advantages, the limitations, the gaps, and the links about each model are extensively reviewed within the chapter.

In chapter three is about methodology. Methodology is an area that major focuses on the procedures and steps that are followed to make the study a success. Within this particular study, there are various methodological steps that are undertaken in order to ensure success within the study. This methodology entails selection of a given number of realizations before data is extracted from them and some crucial features added in order to achieve proper prediction of the results by the smart proxy developed. The methods used in order to arrive at the expected results for the smart proxy are training, calibration, and blind validation.

Chapter 4 is the result section which entails proper running of tests and coming up with appropriate results that ought to be analyzed in correlation with the smart proxy being developed. Various datasets will be used within the study to come up with adequate and comprehensible results that can be analyzed as well as compared with other initial studies to determine whether the model is viable or not. This section contains several charts as well as figures that aim to describe the various
outcomes of predictions that are made by the new model under development. The results are then compared effectively with those from the previous studies as well as from the actual results in order to determine the accuracy levels of the smart proxy in terms of predicting the distribution pressure as well as CO₂ plume.

Chapter 5 is that which entails summary, conclusion, and future research work. This is the section that is mainly concerned with providing properly outlined summary of the whole research study. Moreover, the section seeks to outline the important conclusions that can be drawn from the research work. Most importantly, it provides the readers and researchers with the recommendations that ought to be put in place when conducting future research works of a similar nature so as to improve the future research works regarding the crucial topic of study. This is the section that is majorly concerned with improving and brightening the research works in the coming future.
CHAPTER 2: WHY PROXY MODELS?

A proxy model is a type of model that is majorly a subclass of a database-table in nature. Normally, creation of a subclass of a model result in a wholly new database table that contains a reference dating back to the original model’s table. In order to properly understand the proxy model, one ought to appreciate the fact that proxy models usually don’t create their own database tables but they usually in most cases operate on the original table that was existent. The existent methodology for predicting the pressure distribution and CO$_2$ plume within the petroleum and oil industry is the reservoir numerical simulator model. However, its inefficiency in terms of time has called for a smart proxy model in order to ensure that the rate of prediction of both the pressure distribution and carbon dioxide plume during the process of carbon dioxide sequestration is escalated. In both computational science and engineering, the proxy models offer computationally cheap alternative to the other high-fidelity models. Despite the advances made in terms of computer and modeling tools, the simulation aspects as well as prediction of various variables (in this case CO$_2$ and pressure) are still a pain in the neck and remain to be a challenge up-to date. It is in this view that the smart proxy model to predict the variables have been anchored on.

In most of the sub-surface kinds of modeling and their earth science counterparts, proxy model can be used as a facilitator to the numerical physics-based kinds of approaches such as the CMG model. The reservoir simulation models comprise of the standard modeling tools that are normally used by the petroleum and natural gas industry in order to comprehend the fluid flow behavior in the porous media. Moreover, they are utilized in the petroleum reservoir management for various aspects such as forecasting, optimization of injection scenarios, comprehension of the communication between reservoirs, and proposing infill wells. The prediction of the pressure distribution and carbon dioxide plume squarely lies within the forecasting function of the artificial intelligent models.

Therefore, from the objective of the research is it quite easy and open to tell the importance of the model to the study. It will assist in predicting the pressure distribution after injection of the carbon dioxide into the reservoir for storage purposes as well as predict the carbon dioxide plume within the reservoir.
2.1 Review of Proxy Models

From the basis of a reservoir simulation, a proxy model is a kind of a mathematical model with a unique capability of approximating the output of a simulator using a set parameter or rather a given physical model in place. For example, in this particular project, the proxy model is aimed at predicting the pressure distribution and carbon dioxide plume after injection of the carbon dioxide into the saline aquifer. A proper and useful proxy model ought to be very accurate for the problem at hand while at the same time ensuring less computational effort in comparison to the full order kind of simulation. In most cases, a proxy model can be created through various methods which comprise of data regression, reduced form of order modeling, machine learning, and lower order discretization methods.

For the proxy modeling extensively discussed within this particular section, an advanced approximation or simplification is taking place. Usually, the proxy modeling may be divided into various categories using various criteria. The various techniques may further be subdivided into statistical or the mathematical models regarding their respective approaches and each of the techniques will contain their subcategories. For instance, the response surface is an open example of the proxy models that may be categorized under statistical models. On the other hand, the reduced order models and the up-scaling techniques are categorized under the mathematical proxy models. The contemporary model reduction techniques may be classified as either intrusive or non-intrusive methods based on whether the creation of the reduced order model needs modifications of the governing equations as well as the numerical kind of implementation of the full-fledged simulator. There are various disadvantages associated with the proxy models such as when the inputs are increased; the number of trainings ought to be sharply increased since sampling is of importance in covering the input parameter space. However, the advantages associated with the models outweigh the disadvantages hence have been explored in the recent past to ensure success of various industries especially in the storage of carbon dioxide. Some of the categories of proxy models that are bound to be discussed in this area comprise; reduced physics models, reduced order models, data driven modeling, and the response surface models.

2.1.1. Reduced Physics Models:

Judging from its name, reduced physics model, is a model that aims to reduce the complexity of the model through reduction of run time. On a general scale, the reduced physics model may be categorized as being a simpler version of the full physics model. The reduced physics model makes use of less physical effects or the grid blocks. Tuning procedure may be very important in determining the parameters of the reduced physics model. To effectively use the reduced physics model, there is need for numerical experimentation that acts as a guide to the engineer in determining the amounts of physics required in simplifying the initial numerical reservoir model.
2.1.2 Response Surface Methodology
This is a type of proxy model that is usually based on numerous reservoir realizations and are
usually chosen with strict adherence to the experimental design technique. The various
experimental designs are described as the techniques used in choosing of points for sampling
variables that are usually used in the process of proxy development or designing. An appropriate
example is giving rise to a group of equations with contains the values of parameters at designated
points in an uncertainty space. Examples of this specific type of methodology comprise of space
designs and classical systems. Examples of the experimental design techniques comprise of full
factorial, factorial designs, and space-filling designs. The "curse of dimensionality" poses a
problem for commonly used ROMs (Gholami et al. 2019; He, 2013; Chen et al. 2013). Because of
the nature of the ROM models, the time required to develop it is often comparable to the time
required to execute a numerical simulation. As a result, ROMs can be computationally expensive
at times.

Several the predefined functions are quite okay in a bid to make use of the results of the
simulations. Specific examples of the predefined parameterized functions include the thin plate
splines, the least squares, and the Kriging models. The peculiar thing with this kind of model is
noticeable in the determination of the output where response surface is made use of instead of the
simulator and they employ new inputs to realize the role output determination. Ideally, the
response surface makes use of various mathematical as well as statistical techniques. The
techniques are quite important in determination of functional relationship that exists between
groups of inputs or rather controls (x₁, x₂, …xₖ₁) and their corresponding interest (y). The
relationship is generally unknown but can be replaced in terms of approximation by low-degree
polynomial model which takes the form of.

\[ y = f'(x) + \epsilon \] .................................Equation 1

In which case (x₁, x₂, …xₖ₁), f(x) is a known to be a vector function of the p elements that contain
powers as well as the cross multiplication of the powers of (x₁, x₂, …xₖ₁) up to point given by the
d (≥ 1). The \( \boldsymbol{\beta} \) in the equation represents a vector of p referred to as parameters because it doesn’t
have a known constant while on the other hand the \( \epsilon \) in the equation is a random experimental
error that is usually assumed to possess a zero mean. This is majorly attributed to the fact that
abovementioned model usually gives a proper and satisfactory representation of the response. The
quantity f(x)\( \beta \) is referred to as the expected value of the letter y and it is usually the mean response
that is represented by \( \mu(x) \). Several experiments ought to be performed and the total number of
experiments carried out is denoted by the letter n. Within the various experiments carried out, a
specified number of the inputs ought to be used to realize the response y. The settings may as well
be represented by a matrix which is indicated by a letter D which is of order n x k which is a design matrix.

In this aforementioned matrix, the \( x_{ui} \) is a representation of the \( u^{th} \) design setting of the of \( x_i \) (\( i = 1, 2, \ldots, k; \ u = 1, 2, \ldots, n \)). Each and every row that contains D denotes the design point with a k-dimensional Euclidean space. A good and properly organized design is very crucial for any particular response surface and the quality and accuracy of the prediction in a study squarely depends on it.

The below equation can perfectly match the criterion of being described as response surface. In the equation, the response of interest is denoted by the letter \( y \) while on the other hand the letter \( x \) represents the uncertain parameters which are different in each simulation run. The letters \( i \) and \( j \) represent the run indices of the \( k \) simulation. Various methods such as the least square method of optimization are put into calculation of the weights in the response surface that are created by the use of simulation \( k \)

\[
y = \beta_0 + \sum \beta_i x_i \ k \ i=1 + \sum \beta_i x_i \ k \ 2 \ i=1 + \sum \sum j>i \ \beta_{ij} x_i x_j \ k \ i=1 \ldots \ldots \ldots \text{Equation 2}
\]

The response surface methodology has various methods within it as mentioned earlier in the beginning description part of the model. This part aims to outline brief description of the various methods within the response surface methodology.

Least Squares: With least squares, the main intention is usually based on the construction of a representative function that comprises of simple known functions. The constructions with simple known functions include polynomials which undertake a very unique role of minimizing the sum of the squared residuals that are present between the simulated values and the function values. In most cases, the least square methods are not data exact in nature. In cases where the prior data is of the same value as that of the unknown functional coefficients, then all the prior data points maybe traversed by the least square interpolation. In order to effectively achieve a smooth surface, the total number of coefficients chosen ought to be smaller than the data points at all times. Arguably, the least square algorithm is used to effectively solve over-determined problems.

Kriging: This method usually postulates a mixture of global model and the departures. The equation that is normally used in the kriging method is as written below;
\( (x) = (x) + Z(x) \) \( \ldots \) \text{Equation} 3

In the aforementioned equation, the \( y(x) \) is described to be the unknown function of interest while \( f(x) \) on the other hand is a known function of \( x \) which is usually polynomial in nature. Moreover, the value \( z(x) \) is said to be the realization of a stochastic process and possess a mean of zero, non-zero covariance, and variance \( \sigma^2 \). In this particular equation, the \( f(x) \) term is equal to the polynomial model in a response surface and in most cases usually gives a ‘global’ model of the design space. This Kriging method maybe considered to be a least squares linear regression technique that can be adequately generalized to the multiple dimensions. This method rides on the assumption that various points are spatially correlated to one another. Within the method, a covariance function is specifically used in order to show the extent of correlation that exists between points or otherwise known as the spatial continuity. The covariance model is specifically important in the estimation of the spatial correlation that exists between the sampled and unsampled points. Furthermore, the covariance model is responsible for the determination of weight of each sampled point on the estimation. Usually, spatial correlation is directly proportional to the weight on the location which means that the more spatially correlated a sample is with a particular location, the more likely the weight of the sample on that particular location. Converse to what is in the least squares; the kriging method is data exact which means that it is able to reproduce the observed value at a sampled location. For example, for an experiment which contains \( N \) number of observations, Kriging method requires the inversion of \((N+1) \times (N+1)\) matrix.

Thin-plate splines: The name of this particular method is derived from its physical appearance in which there exists bending of a thin sheet of metal. Just as the normal metal presents with increased rigidity, so is the thin plate spline which in most cases resists bending of whatever nature thereby resulting in negative implications on the smooth fitted surfaces. Making use of a given number of data points, a weighted combination of thin plate splines at the center of each and every data point provides interpolation function that passes via the points while at the same time reducing as much as possible the ‘bending energy’. When provided with \( k \) corresponding points in a 2-D case, the thin plate splines warp is provided by the formula \( 2(K + 3) \) parameters- this is in addition to the six global affine motion parameters and the \( 2K \) coefficients. The aforementioned parameters can effectively be computed by solving a linear system- this is to point out that the thin plate splines possess closed-form solution.

The response surfaces within the thin plate splines have previously been used for various purposes such as optimization, uncertainty quantification, and history matching. The main reason as to why they have been widely employed for those particular roles lies in their ease of implementation at all times and which means obtaining of a model response will be definitely rapid. Conversely, for one to come up with accurate response surface then a huge number of reservoir realizations will be required. The number of reservoir realizations required is directly proportional to the number
of parameters taking part in the study. Therefore, it is worth noting that the response surface methods are normally proper and advisable alternatives only in cases where there are a limited number of parameters usually between five and twenty. It is also worth noting that the required accuracy of the response surface method wholly depends on the nature of the problem at hand. For example, the accuracy levels required for sensitivity analysis kind of problem is far much less than that required for optimization or history matching problem.

Despite the numerous advantages associated with the response surface method, there are various crucial and recognizable limitations that make it quite difficult to be alluring in most cases;

• Each and every response surface is made in such a way that they are able to model only a single response of interest with respect a numerous uncertain parameters. In cases where more than a single response is required, then new response surface models must be built
• The generally accepted assumption is that the response is smoothly varying with the change of the parameters
• In most cases, the response surfaces are put in place in order to work with the continuous parameters but in the oil and gas industry various parameters of discrete nature are dealt with. The examples of discrete parameters that are utilized in the oil and gas industry include; unconformities, permeability curves, faults, and depositional environments just but to mention a few.
• Most of the reservoir models are usually generated by use of the geostatistical methods which normally leads to stochastic noise in response surfaces.

Polynomial Chaos Expansion (PCE): The polynomial chaos expansion was introduced so as to characterize the random fields. In the recent past, the expansions have been quite fundamental in the field of physics as they have been used to solve the stochastic partial differentiation equations. The polynomial chaos expansion methods are divided into two major categories that comprise the intrusive and the non-intrusive approaches. For the intrusive approach of polynomial chaos expansion, the approximations therein are substituted within the governing equation. After that, the Glarekin and the discretization scheme are used in order to obtain accurate coefficients. On the other hand, the non-intrusive approach coefficients can readily be computed through calculations by use of small number of model simulations and without any form of alteration within the governing equations. Within the non-intrusive approach, there are two main methods that are used within it which comprise of projection and regression. Regression, however, has increased popularity as compared to the projection method due to the fact that it contains fewer computations for several system inputs (Zhang and Sahinidis 2013a, b, c).

2.1.3. Data-Driven Modeling
Great and major developments within the field of computational intelligence with strict observance at machine learning have been very instrumental in the expansion of the capabilities related to empirical modeling. Data driven modeling is the field that encompass all these kinds of new approaches. Just as the name goes, data driven kind of modeling is one which is aimed at analyzing
particular regarding a system narrowing down to finding connections that exists between the various state variables within the system. The variables that ought to be compared comprise the input, output, and internal variables. Data modeling as a whole entails a list of fields that are crucial to its functioning:

- The computational intelligence which comprise of fuzzy systems, the neural networks, and evolutionary computing. Moreover, it also entails other areas related to artificial intelligence and machine learning as a whole.
- Artificial intelligence which is a very unique study that seeks to find out the manner in which human intelligence may be incorporated within computers.
- The soft computing which is more or less similar to computational intelligence but which focuses mainly fuzzy rule system.
- Machine learning which was initially categorized under artificial intelligence and concentrates on the theoretical foundations used by both computational intelligence and soft computing.
- Data mining and knowledge discovery within the databases which are aimed at financial applications.

Data modeling as described above is usually focused on computational intelligence as well as machine learning methods that are very instrumental in the process of building models for purposes of either complementing or replacing the physical models in place. Normally, machine learning algorithm is used for purposes of determining the relationship that exists between the system’s input and its output. This is achieved through the use of a training data set that represents all the behavior found within the system. After the training of the model, it may be properly tested through the use of independent data set for purposes of determining how well the model can generalize unseen data. There are numerous data driven modeling techniques within the petroleum and oil industry field. They are discussed in broad length within the coming section.

a) Artificial Neural Networks: These are normally the biologically inspired kinds of computational models which normally attempt to replicate the manner in which the human brain undertakes its roles. There are numerous research works to this effect and most of the studies have indeed echoed the fact that most of the artificial neural networks have been crucial in the petroleum and natural gas field. In a nutshell, the artificial neural models are normally developed through training of the network with the major aim of representing the relationships as well as the processes that are inherent within the data. Artificial neural models are normally non-linear regression models hence they undertake a unique input-output mapping through the use of a simple group of interconnected processing nodes or neurons for that matter. Each of the neurons used in this particular model makes use of the inputs either externally or internally and then passes the inputs via activation or rather a transfer function such as either a logistic or sigmoid curve. The data then enters the neural network via the input units which are arranged in a particular manner referred to as input layers. The data are then progressively fed forward via successive layers which include the
hidden layer situated in the middle and further emerge from the output layer on the right hand side. Importantly, the inputs ought to be any combination of variables without restriction on specifics provided they are thought to be very crucial for prediction of the output. This is the hallmark of the whole process and very fundamental in the process of making proper and sound predictive models.

Ideally, the hidden layer is the most crucial component that gives the neural network the nod to learn the relationship that exists between the data provided. There is usually the use of back propagation algorithm that is very crucial in training of the feed forward neural network. Back propagation kind of algorithm is one of the most used neural networks hence the name multilayer perceptron. The back propagation type of algorithm is very fundamental in reduction of errors registered between the predicted and the actual output values. Additionally, the weighted connections that exist between the neurons are normally adjusted after each and every training cycle up to such a time that the error recorded in the validation data set begins to rise. It is worth noting that the validation data set is usually the second data set given to the network for purposes of evaluation in the training process. If this procedure is not followed to the letter, the network will probably represent the training data set absolutely well but will not be able to generalize to unseen data set or training data set. At the point at which the network is properly trained, it is then eligible for operation when the new input data are passed via the trained network in its non-training mode which results into the desired model outputs. The process of validation of the performance of the trained network before it is actually put into operation is managed through the use of test data set. The most vital manner of assisting in the promotion of generalization to unseen data is through making sure that the training data has at least all of the behavior in the data as highlighted previously in the text. This can actually be achieved through ensuring that all the three main data sets which comprise training, validation, and testing have more or less similar statistical properties. Of great importance is that the test set cannot in anyway be used in changing the present properties of the trained model.

The use of Artificial Neural Networks has many successful applications in the petroleum and natural gas filed. In particular, this network has been fundamental in the prediction of the carbon dioxide storage and oil recovery in the recent past. In order to properly carry out the predictions, uncertainties parameters such as geologic factors have been used in order to come up with training data base from which proper prediction through the use of the network have been accomplished. The results of the previous research works prove beyond any reasonable doubt that artificial neural network have very excellent prediction performance together with a very high correlation coefficient. The accuracy with which the artificial neural network is able to carry out its prediction role is quite high and enables by huge margin to determine the amounts of gas that can be stored to prevent environmental pollution. When artificial neural network is used, there is increased generation of scenario predictions thereby resulting in the maximization of the reservoir management decision making towards reservoir optimization.
Application Perspective of Artificial Neural Network Models

The applicable artificial neural network ought to be considered both in the technical manner as well as on economic grounds due to the fact that oil price emanate as a substantial barrier for CO2-EOR and storage project. In this particular study, the CO2 injection process was adequately optimized by the use of PSO otherwise known as the Particle Swarm Optimization and the developed artificial neural network was then used in order to achieve optimum oil production levels, CO2 storage, as well as economic aspects such as the Net Present Value. The financial metrics for the injection of the gas were then highlighted in order to provide a rough overview of the role of artificial neural network.

b) Fuzzy Rule-Based System: This particular model makes use of the fuzzy logic for its inference. This logic is based on what is referred to as fuzzy set theory whereby the binary set membership is extended for purposes of including partial membership ranging between zero and one. The fuzzy sets have got gradual transitions between the defined sets which provide the ground for the uncertainty related those particular concepts to be modeled directly. In situations where there have been definitions of each model variable together with a group of overlapping fuzzy sets, the mapping in regards to inputs to outputs may be put as a set of IF-THEN rules which may then be specified from data or even from the expert knowledge. Conversely, unlike the above discussed neural networks, fuzzy models are often susceptible to rule explosion which means that an increase in the number of fuzzy sets results to a direct increase in number of rules thereby making it quite difficult to specify the model in basing on expert’s knowledge alone.

The group of fuzzy sets as well as their accompanying rules is normally referred to as the fuzzy model knowledgebase. The inputs of the model are initially fuzzified through the knowledgebase and later a fuzzy inference engine is made use of to process the rules in parallel through a fuzzy inference procedure such as the max-min or the max-product operations. The fuzzy solution surface that is normally due to the execution of the rule base is defuzzified in order to produce what is known as the system outputs. The IF-THEN rules may at some point contain consequents usually of a liner or polynomial form contained in a formulation referred to as TSK model. The inputs are typically fuzzified with regards to the fuzzy set definitions which are combined through inference engine as well as functional consequents. The overall result is usually a weighted average of the equations due to the fact that more than a single rule may fire positively in a single pass of the rule base.

Ideally, fuzzy logic has overtime found numerous successful applications majorly in control theory. As previously discussed, it is true that fuzzy rule-based system may be created primarily through either interviewing of human experts or otherwise through processing of the past data hence the hallmark of data driven model. The basics of the latter in petroleum and oil industry field is vast and may be found in various research works in the recent past. It has been very fundamental
in carrying out various predictive operations within the petroleum and oil industry. One of the milestone projects of the fuzzy rule based system in the recent past was its role in the prediction of carbon dioxide emission from the energy sector and the global temperature increase. This model may be very instrumental in the prediction of carbon dioxide storage and pressure distribution during and after injection of carbon dioxide.

c) Genetic Algorithms in Model Optimization: These are the non-linear search as well as optimization ways that are normally attributed to the biological processes of mainly natural selection and survival for the fittest. They do not primarily belong to the data driven models but since they are widely used as optimizing models, they are broadly categorized in this section. Genetic algorithms normally lie towards the sector of computational intelligence. Genetic algorithms just like their counterpart randomized search algorithms portray implicit parallelism hence by huge margin reduce the chances of converging to a local optimum. Additionally, the genetic algorithms make use of probabilistic rules within their search processes thus can easily outdo conventional optimization techniques on very hard, discontinuous, and multimodal kinds of functions. However, their unique and adaptive search dimensions do not guarantee proper global solutions but they are said to be quite fast in producing acceptable solutions.

Normally, the basic unit used in genetic algorithm is the gene which is a biological terms used to refer to the basic characteristic of a particular individual. Specifically, the gene in genetic algorithm represents a particular parameter that is being optimized which is then required to produce the desired solution. In order to facilitate the initial process of the search in genetic algorithms, a population of the individuals is normally randomly generated. This is followed by the string being evaluated by a fitness or rather objective function in strict relation to the performance measure. This, therefore, represents the success of the solution and is very critical to the survival of any individual within the population. In order to enhance the solutions of the population, it is quite necessary to ensure selection of the fittest members of the population before being exposed to genetic operators. The least fit members of the population will then die due to natural selection as they will be replaced by new and recombined individuals of the solution.

The major genetic operator is what is referred to as the crossover in which random choice along the operator is made that is responsible for cutting two parent chromosomes into two major segments which are then swapped. Genetic algorithm is a very crucial tool that is that is aimed at handling very difficult problems that conventional techniques may offer zero or minimal help. Also, the techniques may be used in order to improve existing methodology through the process of hybridization. For instance, the fuzzy model may be adequately optimized by the genetic algorithm via an inductive approach or alternatively an expert’s knowledge may be used in order to set the rules for the same.
In the recent past, genetic algorithm has been very fundamental in the process of modeling solubility of carbon dioxide in the reservoir brine in a bid to ensure successful storage of carbon dioxide. In this case, the genetic algorithm can be said to obtain the most appropriate route after numerous loop computations. The search processes comprise of cross-over, selection operators, and artificial mutation.

2.1.4. Reduced Order Models

Ideally, the reduced order models are the projections of the full order numerical descriptions into a low dimensional sub-space. This, therefore, reduces by a huge margin the number of unknowns that ought to be computed at particular time steps. The reduced order models may primarily be classified as either system-based, grid-based, and the snapshot methods. To begin with, the grid based technique is that which the dimensions of the problem is hugely reduced through the process of changing the dimensions of the grid and further solving the problems with special regards to the coarser grids. The most common examples associated to this particular technique comprise up scaling and the multi-scaling methods. In up scaling, for instance, can be performed through the calculation of permeability, transmissibility, and porosity. The fact that up scaled kinds of models may be solved via reservoir simulator makes the method quite appealing as a method of reduced order model. Conversely, this particular method may not be very good in the provision of fine scale kinds of solutions but may provide proper solution for coarse-scale model. In terms of the compositional kinds of models, up scaling has proven to be a very satisfactory technique. In the multi-scale methods, the functions that take control of fine scale effects are duly constructed for each and every coarse grid block. The coming up with a dual-grid method which encompasses a number of multi-scale features proved that there may be application of multi-scale for Darcy flow. In these kinds of approaches, the pressure kind of equation is taken care of in the coarse scale while the transport equation is solved via the fine scale. These kinds of approaches have been submitted to speed up the processes between factor 10 and 20.

The system based kinds of approaches emanate from the system control theory and marked as the second type of reduced order models. In this group, the main methods employed are the balanced truncation methods and the krylov subspace. The most fundamental aspect that ought to be taken care of in this particular field is the fact that not all of the outputs in a model are relevant and crucial for the problem that is in place. These particular methods are normally derived through special consideration of the linear time invariant and then converting the full order to reduced order which can then be adequately solved without many problems. Krylov subspace methods prove to be very fast as they entail more of multiplication of matrices which are normally sparse. Their undoing, however, is the fact that they usually generate larger projection subspaces. On the other hand, balanced truncation methods are normally pegged on observability and controllability. In terms of definition, controllability entails sensitivity to a state while observability refers to the
influence of a state on the output or otherwise the output energy that is normally produced by the state.

The main difference that lies between the snapshot and the system-based methods is in the fact that the matrices in the snapshot methods are from the snapshots. In a layman’s language, snapshot refers to the provision of full-state information at a particular within the duration of training simulation. The main method used in snapshot system is the proper orthogonal decomposition. It is a very important technique that is used by the snapshot method in order to produce low order models by use of snapshots from a forward type of simulation. Some of the reduced order models are properly discussed below;

a) Proper Orthogonal Decomposition: This is a very crucial part of the reduced order model that is mainly aimed at reducing the high order models to the reduced order models for purposes of easier simplification. It can, therefore, be said to be aimed at reducing the dimension of the problem at hand. The simulation used in this particular method for provision of snapshots is referred to as training cases while those required for prediction are the test cases.

High-Order Reservoir Model: So as to adequately produce a reduced order model, it is prudent to initially run a full order simulation model and produce snapshots. Putting into consideration a two-dimensional two-phase reservoir model which measures m by n grid blocks and which possesses absolutely no flow boundaries at all of the sides; then the behavior of the reservoir model may be described a group properly outlined differential equations. The group of model equations may be derived from oil and water material balances, Darcy’s law, and two various closure equations of saturation and pressure.

POD Applied on Two-Phase Reservoir Simulation: During the process of reservoir simulation run, both the pressure and the saturation at each of the grid block and each time step is kept as a vector (state vectors $x_p, x_s$). The full order state vector may then be found through multiplication of the reduced vector by the transformation function.

b) Trajectory Piecewise Linearization: In order to accurately obtain higher levels of speedups, there ought to be proper treatment of the nonlinearities associated with the proper orthogonal decomposition. There are two major ways that can be used in order to effectively achieve the above requirement; trajectory piecewise linearization and the discrete empirical interpolation method. The former involves the combination of both the
POD and the interpolation method in order to solve the reduced linear terms. Snapshot matrix is very crucial for the purposes of construction of the basis matrix during the training simulation. Additionally, the nonlinear terms are gauged first through the evaluation of their set of grid blocks before being interpolated through the use of the basis matrices. On the contrary, the discrete empirical interpolation method is quite intrusive as it entails the computation of the full order non-linear terms through the help of a simulator when carrying out the reduced order runs. As a result, the method does not produce a surrogate model that can be run outside the full-order simulator hence its disadvantage. On the other hand, making use of the trajectory piecewise linearization method ensures that the solution arrived at each time set of the test simulation is adequately represented.

The POD-TPWL procedure may be divided into two major ways; the offline and online processes.

**Offline Processing:**
- Running of the reservoir simulator in order to adequately prepare the state vector as well as the matrices (X). Save the derivative terms $J^i$, $B^i$, and $C^i$ in the training runs.
- By adequately performing the SVD’s on the snapshot matrices ($X_p$, $X_s$) construct the basis matrix $\phi$.
- Compute the reduced state denoted by $Z^i_r$ as well as the reduced derivative matrices $J^i_r$, $B^i_r$, and $C^i_r$ from the training runs in order to be utilized in the process of linearization.

**Online Processing:**
- Given the denotation of $Z^n$, determine the training point $i$ to linearize through minimization of the $d^{n,j}$.
- Obtain the value of $Z^{n+1}$.
- Reconstruct the full order state $X^{n+1}$ from $Z^{n+1}$ at the appropriate locations.
- Calculate the flow rates.
- Repetition of the above mentioned steps up to the point at which the final simulation time is achieved.

### 2.2 Proxy Modeling in Literature

In the geological fields, numerous experiments and studies have been conducted in relation to pressure distribution as well as CO$_2$ plume. Some of the studies have come out to be a success while others do not have supporting evidence of being solid in nature. It is, however, noticeable that strides have been made as far as pressure and CO$_2$ predictions are concerned. Some studies have carried out the numerical simulation of the CO$_2$ injection in homogenous formations while others have numerically simulated the three-dimensional heterogeneous formations for the CO$_2$ sequestration. In one study conducted by Thanh et al. (2020), artificial neural network was used in order to undertake the role of predicting the performance of CO$_2$ enhanced oil recovery as well as storage in residual oil zones. The study had numerous objectives which comprise; creation of predictive models for the CO$_2$-EOR as well as storage in the residual oil zones. Another important objective was to generate rapid tool that is essential in reducing time consumable compositional reservoir simulation. Moreover, it aimed at validating the stability as well as the accuracy of the
artificial neural network by use of residual oil zones field located in the Permian Basin. Lastly, it was aimed at demonstrating the application of artificial neural network models for the optimization of the CO$_2$ injection process. In most cases, the residual oil zones are known to be potential formations for the carbon capture, utilization, and storage otherwise abbreviated as CCUS. In the study, uncertainty parameters such as the geological factors as well as well operations were used in order to come up with training database. Within the study, approximately 351 numerical forms of samples were simulated. The numerical samples were aimed at providing answers to the questions regarding artificial network. Within the results of the study, the training simulation results for the cumulative oil production, the cumulative CO$_2$ retained, and that of the CO$_2$ injection are provided. The model used approximately 300 samples for the training network. For the blind testing network, approximately 51 samples were used in order to avoid any issue related to overfitting. By use of the MATLAB, the artificial neural network model was effectively created by strict adherence to the 80%-10%-10% training plan. The exact training plan was further used for the three different targets. A population size of 240 which constitute about 80% of the population was used to determine the gradient as well as to update the weight and biases. A sample size of around 30 which translate to about 10% was then used in order to evaluate the network generalization. Another sample size of 30 which equally translates to a 10% was used to compare various models. The verification scheme may also double up as a means to evaluate the neural network performance during the training models.

For the effective performance of the artificial neural network, successful training is number one consideration in as far as its performance is concerned. Successful training is completed up to the point at which the verification and their testing curves are more or less similar to each other with comparison to the epoch numbers. Within the study, the cumulative oil production converged to a mean square of about 0.02519 at the fiftieth iteration. For the CO$_2$ injection the validation performance was at 0.30279 at the twentieth iteration while for the CO$_2$ retained, the validation performance was at 0.1259 at the fortieth iteration. Despite the revelations from the results pointing out at the artificial neural network prediction model being excellent in relation to the RMSE and the $R^2$, it was quite important to undertake a blind dataset test prior to putting the predictive model into real use in the Permian Basin. The blind dataset was put at a matrix of forty-one rows and eight columns. The results of the test portrayed that the value of $R^2$ was greater than 0.98 value which depicted the success associated with the blind testing validation for three data-driven models based on artificial neural network.

The results of the study outlined clearly that there is need for developing an artificial neural network predictive tools in order to evaluate the performance of CO$_2$-EOR as well as storage in the ROZs. The study further indicated that the artificial neural network had the capability of accurately predicting results by comparison with the initially reported data from real fields in Permian Basin. The correct and excellent evaluation capability of the model is anchored on the careful selection of the specific uncertainties parameters for training of the artificial neural network model. Furthermore, it is worth noting that the blind database testing process was quite important in the
bid to verify the accuracy of the artificial neural network models. In order to ensure success in developing any artificial model, it is just prudent for those involved to take care of the range of uncertainty parameters. It was also an important observation that the number of simulations is a very crucial aspect for hand-shaking reservoir simulation as well as the machine learning tools in a bid to develop artificial neural network models. Of particular interest was that the study made use of 351 samples for both the training and blind testing and that increasing the numerical value was directly proportional to improved predictive models- this means that an increase in number of population sample led to an increase in predictive models and vice-versa. As a result, the difference that was recorded between the predictive result and that of the field report was at about 8%. It is, therefore, worth noting that the number of samples in a study or experiment is very crucial as far as accuracy of the results is concerned. The previous studies prior to the development of artificial neural network did not take into consideration such a crucial account hence the results that were formulated were not up to standard as compared to those produced by the artificial neural network. Nonetheless, the study was very instrumental in providing the application perspective of the developed artificial neural network model by reinforcing it with SPO in order to hasten the optimization process. The study also highlighted the advantage associated with artificial neural network as that which could support conventional reservoir simulator in order to reduce by a huge percentage the time-consuming lapse for engineering applications which comprise optimization as well as sensitivity analysis. In a nutshell, the study proposes in a special way an innovative framework that ought to be used in the present generation in order to fasten the processes of oil and petroleum industry. The model proposed is of top-notch level and in a special way provides high levels of accuracy not experienced before. In order to achieve the reproduction of the model, it is highly recommended for the CO₂ storage formation which comprise of already used hydrocarbon reservoirs, saline aquifers, and the unconventional reservoirs. It is a complete model with specific characteristics that are useful within the petroleum and oil industry as a whole. This particular methodology can be used in various aspects of the CCUS, reservoir engineering, enhanced oil-recovery, as well as other important science disciplines.

In another study, Van Doren et al. (2206) made a proposal on reduced order model for purposes of production optimization in water flooding process through the use of the proper orthogonal decomposition. They produced the reduced order models through the use of data obtained from the numerous snapshots from the various model states- water saturation distribution and pressure distribution derived from the full order simulation model. In the experiment, the proper orthogonal decomposition was adequately used to summarize the dynamic variability of the full order reservoir model within the reduced subspace. Despite the fact that the number of state vectors is reduced through the help of the proper orthogonal decomposition, the difference exhibited from penta-diagonal to a full matrix acts against the computational advantage achieved by the vector size kind of reduction.

In order to confirm the efficiency of the study, the methodology was adequately tested on the water flooding scenario within a 2-D, 2-Phase model that contain the following characteristics; 2025 grid
blocks in numbers and 2 horizontal wells containing the control valves within each of the grid blocks. A comparison between the reduced order model and the original order model was carried out. This was achieved through simulating the reduced order model with the same controls as the full order model. The results indicated that the states were almost similar but upon tampering with the controls (blind set) the sets of the full order were observed to be less represented as compared to those of the reduced order model. In conclusion, it was worth noting that despite the fact that the POD methodology produces reduced order models with very low complexity, the real speed up on the simulation is modest in comparison with the size of the models. This is normally attributed to the fact that non-linear function used for the estimation of the state vectors will automatically be evaluated at the full order number of states. The main challenge that is connected to proper orthogonal decomposition is the fact that projection is directly dependent on the training inputs as well as the time scale of snapshot take.

Another attempt in the use of the proxy models is connected to National Resistance Assessment Partnership. In this example, the body majorly focused the use of the tool for risk assessment of carbon storage. This is achieved through the division of the carbon storage through various components namely; seals, wells, reservoirs, atmosphere, and groundwater. After the division, the proxy model is used for each component and all other models are duly integrated for purposes of assessing the success probability of carbon storage through the use of Monte Carlo Simulation. Various proxy models are put into practice such as the look-up table, PCE, AI surrogate reservoir models, and the response surface. The first model known as the look up table is quite simple but needs several runs of the high-fidelity model regarding various inputs. The aforementioned table is created with special regards to inputs, the results obtained after successful simulation runs, and another dimension representing the time step. The outcome as a result of the new scenario can adequately be found through interpolation-kind of approach from the created table. Despite the fact that this particular type of method is very fast, the problem comes in when coming up with the table as it requires full-physics models. In the work carried out by the National Risk Assessment Partnership Zhang and Paul (2012), a figure above three hundred simulation runs were required to adequately come up with a table for predicting the pressure as well as saturation at each and every grid block. The characteristics of the model were as follows; it was a 2-D, 2-Phase model containing ten thousand grid blocks and with only three variable parameters which comprise; reservoir permeability, reservoir porosity, and seal permeability. It is worth noting, however, that the heterogeneous field cannot apply in this particular approach and that permeability must be varied through a scalar multiplier. Within the work, various time snapshots were selected in the interval of 1000 years of post-injection. It is a fact that the accuracy of the model as well as the size of the look up table depends primarily on the selected snapshots as well as the duration between them.

Furthermore, Zhang and Paul (2012) came up with a reduced order model for purposes of carbon dioxide storage within the brine reservoirs. The main objective of the two was to actually use the reduced order model for the risk assessment of the geologic sequestration of carbon dioxide which
was actually a rising topic due to the effects of the gas on the environment. The study entailed building of a response surface derived majorly from a group of high-fidelity forward simulations for a particular set of parameters. The various approaches that were put into practice to ensure the success of the study included Gaussian process regression, a look up table in conjunction with a linear interpolation, and a radial basis function. The main goal of the study was to properly predict the pressure value at specific locations within the reservoir as well as at specific time frame in the reservoir. To further escalate the chances of accuracy, a relative error was used in order to quantify the accuracy in relation to approximation. Within the study, three parameters were used for purposes of building realizations which comprise: permeability of the sand layer, porosity of the sand layer, and permeability of the cap rock. The total sum of simulations used within this particular approach was at 57 which are somewhat lower in comparison to a typical surface response approach. One of the major limitations associated with this particular approach is that the reduced order model can only predict one output of interest. This, therefore, means that for purposes of predicting each parameter then the number of reduced order models required to be created will be equal to the number grid blocks and time steps hence rendering the approach to be prohibitive in nature.

Nonetheless, He et al. (2011) have put into practice the POD-TPWL which is a hybrid of both the trajectory piecewise realization and the proper orthogonal decomposition. In their study, they aimed at building a reduced order model to help carry out certain fundamental tasks. It is common knowledge that employing of proper orthogonal decomposition as a single entity may primarily result into high order complexity due to the requirement of projection and construction of the nonlinear terms. As a result of that, the trajectory piecewise realization was used in order to address the problem. One of the most common limitations of this kind of approach is that the system stability is attributed to the type of projection scheme employed. Moreover, this particular method requires huge levels of offline processing in order to properly construct the POD-TPWL model. For purposes of registering more accurate results, numerous simulations runs ought to be carried out which in a negative way results in computational and storage problems. It is a necessity that some modification ought to be done on the reservoir simulator thus it is inapplicable through the use of the conventional reservoir simulator. Also, it is open that the number of the variables is the same as the multiplication of the number of the fluid components and the number of the grid blocks.

In the methodology, the offline computation entails running the full order training simulations, saving as well as reading of not only the states but also the derivatives, producing of the basis matrices, and reducing of both the states and the derivatives. As such two or three training simulations may require the equal amount of time taken for a one full order simulation run. This particular kind of technique has been tested on a small reservoir with very few wells which are approximately less than ten in number. The main variables for this kind of technique remain to be pressure and the component mole fraction. In order to adequately achieve the flow rate, the full order primary variables ought to be modified at specific locations and time while on the other hand
the secondary variables ought to be adequately calculated through the performance of the flash calculations together with the primary variables. Also, adjusting the number of wells may result in increased variability in states which may produce a tangible effect on the model results as well as computational expenses. These techniques are unlikely to provide a significant breakthrough in this field. Furthermore, most of these methods have only been used on academic situations involving a small number of wells. The real test will come when they try to illustrate the capabilities of these techniques by using them to develop proxies for full-field industry-based numerical models with hundreds of wells and millions of grid blocks, such as the numerical reservoir model discussed in this study (Mohaghegh et al. 2015).

2.2 Smart Proxy Modeling in Literature

Smart Proxy Technology (also known as surrogate reservoir modeling or SRM) was first designed by Prof. Shahab Mohaghegh at the West Virginia University in 2006. Since then, there has been several studies focused on application of such technology by the Laboratory for Engineering Application of Data Science (LEADS) at the West Virginia University Department of Petroleum and Natural Gas Engineering (PNGE) (Jalali, Mohaghegh, 2009; Kalantari-Dahaghi, Mohaghegh, et al. 2011; Mohaghegh, Amini et al. 2012; Amini, Mohaghegh, et al., 2014; Shahkarami, Mohaghegh, et al, 2014; Gholami, Mohaghegh, et al, 2014; Mohaghegh, Gaskari, et al. 2014; Haghighat, Mohaghegh, 2015; Alenezi, Mohaghegh, 2017; Alabboodi, Mohaghegh, 2021). The technology can be characterized as well-based (Mohaghegh et al. 2012; Shahkarami et al. 2014) or grid-based (Mohaghegh et al. 2012). (Gholami et al. 2019; Alenezi and Mohaghegh 2017). A well-based SRM's purpose is to simulate reservoir response in terms of production at the well location (or injection). The grid-based SRM, on the other hand, enables you to simulate any dynamic reservoir parameter, such as pressure, phase saturations, or fluid component composition, at any time or location (grid block). The Smart Proxy Model differs from other types of proxy models in its construction. It employs artificial intelligence and data mining techniques. The advantage of using this proxy technique is that it generates an approximate replica of a numerical reservoir simulation in a relatively short period of time. Based on its purpose, a Smart Proxy Model may be divided into two types: a well-based smart proxy model and a grid-based smart proxy model that can be paired with a well-based model in specific cases. The method is known as "Smart Proxy Modeling" because of its unique ability to properly duplicate the pressure and saturation distribution across the reservoir at the grid block level and at each time step (hence Dynamic Smart Proxy Modeling) without affecting the original numerical simulation model's physics or resolution. Understanding the amount of pressure and saturation fluctuations across the geological formation used for CO2 storage, especially beyond the injection wells, is crucial in the design of the multiple reservoir engineering operations involved with CO2 geological storage. To achieve this goal, a mix of reservoir engineering and reservoir modeling domains, as well as machine learning and data mining in SPM, is required (Smart Proxy Modeling). The technology's goal is to learn the mechanics of fluid flow in porous media from numerical simulation model data in order
to replicate the outcomes of various scenarios. SPM accurately copies the pressure, CO₂ saturation, and temperature across the reservoir with a very high resolution (at every grid block). This is an important aspect of SPM. These models are far more valuable now that they can be created in seconds rather than hours or days (Amini et al. 2014; Alenezi and Mohaghegh 2017; Mohaghegh 2015; Mohaghegh et al. 2015; Gholami et al, 2019). While there have been multiple studies on well-based proxy modeling, there have been very few studies on grid-based proxy modeling for a black-oil reservoir model, much alone compositional simulations. In a study by Alenezi and Mohaghegh (2017), they used smart proxy modeling at both the grid block and well level in another investigation in the SACROC unit field in Scurry County, Texas. They focused on a cascade training and validation method, in which the inputs for each time step are generated from the output of the previous time step until the final time step is reached (except for the first time step). This feature enables the smart proxy model to feed itself from dynamic data sources that it has selected. The intricacy of the production performance and its geological classification make it an ideal candidate for assessing smart proxy performance. To build a smart proxy model, multiple reservoir simulation scenarios were necessary. For various operating restrictions and geological conditions, multiple reservoir modeling scenarios were developed. Geological parameters and results from the desired simulation runs were gathered to generate the spatial-temporal database. To select the data needed to create the smart proxy model, key performance indicators were chosen as input parameters. After training the smart proxy model, it was utilized to validate it using a blind numerical simulation run. In order to acquire a better understanding of field performance, a second proxy model was built to construct the well-production profile. Both smart proxy models delivered highly accurate findings.

Smart proxy models, also known as surrogate reservoir models, are high-fidelity reservoir approximations capable of accurately reproducing the behavior of full field models in response to changes in all input parameters (reservoir characteristics and operating constraints) in seconds. Reservoir engineering and modeling are combined with data mining and machine learning in SRM. According to Mohaghegh, SRM is "an ensemble of multiple, interconnected neuro-fuzzy systems that are trained to adaptively learn the fluid flow behavior from a multi-well, multilayer reservoir simulation model, such that they can reproduce results similar to those of the reservoir simulation model (with high accuracy) in real-time" (Mohaghegh et al. 2012).
CHAPTER 3: RESERVOIR SIMULATION MODELING (CMG)

3.1 Overview:

Petroleum engineers employ numerical reservoir modeling as a typical tool to model hydrocarbon reservoirs. The reservoir's performance is frequently forecasted using a successful history matched model. They're also utilized to create the best field development strategy by simulating numerous different situations. Numeric reservoir simulation was employed in our study to produce the necessary data for the artificial neural network development and testing.

3.2 Reservoir Geological Structures Description:

Understanding fluid dynamics via porous media in subsurface reservoirs necessitates creating a reservoir simulation based on geology and petrophysical parameters. The reservoir model construction approach used in this work is discussed in the sections below. The GEM simulator was created by the Computer Modeling Group (CMG). GEM is a sophisticated general equation of state compositional simulator with features including equation of state, dual porosity, CO₂, miscible gases, volatile oil, gas condensate, horizontal wells, well management, complex phase behavior, and more. This study used the CO₂ module of the simulator to replicate CO₂ injection and sequestration into an aquifer formation.

The model consists of 211 x 211 grid blocks in X and Y direction. In addition, the model's geological structure includes 30 layers in Z direction, 28 of which were mixture of conductive and non-conductive rock types. The layers 1 to 2 define the top shale barrier layers. The reservoir structure and well locations are depicted in Figures Figure 1. Cross-sectional view for the reservoir model. and Figure 2. 3D View for the Porosity Model Geometry and the relative position of each injection well. All the layers have 10 ft thickness.
Figure 1. Cross-sectional view for the reservoir model.

Figure 2. 3D View for the Porosity Model Geometry and the relative position of each injection well.
3.3 Geological Reservoir Realizations Development

Algorithms in (artificial neural network) ANN must be trained on enormous volumes of data to learn the complicated patterns of the fluid in the numerical reservoir simulation in order to construct the Smart Proxy Model. During the design of the numerical reservoir simulations, two factors were considered as variables to be investigated further. The first was porosity, and the second was permeability distribution. The geological realizations have the same broad porosity distribution pattern and a similar range of values, but their distributions are distinct and varied. The purpose of these models was to create a convincing, hypothetical heterogeneous reservoir with reservoir characteristics that differed geographically between layers.

The National Energy Technology Laboratory (NETL) started work with a number of universities and national laboratories across the country. They constructed 100 realizations by altering porosity and permeability across seven categories (P5, P10, P25, P50, P75, P90, P95), with P5 and P95 having the lowest and greatest average porosity and permeability, respectively, using a single numerical reservoir simulation model. Because the abandoned realizations' porosity and permeability values were unrealistically low / high, NETL decided to divide the study realizations into three midrange groups (P25, P50, and P75). From the remaining 61 possibilities, 46 are selected. 40 realizations are used for smart proxy training and calibration, and 6 are used for blind validation in those three categories.

3.3.1 Porosity:
The percentage of vacant space in a porous media is called porosity. It has no units and is measured as a fraction. The degree of cementation determines the porosity of cemented materials.

Figure 3. Porosity Distribution for the 46 realizations demonstrates the porosity distribution map for all the 46 realizations used in this study. It's worth noting that some sections of the reservoir, as well as all realizations in general, have a high porosity distribution; nevertheless, no two realizations' porosity distributions are similar.
3.3.2 Permeability:

The ability of a rock to transfer fluids is measured by its permeability. It can also be described as a measure of pore space connectivity. Darcy (d) or millidarcy (m) are the most frequent units of measurement (md). Permeability is a rock attribute that varies depending on the type of rock. In contrast to porosity data measuring methods, which are inexpensive, directed permeability measurements are expensive and difficult to collect. Figure 4 demonstrates the permeability distribution map for all the 46 realizations used in this study.
Since the permeability skewed in low values, the plot is zoomed in to show the distribution (Figure 5. Zoomed in Permeability Distribution for the 46 realizations).
3.3.3 Relative Permeability Curves:

When only one fluid flows through the porous medium, the absolute permeability of the medium is the permeability of the porous media to the fluid. When two or more fluids flow through the porous media at the same time, the flow path must be shared by all of them. As a result, the porous medium's effective permeability to each fluid is less than its absolute permeability. The effective permeability of each phase is measured in terms of relative permeability, which is a dimensionless quantity. The relative permeability value ranges from zero to one. In a black oil model, relative permeability of each phase is usually computed from two-phase relative permeability. Endpoint saturations relative permeability at endpoint saturation is used to build two-phase relative permeability curves, oil-water relative permeability, and oil-gas relative permeability.

There are 3 rock types extracted from the numerical reservoir simulator. However, when we plot the relative permeability curves, we noticed that rock types 1 & 2 acting the same under water / gas saturations condition. Therefore, there are only two rock types that exists in the reservoir. The relative permeability curves are shown in the following Figures Figure 6. Relative permeability curves for rock type 1 to Figure 14. Water & Gas Relative permeability curves for rock type 3.
Figure 6. Relative permeability curves for rock type 1

Figure 7. Oil & Gas Relative permeability curves for rock type 1
Figure 8. Gas & Water Relative permeability curves for rock type 1

Figure 9. Oil & Water Relative permeability curves for rock type 2
Figure 10. Oil & Gas Relative permeability curves for rock type 2

Figure 11. Water & Gas Relative permeability curves for rock type 2
Figure 12. Oil & Water Relative permeability curves for rock type 3

Figure 13. Oil & Gas Relative permeability curves for rock type 3
It is important to note that these parameters including the relative permeability curve were all kept constant across the 46 realizations. The only data that was different across each realization is porosity, permeability.

3.4 Reservoir Simulation
The GEM simulator from the Computer Modeling Group (CMG) was used for numerical reservoir simulation part of this study. GEM is a sophisticated general equation of state compositional simulator with features including equation of state, dual porosity, CO₂, miscible gases, volatile oil, gas condensate, horizontal wells, well management, complex phase behavior, and more. This study used the CO₂ module of the simulator to replicate CO₂ injection and sequestration into an aquifer formation.

The modeling of CO₂ storage in saline aquifers entails solving the component transport equations, the equations for thermodynamic equilibrium between the gas and the aqueous phase, and the geochemistry equations, which involve reactions between the aqueous species and mineral precipitation and dissolution.

For the numerical reservoir simulation section of this work, the Computer Modeling Group (CMG GEM )'s simulator was used. GEM is a powerful general equation of state compositional simulator
that includes features like equation of state, dual porosity, CO₂, miscible gases, volatile oil, gas condensate, horizontal wells, well management, complex phase behavior, and more. CO₂ injection and sequestration into an aquifer formation, were all replicated using the simulator’s CO₂ module. Solving the component transport equations, the equations for thermodynamic equilibrium between the gas and the aqueous phase, and the geochemistry equations, which involve reactions between the aqueous species and mineral precipitation and dissolution, are all part of the modeling of CO₂ storage in saline aquifers (Ngiem 2004).

3.4.1 CO₂ Injection Design (Number of Injector, Injection Time)
The reservoir simulation model includes four vertical injection wells, as previously described. In all 28 sand layers (#3-30), these injection wells were pierced and completed. The injection scheme was created with two primary and secondary limitations in mind. Bottom Hole Pressure is a well-level main constraint, and Max Injection Rate is a group-level secondary constraint. The group rate and bottom-hole pressure (BHP) operational constraints for the CO₂ injection well are 103.6 MCF/day (standard condition at surface) and 2,814 psi, respectively. Based on a 10% buffer from the reservoir fracture pressure, the maximum permissible BHP of 2,814 psi was chosen. The injection began on January 1, 2020 and will last 10 years until January 1, 2030.

The simulation was conducted during this 10-year period and four timesteps (01-01-2021, 01-01-2025, 01-01-2030 and one timestep from the post-injection period 01-01-2100) were selected for prediction by the smart proxy model and the pressure and saturation values were extracted in order to be later compared with the Smart proxy Model outputs. The distribution of Pressure and Saturation for all the 46 realizations are presented in Figures Figure 15. Pressure Distribution for the 46 realizations to Figure 17. Zoomed in CO₂ Saturation Distribution for the 46 realizations.
Figure 15. Pressure Distribution for the 46 realizations

Figure 16. CO₂ Saturation Distribution for the 46 realizations
3.5 Summary:

Each of the aforementioned reservoir realizations has over 300 static and dynamic features (attributes) retrieved from the output of the numerical reservoir simulation as inputs to the ANN models, each with a set number of wells and rock types. At the end of the injection timestep, the runs are used to generate reservoir pressure and CO$_2$ saturation data. Because the reservoir’s wells have all been put into injection since the start of the injection program (01-01-2020), the pressure has already developed and propagated to a great extent surrounding the injection wells at the given timestep.

The extracted injection, pressure data from the simulation runs, coupled with some additional attributes generated by petroleum engineering domain expertise, are used to educate the artificial neural network the foundations of fluid flow through porous media and improve its performance. The next chapter will explain and show such a feature engineering technique.
CHAPTER 4: METHODOLOGY (SMART PROXY MODELING)

4.1 Review:
This chapter examines the methodologies employed in the design of artificial neural networks in general and Smart Proxy Modeling (SPM) in particular. During the training development and blind deployment procedures, some of the challenges encountered and the solutions discovered are highlighted. This chapter will give a high-level overview of the procedures used to create the Smart Proxy for this project. Guidelines for creating an ANN prediction tool and feature creation strategies are also offered, based on observations made during the development and training process. A blind simulation scenario was created and conducted in model’s primary evaluation phase. If the evaluation is satisfactory, the SPM will be exposed to the blind-validation dataset, which is again a simulation scenario that has “never been” seen by the neural network. This step would be the main performance indicator of the SPM.

4.2 Design of Artificial Neural Network and Smart Proxy Model
Today, artificial neural networks are widely used in variety of petroleum engineering applications. However, due to a lack of documented criteria for designing ANN prediction networks, users have been forced to rely on trial-and-error at every stage of the development process. A set of rules has been produced based on observations made during the creation of an ANN prediction tool for history matching. The design of an ANN prediction tool can be separated into three groups:

1. Network structure,
2. Input/output parameters,
3. Data formulation and structure.

Each of the aforementioned parameters will be addressed in further depth in the following.

4.2.1 Network structure:
ANN design entails optimizing a number of parameters. The variables which define the general structure of a neural network are called hyper-parameters. The following variables are taken into consideration when determining the best architecture for ANN:

• Number of neurons in each layer – the number of neurons within the hidden layers of the network that produce the desired characteristics of the output layer
• Transfer functions – a linear or non-linear function in each layer that is chosen to satisfy some specification of the problem that the neuron is attempting to solve

• Training algorithm – A function that measures the network’s performance and changes the weight and bias values in each layer of the network.

As mentioned, the structure and topology of a neural network are controlled by a range of parameters, and it can theoretically take on an infinite number of possible shapes. Almost all of them, on the other hand, consider a number of factors such as the number of hidden layers, the number of hidden neurons in each hidden layer, the activation functions combined, and the type of the connections between neurons. The purpose of this section is to provide a quick overview of some of the most prevalent structures, with a focus on those that have proven to be effective when used to build data-driven models for oil and gas applications. Fully connected neural networks are the architectures that have been used most successfully in data-driven models in terms of neuron connectivity. As shown in Figure 18. A simple schematics of a single-layer fully connected neural network structure with one output, each input neuron is connected to each hidden neuron, and each hidden neuron is connected to the output neuron in the same way. This is known as a fully connected network. The NNs in this study were built using only one hidden layer. The most basic and extensively used type of neural network for data-driven model construction, serving as the data-driven model's main engine. This is a three-layer neural network that is fully connected. The three layers are the input layer, the concealed layer, and the output layer. Furthermore, while the output layer can contain several output neurons, our data-driven model experience has shown that, with a few exceptions, a single output neuron in the output layer performs best. Artificial neural network has several neurons which are arranged in layers and the connections between these layers constitute the network architecture. They can be either a single layer network or a multilayer network. Single layer networks have one input layer and one output layer. Most problems do not require more than four hidden layers (Fausett, 1994). For this study, it has been shown that single-layer feedforward network with back propagation works well (Alaboodi, 2021, Faisal 2017). The actual number of neurons in each hidden layer depends on the number of input and output neurons.

The WVU LEADS (Laboratory for Engineering Application of Data Science) collective experience has shown that the quality and information content of the spatio-temporal database, rather than the topology or structure of the neural network, dictate the likelihood of a data-driven project succeeding.
4.2.1.2 Backpropagation

Feedforward networks with backpropagation are simple to set up, train faster than other networks, and accurately address a wide range of problems (Centilmen, 1999). They work in two stages. The feedforward step is the first. During this step, the input pattern is delivered to the input layer, and the data is conveyed to the output layer via hidden layer(s). As they go from layer to layer, transfer functions process the data. The second phase is the backpropagation step, which involves calculating the gradient of the network's error with respect to the network's modifiable weights using backpropagation. The response of the network is compared to the desired output, and mistakes are transferred from the output layer to the inner layers during this process. The network weights are adjusted using this error signal. According to Ali (1994), each intermediate layer receives a portion of the overall error signal based on the proportionate contribution of the unit to the original output. As a result, after multiple cycles of this operation, the resulting error signal becomes minimal. The network is considered trained for the intended function at this point. It must be capable of making predictions based on a new set of data.

Each neural network's hyperparameters were fine-tuned based on the size of its spatio-temporal dataset and the initial training observation. Some hyperparameters have been discovered to have a major impact on the training performance of any of those neural networks, including but not limited to learning rate, number of neurons in hidden layers, and activation function. However, setting the hyperparameters is mostly determined by the nature and scope of the problem as well as the data, as well as the level of experience of the machine learning engineer. A sample of...
hyperparameters utilized for one of the NNs in this investigation is shown in Table 1. Typical hyperparameters used in this study for one of the SPMs.

Table 1. Typical hyperparameters used in this study for one of the SPMs

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Values \ Type</th>
</tr>
</thead>
<tbody>
<tr>
<td># Hidden Layers</td>
<td>1</td>
</tr>
<tr>
<td># Neuron in Hidden Layer</td>
<td>5000</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.00001</td>
</tr>
<tr>
<td>Activation Function (Input Layer-Hidden Layer)</td>
<td>Relu</td>
</tr>
<tr>
<td>Activation Function (Hidden Layer-Output Layer)</td>
<td>Sigmoid</td>
</tr>
<tr>
<td># Epoch</td>
<td>100,000</td>
</tr>
<tr>
<td>Batch size</td>
<td>10,000</td>
</tr>
</tbody>
</table>

An epoch of training is completed when the neural network has processed all of the data records in a training set and the error between the neural network output and the real output has been established. To put it in another way, the epoch refers to the process of passing an ENTIRE training dataset through the neural network ONCE. In general, stopping criteria are used to identify when the training process should come to an end. When the difference between the projected and actual output is less than a pre-determined threshold, the training procedure ends. The training data can be broken into smaller bits of data termed batch size during the training process.

4.2.3 Input/output parameters:
This step involves making several critical decisions about the data being presented to the network during training phase of the process. Some of the significant ones are, 1. Data generation strategy, 2. Number of data sets to be presented to the network during training phase

1. Data generation strategy: To learn the complicated patterns of the fluid flow in the numerical reservoir simulation, the SPM algorithms in Artificial Neural Networks (ANN) of the Smart Proxy must be trained on vast volumes of data. Neural network generally works better at interpolation than extrapolation. So, the data presented to the network during training stage should cover the whole spectrum of parameter values in the output layer.

2. Number of data sets: Number of data sets used in the training phase of the network development is very critical. Use of too few data sets may lead to a network that has not captured the actual relationship between input and output parameters. Use of too many data sets may lead to a network that memorizes the values rather than understanding the relationship between input and output parameters. Both scenarios are undesirable since neither will give good predictions for actual problem. Apart from this, time required to
train the network also increases as the number of data sets increase. So, there is always incentive to reduce the number of data sets. It is essential to find an optimum number of data sets that give good results. In this study, 40 realizations were used in training.

4.1.3 Data formulation and structure:

Data formulation is an important step that determines the performance and accuracy of predictions made by the network. The next step is to quality-check the data formats and patterns after performing the numerical reservoir simulation runs that were designed in the previous chapter and therefore collect all the necessary data from each of them.

The spatio-temporal database that will be discussed in the Section 4.1.3.4 will serve as the smart proxy model's foundation and the main input to the ANN using the extracted data. The 40 realizations used in the training consists of about 53 million cells and each cell has 296 features or attributes (columns) that are numerical values. The output target for prediction (Pressure or CO₂) is also one of these columns that is kept separated from the training process.

4.1.3.1 Data Preparation:

Data extracted from numerical reservoir simulator (CMG). some data has daily values need to be converted to monthly bases to end up with uniformed data unites.

- Bottom hole pressure (Bhp) extracted as cell based daily, needs to be converted to monthly by averaging the daily records for each month. (unit PSI)
- Injection rate extracted as cell based daily for each injector, needs to be converted to monthly by summating the daily records for each month. (unit cubic feet)
- Grid bottom (it shows the bottom depth of each cell and it’s a static value across all realizations). (unit ft)
- Grid Top (it shows the top depth of each cell and it’s a static value across all realizations). (unit ft)
- Grid centroid X (it shows the location of each cell in x axis and it’s a static value across all realizations). (unit ft)
- Grid centroid Y (it shows the location of each cell in Y axis and it’s a static value across all realizations). (unit ft)
- Grid Paydepth (it shows the depth of each cell center in Z axis and it’s a static value across all realizations). (unit ft)
- Grid Thickness (it shows the thickness of each cell and it’s a static value across all realizations and all cells). (10 ft)
- Permeability extracted for each cell and it’s changing across realizations. (unit darcy)
• Porosity extracted for each cell and it’s changing across realizations. (unit %)
• Rock type (it shows the rock type of each cell and it’s a static value across all realizations). There are 2 rock types in this project (unitless)
• Pressure extracted as cell based monthly. (unit PSI)
• Gas Saturation (CO₂ Saturation) extracted as cell based monthly. (unit %)

Additionally, normalizing the data standardizes the numerical range of the input data and enhances the fairness of training by preventing an input with large values from swamping out another input that is equally important but with smaller values (Al-Fattah, 1994).

4.1.3.2 Data Processing:
The numerical reservoir simulation runs (CMG) generated a considerable amount of raw data (outputs and inputs) that were analyzed and then converted into a usable format. Python starts by reading the simulation output/input, then extracting and saving any usable data in a specified manner. At first glance, this may appear to be a straightforward process, but it requires a large amount of calculation time as well as a full comprehension of the important reservoir simulation data. Choosing the input parameters that will be used to train a neural network from the variables (potential inputs) that have been assimilated in the database is not a simple operation. A project database often contains a large number of parameters, all of which might be used as input parameters for the neural networks that will be used to train the data-driven model. There are static and dynamic characteristics supplied, as well as equivalent data for various offset wells. These values are placed in a flat file as columns and will be used to train data-driven models (neural networks).

4.1.3.3 Features Generation and Engineering:
Feature engineering refers to the process of applying domain expertise to choose and transform the most essential variables from raw data while constructing a predictive model using machine learning modeling. Furthermore, feature engineering and selection is used to improve the performance of machine learning (ML) algorithms by teaching them reservoir engineering expertise using data from the numerical simulation model. The feature engineering in this project supplies features or information to the machine learning in order for it to learn the mechanics of fluid flow in porous media.

In addition to the static and dynamic features that were extracted from the CMG model, new futures were created after analyzing and understanding the data structure in order to improve the ability of models learning to make exact forecasts. Because the data is organized in a cell-based format, each row in the dataset will represent a focus cell in the model; the model’s dimensions are 211 cells on the "I" axis, 211 cells on the "J" axis, and 30 layers on the "K" axis. x, y, and z are represented by i, j, and k, respectively. Each realization has a total of 211x211x30 = 1,335,630
cells. Each new future will be a column with a value for each focus cell, implying that the characteristics will be generated as columns.

Distance from the focal cell to the (closest, 2nd closest, 3rd closest, 4th closest) injectors (4 new attributes): The position of each cell should be individually defined and used as input to the ANN Model in order to offer enough information for the ANN Model to learn the behavior of each cell in the system. Calculating the distance between each cell and the reservoir simulation model boundary can be used to depict the location of each cell in the reservoir simulation model (which include distance to top, bottom, east, west, north, and south boundary). In addition, each cell's three indexing I, J, and k) is utilized to identify its position. Furthermore, each cell in the reservoir simulation model was given a unique number (cell ID) as an input to the ANN to aid in recognizing the unique location for each cell (Figure 19. Schematics showing the generation of new features).
After generating the distance to each injector, the best way to represent injection rate, cumulative injection and bhp is to arrange them in the same fashion (injection rate of the closest injector, cumulative injection of closest injector, bhp of the closest injector… etc) (12 new attributes)

We also apply the same method (injection rate of the closest injector, bhp of the closest injector for layers above and below the layer of focal cell, that means if the focal cell in layer 4, layers 3 and 5 values will be included also as new attributes to account for any chance of communication between layers.

Rock type was extracted from the numerical simulator as one attribute has (1, 2 and 3), we converted that attribute to 3 new attributes, each column will represent one rock type as binary (either 1 or 0)

Tier cells are cells surrounding the focal cell, there are three types of tier cells (face, line, point).

Face tier cells have face contact with the focal cell and for each focal cell there are 6 face tier cells
Line tier cells have line contact with the focal cell and for each focal cell there are 12 line tier cells
Point tier cells have point contact with the focal cell and for each focal cell there are 8 point tier cells (Figure 20. Schematics showing the implementation of tier cells).

From tier cells, new attributes were generated. Permeability, porosity, initial pressure, rock type (3 attributes) so we end up adding 6x26 = 156 new attributes.

While running the numerical reservoir simulator, there were some instances that the cell value will be very low value or empty due to difficulty of converging a very small permeability. Therefore,
4 new attributes generated to show the number of inactive face tier cells, inactive line tier cells, inactive point tier cells and total number of inactive tier cells for each focal cell (Figure 21. Schematics showing the implementation of tier cells and inactive cells).

Since there are only 2 relative permeability curves for the 3 rock types, we can consider rock type 1 & 2 as sealed (very low permeability) and rock type three as conductive, therefore we generated a new attribute will show a binary value for each cell, if the rock type in the focal cell is 1 or 2 the value of the new attribute will be 0 and if it is rock type 3, the value will be 1.

There are 2 uncompleted or not perforated layers (layer 1 and 2), and layers 3-30 are completed. So a new attribute is generated with binary value to show if the focal cell is in a completed layer or sealed (1 or 0) respectively.

To strengthen the understanding of different rock types in the model, effective relative permeability was introduced to communicate the relation of relative permeability curves with each rock type. 4 relative permeability values were taken on each curve at similar interval resulting in 8 values with 4 krg values and 4 krw values (Figure 22. Relative permeability as input to the model).
\[ K_{eff} = K_{abs} \times K_{rg} \]

Equation 28

Figure 22. Relative permeability as input to the model

That was applied for focal and face tier cells (8 attributes for focal cell and 8 for each face tier cells 8x6) 56 new attributes.

Most of permeability values are laying below 100 md while some values go up to 12,000, this makes normalization process less efficient. Therefore, to overcome this issue we implemented a method called “power transformation”.

Power transforms are a family of parametric, monotonic transformations that are applied to make data more Gaussian-like. This is useful for modeling issues related to heteroscedasticity (non-constant variance), or other situations where normality is desired (Figures Figure 23. Violin plots of permeability distribution to Figure 26. Power-transformed permeability data).
Figure 23. Violin plots of permeability distribution

Figure 24. Zoomed in Violin plots of permeability distribution
When permeability was normalized, we noticed skewness issue

![Figure 25. Raw permeability data (non-transformed)](image)

After applying power transformation to permeability and normalization

![Figure 26. Power-transformed permeability data](image)

This technique was implemented to effective permeability attributes as well since they are derived from permeability.
4.1.3.4 Spatio-Temporal Dataset Construction

As previously mentioned, the spatio-temporal dataset refers to a set of data that includes both spatial and temporal parameters for the research topic. The geological features, as well as created features and simulation outputs, were collected in this study to create a spatio-temporal dataset. Using the produced spatio-temporal database, the Smart Proxy is then taught the principles of fluid flow through porous media, as well as the complexities of the heterogeneous reservoir represented by the geological model and its impact on fluid flow and pressure variations in the reservoir. Running design with a methodical approach aid in obtaining the most information while decreasing computational time. Several artificial neural networks were built using the produced input file, including networks for pressure, gas saturation. The training approach was the "Back Propagation" algorithm. By the end of each training epoch, the mistake is given back to the network in this approach (C. M. Bishop 1995). There is one hidden layer and one output in each network. Figure 16 shows a typical neural network architecture employed in this study.

The rows in a spatio-temporal database represent the records/samples or realizations in a dataset, whereas the columns represent the features or qualities. Each model (realization) of a numerical reservoir simulation generates more than one million records (grid cells). Because there are 46 realizations in total, the total number of records in the spatio-temporal database will exceed 61 million. Each entry in the spatio-temporal database represents the static and dynamic properties of a given grid block in a specific run and time-step.

4.1.4 Data Partitioning:

Each SPM's development data is separated into three pieces at random: training, calibration, and validation, in the following order: 80% Training, 10% Validation, and 10% Calibration for the pressure model but in gas model no validation dataset was used however, we used a 20% Calibration. By far the largest of the three components is the training dataset. This data is used to train the neural network and build relationships between the input and output parameters. The training dataset must contain everything that a data-driven model needs to learn. It's vital to remember that the range of parameters in the training set determines the applicability of the data-driven model.

The calibration data is used to analyze the ANN's performance throughout the training phase, while the training data is used to train the ANN. After the training phase is done, the trained ANN is validated using the Validation set. The calibration dataset works as a watch dog that watches the training process and decides when to stop it, because the network is only as good as its prediction of the calibration dataset (a randomly picked dataset that is actually a blind dataset). In addition, six of the reservoir simulation runs were designated as blind runs. These blind runs are never used during training; instead, they are stored for to be use in the verification process, presuming that they are new reservoir simulation runs.
The validation dataset is the last but possibly most important (segment). This dataset has no bearing on the training or calibration of the neural network. It was picked and set aside from the beginning to be utilized as a blind dataset. It sits on the sidelines till the session is over, doing nothing. The generalization ability of the trained neural network is tested firstly using this dataset before it goes to the final test (blind validation).

4.2 Training Process of ANN and SPM:
For neural network training, training simulation runs are simulation scenarios that give a sample space of model input-output interactions. The training simulation runs were meant to take into consideration both areas of concern because the surrogate reservoir model was constructed to monitor CO₂ injection under various operational limitations and geological realizations. The goal of this sort of proxy modeling is to use data to represent the complete reservoir system as well as the processes that occur inside it. To adequately introduce the entire system and illustrate the input-output relationship, data from many domains is required. The static data does not change over time and contains information about the reservoir structure, such as grid block locations and reservoir metrics like porosity and permeability. Other forms of data that must be determined are the distance of the grid blocks from the injection well, the relative grid distance from the reservoir's limits, and so on. Other parameters, such as the starting pressure values, phase saturation, and CO₂ model fraction of each grid block, are treated as constant variables over time. The parameters in the well domain and the grid block domain make up dynamic data. The well domain parameters are related to the well limitations, which often include production/injection rates or changing well bottom-hole pressures over time. The state variables of the systems (pressure, gas saturation, and CO₂ mole fraction) that change over time are known as dynamic parameters at the grid blocks.

A network is ready to be trained once it has been structured for a specific application. The starting weights (explained in the next section) are picked at random to begin this procedure. The training, or learning, process then begins. We're simply trying to solve an optimization problem when we train a model. We're attempting to optimize the model's weights. Our goal is to determine the weights that transfer our input data to the correct output value, the most accurately. This is the mapping that the network needs to learn. The iterative learning process in which data cases (rows) are provided to the network one at a time and the weights associated with the input values are modified each time is a critical aspect of neural networks. After all of the instances have been given, the process is frequently restarted. The network learns during this phase by modifying the weights in order to anticipate the right output value of input samples. Because of the connections between the units, neural network learning is also known as "connectionist learning." Neural networks have several advantages, including a high tolerance for noisy input and the capacity to classify patterns on which they have not been trained.

The network uses the weights and functions in the hidden layers to process the records in the training data one at a time, then compares the generated outputs to the desired outputs. The system then propagates the errors back through the system, causing the weights to be adjusted for the next
record to be processed. As the weights are regularly changed, this process repeats again. The same set of data is processed several times throughout the training of a network as the connection weights are continually improved.

4.3 Error Measurements:

The error term essentially indicates how well your network performs on a specific (training/testing/validation) piece of data. A low error is desirable, while a high error is unquestionably undesirable. A loss function, of which there are several, is used to determine the error. The Mean Squared Error, for example, calculates the distance between the desired output and the actual output before squaring the result.

A validation dataset is used to validate the proposed smart proxy model. The model's precision in regard to the blind set must be determined. In this study, the precision is calculated by subtracting the numerical simulator result from the smart proxy model output. Because the created smart proxy generates output at each grid block, the inaccuracy should be checked at each grid block. Different error calculation formulas are used depending on the type of output data. The quality of ANN model would be validated using the following formulas. For the pressure output, the following error formula is used:

\[
\text{Absolute Error Percentage} = \left(\frac{\text{absolute (ANN Output - CMG Output)}}{\text{CMG Output}}\right) \times 100
\]

The nature of data for CO₂ saturation data is different. Because the values are between 0 and 1, the error formula which was used is as follows:

\[
\text{Absolute Error} = \left(\frac{\text{absolute (ANN Output - CMG)]]}}{\text{CMG Output}}\right) \times 100
\]

The error between the predicted ANN output and the numerical reservoir simulation model outputs illustrates the SPM's accuracy.

4.4 Validation of the trained ANN with validation dataset:

The next step in the Smart Proxy development process is to validate the neural network using validation datasets after it has been acceptably trained. The validation dataset was never used during training; rather, it is used after training is completed and before deploying the Smart Proxy on completely new simulation runs (blind validation), as previously stated. The final step in the SPM development process is to test the results of the developed model on a blind validation. A blind validation is a simulation scenario that has “never been” used to train neural networks.
4.5 Smart Proxy Deployment:

After the training processes for both SPMs (pressure and saturation models) were finished, the models were calibrated and validated to internally test their performance. When the trained NN's performance has been confirmed to be satisfactory (by stopping criteria), the deployment procedure is started using the blind validation dataset, which consists of six numerical simulations that were never used during the neural network's training. The deployment method employs the six blind-validation realizations that were set aside. It must be noted that the neural network had never viewed the blind-validation runs throughout any stage of the model's training and development.

4.6 Summary:

In summary, deep learning artificial neural network models are used to replicate the numerical reservoir simulation outputs (pressure and CO$_2$ saturation). The general procedure and step-by-step workflow for data extraction, data preparation, feature engineering and ANN development and deployment and validation was briefly presented. The following are the four major stages of SPM development (see Figure 27. The general workflow utilized for SPM in this study):

1. A number of reservoir simulation scenarios are run based on the study's purpose.
2. The relevant static and dynamic data is retrieved and structured to produce a comprehensive spatio-temporal dataset, which is then used to construct the input data set for neural network training.
3. A number of neural networks are built and trained.
4. The neural network models are tested by running them through a series of blind scenarios.
There are two SPM networks in this study. One SPM focuses on predicting reservoir pressure due to CO₂ injection, while the other focuses on predicting CO₂ saturation. Each SPM has a methodical approach that differs from the others and the majority of the neural network characteristics and inputs are the same in both variants. These two SPMs are built in such a way that they can communicate with one another. In the next chapter, results and discussion of the predicted outputs will be presented and comparison with the numerical reservoir simulation outputs for both the pressure and CO₂ saturation will be provided.
CHAPTER 5: RESULTS AND DISCUSSIONS

5.1 Review:
The objective of this chapter is to see if the smart proxy modeling technique developed in this study can accurately reproduce the numerical simulation pressure and saturation outcomes. The geological heterogeneity in the porosity and permeability distribution would provide a good setting for determining the predictive power of this study. As discussed in the previous chapter, once the training phase of the neural network is ended, the model will be undergone a series of blind data that has been never seen by the model before. There are an overall 46 geological realizations developed in this study that six of them were kept as blind realizations at the beginning of the study. The remaining were included in the training phase of the smart proxy model. A list of the blind realizations as well as training realizations that are shown in this chapter is provided in Table 2. A list of the presented training and blind realizations in this Chapter. The numbers associated with the naming of each realization is representative of the possibility or frequency distribution of the porosity/permeability distribution. For example, a P25 case means that this particular model was placed in the lower 25 percentage frequency distribution of the entire porosity/permeability ranges. The second number indicates different distribution of the model but having the same min and max of the porosity and permeability values. A certain criterion was followed to select those blind cases that would not only cover the min and max values of the entire porosity and permeability domain of all the realizations, but also placed fairly in different frequency distributions rankings. The Due to the fact that there is 30 layers in each simulation run, and therefore, extent of the training results for Pressure as well as CO\textsubscript{2} models, only three training realizations are presented in this chapter, and the remaining the training results are provided in the Appendix section for brevity purposes.

The simulation run was intended for a duration of ten years (from 2020 to 2030), with a total of ten yearly timesteps, as detailed in Chapter 3. Four timesteps (01-01-2021, 01-01-25, 01-01-2030, and 01-01-2100) were chosen to assess the model's performance out of all of them. Due to the fact that there is 30 layers in each simulation run, and therefore, extent of the blind results, the selected three blind results are only shown in two time-steps for the Pressure as well as CO\textsubscript{2} and the remaining blind results are provided in the Appendix section for brevity purposes.
The ANN was constructed in three layers, an input layer, a hidden layer, and an output layer. The input layer for the pressure model contains 296 selected parameters or features. These features included static and dynamic properties relevant to the scope of the smart proxy models. The hidden layer has 5,000 nodes and there was one output in the output layer. A list of the selected parameters that were used for one of the runs is presented in Figure Figure 28. A list of selected features as input to the neural network a particular SPM mode

<table>
<thead>
<tr>
<th>Realization Number</th>
<th>Realization Type</th>
<th>Where to find Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>P25-3</td>
<td>Train</td>
<td>This chapter</td>
</tr>
<tr>
<td>P25-5</td>
<td>Blind</td>
<td>This chapter</td>
</tr>
<tr>
<td>P25-15</td>
<td>Blind</td>
<td>Appendix</td>
</tr>
<tr>
<td>P50-2</td>
<td>Train</td>
<td>Appendix</td>
</tr>
<tr>
<td><strong>P50-6</strong></td>
<td><strong>Blind</strong></td>
<td><strong>This chapter</strong></td>
</tr>
<tr>
<td>P50-14</td>
<td>Blind</td>
<td>Appendix</td>
</tr>
<tr>
<td>P75-1</td>
<td>Blind</td>
<td>Appendix</td>
</tr>
<tr>
<td>P75-8</td>
<td>Train</td>
<td>Appendix</td>
</tr>
<tr>
<td><strong>P75-14</strong></td>
<td><strong>Blind</strong></td>
<td><strong>This chapter</strong></td>
</tr>
</tbody>
</table>
5.2 Pressure Results:
The pressure results for selected Trains and Blind realization are presented in this section. It must be noted that the first two layers are shale barriers, and the remaining 28 layers of the reservoir are the actual reservoir layers.

5.2.1 Results of Train Runs:
The results are presented for two selected timesteps [01-01-2021 and 01-01-2030] and for each layer in Figures 29 through 44 for one of the selected training realizations (P25-3). The remaining two timesteps [01-01-2025 and 01-01-2100] are provided as Appendix. In each triple plot, the leftmost plot shows the numerical reservoir simulation actual results, the middle plot shows the results predicted by the Smart Proxy Model developed in this study and the rightmost plot shows the percentage of the error calculated from 0 to 0.1 percent. The training results demonstrates that the numerical simulation model and the smart proxy model are good fit. Between the two models, the absolute average inaccuracy was less than 0.1%. Some of the layers yielded exceptionally good results, while some layers were satisfactorily good. It must be mentioned that the Layers #1 and #2 are excluded from the results as they act as seal for the saline formation. The actual is the results from the Numerical Reservoir Simulation (CMG) and the Predicted is the result from the Smart Proxy Modeling (SPM).
P25-3 at time step 01-01-2021
Figure 29. Training pressure model P25-3 results for timestep 01-01-2021
Figure 30. Training pressure model P25-3 results for timestep 01-01-2021
Figure 31. Training pressure model P25-3 results for timestep 01-01-2021
Figure 32. Training pressure model P25-3 results for timestep 01-01-2021
Figure 33. Training pressure model P25-3 results for timestep 01-01-2021
Figure 34. Training pressure model P25-3 results for timestep 01-01-2021
Figure 35. Training pressure model P25-3 results for timestep 01-01-202
P25-3 at time step 01-01-2030 (End of Injection)
Figure 36. Training pressure model P25-3 results for timestep 01-01-2030
Figure 37. Training pressure model P25-3 results for timestep 01-01-2030
Figure 38. Training pressure model P25-3 results for timestep 01-01-2030
Figure 39. Training pressure model P25-3 results for timestep 01-01-2030
Figure 40. Training pressure model P25-3 results for timestep 01-01-2030
Figure 41. Training pressure model P25-3 results for timestep 01-01-2030
Figure 42. Training pressure model P25-3 results for timestep 01-01-2030
5.2.2 Results of Blind Runs:

Each smart proxy model was tested on a set of blind data sets. Blind data sets for the smart proxy model were static and dynamic reservoir data from years 2021, 2025, 2030 (end of injection) and 2100. The results are presented for two selected timesteps and for each layer in Figures 45 through 92 for the six blind realizations (P25-5, P50-6, and P75-14). As mentioned in the summary section, the remaining two timesteps for the training cases as well as the remaining blind realization results are provided in the Appendix section. It should be noted that for the timestep 01-01-2100 which is 70 years after the injection has been stopped, the CO$_2$ plume which was formed by the end of the injection period has now reached an equilibrium and dissolved into the water phase throughout the reservoir and that is the reason there is no CO$_2$ observed as a single separate liquid phase in the model at that time. The blind results demonstrates that the numerical simulation model and the smart proxy model are good fit. Between the two models, the absolute average inaccuracy was less than 0.1%. Some of the layers yielded exceptionally good results, while some layers were satisfactorily good.

P25-5 at time step 01-01-2021
Figure 43. Blind pressure model P25-5 results for timestep 01-01-2021
Figure 44. Blind pressure model P25-5 results for timestep 01-01-2021
Figure 45. Blind pressure model P25-5 results for timestep 01-01-2021
Figure 46. Blind pressure model P25-5 results for timestep 01-01-2021
Figure 47. Blind pressure model P25-5 results for timestep 01-01-2021
Figure 48. Blind pressure model P25-5 results for timestep 01-01-2021
Figure 49. Blind pressure model P25-5 results for timestep 01-01-2021
P25-5 at time step 01-01-2030 (End of injection)
Figure 50. Blind pressure model P25-5 results for timestep 01-01-2030
Figure 51. Blind pressure model P25-5 results for timestep 01-01-2030
Figure 52. Blind pressure model P25-5 results for timestep 01-01-2030
Figure 53. Blind pressure model P25-5 results for timestep 01-01-2030
Figure 54. Blind pressure model P25-5 results for timestep 01-01-2030
Figure 55. Blind pressure model P25-5 results for timestep 01-01-2030
Figure 56. Blind pressure model P25-5 results for timestep 01-01-2030
P50-6 at time step 01-01-2021
Figure 57. Blind pressure model P50-6 results for timestep 01-01-2021
Figure 58. Blind pressure model P50-6 results for timestep 01-01-2021
Figure 59. Blind pressure model P50-6 results for timestep 01-01-2021
Figure 60. Blind pressure model P50-6 results for timestep 01-01-2021
Figure 61. Blind pressure model P50-6 results for timestep 01-01-2021
Figure 62. Blind pressure model P50-6 results for timestep 01-01-2021
Figure 63. Blind pressure model P50-6 results for timestep 01-01-2021
P50-6 at time step 01-01-2030 (End of injection)
Figure 64. Blind pressure model P50-6 results for timestep 01-01-2030
Figure 65. Blind pressure model P50-6 results for timestep 01-01-2030
Figure 66. Blind pressure model P50-6 results for timestep 01-01-2030
Figure 67. Blind pressure model P50-6 results for timestep 01-01-2030
Figure 68. Blind pressure model P50-6 results for timestep 01-01-2030
Figure 69. Blind pressure model P50-6 results for timestep 01-01-2030
Figure 70. Blind pressure model P50-6 results for timestep 01-01-2030
P75-14 at time step 01-01-2021
Figure 71. Blind pressure model P75-14 results for timestep 01-01-2021
Figure 72. Blind pressure model P75-14 results for timestep 01-01-2021
Figure 73. Blind pressure model P75-14 results for timestep 01-01-2021
Figure 74. Blind pressure model P75-14 results for timestep 01-01-2021
Figure 75. Blind pressure model P75-14 results for timestep 01-01-2021
Figure 76. Blind pressure model P75-14 results for timestep 01-01-2021
Figure 77. Blind pressure model P75-14 results for timestep 01-01-2021
P75-14 at time step 01-01-2030 (End of injection)
Figure 78. Blind pressure model P75-14 results for timestep 01-01-2030
Figure 79. Blind pressure model P75-14 results for timestep 01-01-2030
Figure 80. Blind pressure model P75-14 results for timestep 01-01-2030
Figure 81. Blind pressure model P75-14 results for timestep 01-01-2030
Figure 82. Blind pressure model P75-14 results for timestep 01-01-2030
Figure 83. Blind pressure model P75-14 results for timestep 01-01-2030
Figure 84. Blind pressure model P75-14 results for timestep 01-01-2030
The reservoir pressure results of blind cases show that the Smart Proxy Model is capable of reliably replicating the outcomes of numerical reservoir simulation across all layers and several time steps both within the injection period as well as post-injection period. This means that the ANN algorithm has been successfully trained and taught to predict pressure patterns in several brand-new geological runs (blind runs). Some of the layers delivered exceptional results, while others worked well. Among the blind runs, some exhibit reasonably good match while some exhibit great results.

5.3 CO₂ Saturation Results:
The same neural network topology and architecture as in the CO₂ case was used in the construction of the ANN for saturation. The ANN was constructed in three layers, an input layer, a hidden layer, and an output layer (CO₂ saturation). The input layer for the CO₂ model contains about the same number of input features as the pressure models, with the exception that there exists a scenario in which the number of inputs were reduced to 54 selected parameters or features in order to optimize the number of records with the number of columns for the CO₂ saturation.

For the CO₂ models, many different approaches were investigated, 3 of those approaches are presented in this study as scenarios # 1, 2, and 3. In the following section, these scenarios are briefly introduced, and their respective results are presented. In all of these scenarios, CO₂ plume is defined as any cell with CO₂ values more than 1%.

Due to dynamic nature of the saturation changes in the plume, it was decided to build additional smart proxy models that can utilize the actual CO₂ values as an input from the numerical reservoir simulation at the previous timestep (t-1) for prediction and deployment on the current timestep (t) (scenario #2). This would provide the smart proxy models with advantages of having extra information about the previous timesteps. It is worth mentioning that among the 54 inputs that were included in the base CO₂ model, there were 12 features related to the effective permeability. As we know, the effective permeability is a function of saturation of the mobile phase, and by incorporating such information, the smart proxy can pinpoint the relative saturation region of the target cell saturation value.

Since the results from the scenario # 2 were satisfactory, in another scenario (# scenario #3), a different smart proxy model was developed that can utilize the predicted CO₂ values at t-1 timestep as an input for prediction and deployment on the current timestep. The key difference between this scenario and scenario #2 is that in this one, t-1 CO₂ values are predicted rather than actual values from the numerical simulator. This approach is also called cascading.

Furthermore, the deployment technique is modified in that the benefits of having extra information from the effective permeability inputs are cancelled out by the error accumulation caused by using
the first timestep moving forward when using the predicted t-1 CO₂ values. For example, a difference of only 5% in the CO₂ saturation can have a huge impact on the effective permeability curve and therefore, totally misguiding the smart proxy model (confuse the neural network).

**Scenario #1:** In this scenario, the CO₂ model was trained and deployed with the following conditions:

- Train and deploy with CO₂ saturation plume values for the end of injection timestep

**Scenario #2:** In this scenario, the CO₂ model was trained and deployed with the following conditions:

- Train with CO₂ saturation plume values plus two injection wells surrounding cells using 10 yearly timesteps data (2021 – 2030).
- Added (CO₂ t-1) and used that value to calculate effective permeability.
- Deploy the model on one timestep (2030) of the blind realizations on only the plume.
- In this scenario the CO₂ t-1 used is the actual value from CMG.

**Scenario #3:** In this scenario, the CO₂ model was trained and deployed with the following conditions:

- Train with CO₂ saturation plume values plus two surrounding cells using 10 yearly timesteps data (2021 – 2030).
- Added (CO₂ t-1) and used that value to calculate effective permeability.
- Deploy the model on 10 timestep (2021-2030) blind realizations (cascading) on all the cells
- In this scenario the CO₂ t-1 used is predicted value.

### 5.3.2 Results of CO₂ Models:

The CO₂ results for selected Trains and Blind realization are presented in this section. As previously mentioned, the first two layers are shale barriers, and the remaining 28 layers of the reservoir are the actual reservoir layers.

#### 5.3.2.1 Results of Train Runs:

The results are presented for the considered timesteps and for each layer for one of the selected training realizations (P25-3). The remaining training realizations are provided as Appendix.

In each triple plot, the leftmost plot shows the numerical reservoir simulation actual results, the middle plot shows the results predicted by the Smart Proxy Model developed in this study and the rightmost plot shows the error. In Scenarios #3 where there is cascading deployment, there error plot shows the binary classification where there exists CO₂ or not, based on a threshold value of 10% in the plume. For brevity purposes, only selected timesteps are presented for several layers.
Scenario #1:
P25-3 at time step 01-01-2030 (End of injection)
Figure 85. Scenario 1: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 86. Scenario 1: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 87. Scenario 1: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 88. Scenario 1: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 89. Scenario 1: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 90. Scenario 1: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 91. Scenario 1: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Scenario #2:
Realization P25-3 end of injection 01-01-2030
Figure 92. Scenario 2: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 93. Scenario 2: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 94. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030
Figure 95. Scenario 2: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 96. Scenario 2: Train CO$_2$ model P25-3 results for timestep end of injection 01-01-2030
Figure 97. Scenario 2: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Figure 98. Scenario 2: Train CO₂ model P25-3 results for timestep end of injection 01-01-2030
Scenario #3:
Realization P25-3 per layer in a 10-year period
Layer #5

01-01-2021

p25-3: Layer-5 [R2= 0.76]

01-01-2025

p25-3: Layer-5 [R2= -3.64]

01-01-2030

p25-3: Layer-5 [R2= -3.02]

Figure 99. Scenario 3: Layer 5: Train CO₂ model P25-3 results
Layer #10

01-01-2021
p25-3: Layer-10 [R2 = 0.31]

01-01-2025
p25-3: Layer-10 [R2 = -2.56]

01-01-2030
p25-3: Layer-10 [R2 = -2.67]

Figure 100. Scenario 3: Layer 10; Train CO₂ model P25-3 results
Figure 101. Scenario 3: Layer 15: Train CO$_2$ model P25-3 results
Figure 102. Scenario 3: Layer 22: Train CO$_2$ model P25-3 results
Layer #30

01-01-2021

p25-3: Layer-30 [R2= 0.18]

CMG Model, 10 % Threshold

01-01-2025

p25-3: Layer-30 [R2= -180.57]

CMG Model, 10 % Threshold

01-01-2030

p25-3: Layer-30 [R2= -70.8]

CMG Model, 10 % Threshold

Figure 103. Scenario 3: Layer 30: Train CO₂ model P25-3 results
5.3.2.1 Results of Blind Runs:

Scenario #1:
P25-5 at time step 01-01-2030 (End of injection)
Figure 104. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 105. Scenario 1: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
Figure 106. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 107. Scenario 1: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
Figure 108. Scenario 1: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 109. Scenario 1: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
Figure 110. Scenario 1: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
P50-6 at time step 01-01-2030 (End of injection)
Figure 111. Scenario 1: Blind CO₂ model P50-6 results for timestep end of injection 01-01-2030
Figure 112. Scenario 1: Blind CO₂ model P50-6 results for timestep end of injection 01-01-2030
Figure 113. Scenario 1: Blind CO$_2$ model P50-6 results for timestep end of injection 01-01-2030
Figure 114. Scenario 1: Blind CO₂ model P50-6 results for timestep end of injection 01-01-2030
Figure 115. Scenario 1: Blind CO₂ model P50-6 results for timestep end of injection 01-01-2030
Figure 116. Scenario 1: Blind CO₂ model P50-6 results for timestep end of injection 01-01-2030
Figure 117. Scenario 1: Blind CO₂ model P50-6 results for timestep end of injection 01-01-2030
P75-14 at time step 01-01-2030 (End of injection)
Figure 118. Scenario 1: Blind CO₂ model P75-14 results for timestep end of injection 01-01-2030
Figure 119. Scenario 1: Blind CO₂ model P75-14 results for timestep end of injection 01-01-2030
Figure 120. Scenario 1: Blind CO₂ model P75-14 results for timestep end of injection 01-01-2030
Figure 121. Scenario 1: Blind CO₂ model P75-14 results for timestep 01-01-2030
Figure 122. Scenario 1: Blind CO2 model P75-14 results for timestep 01-01-2030
Figure 123. Scenario 1: Blind CO$_2$ model P75-14 results for timestep end of injection 01-01-2030
Figure 124. Scenario 1: Blind CO₂ model P75-14 results for timestep end of injection 01-01-2030
Scenario #2:
P25-5 at time step 01-01-2030 (End of injection)
Figure 125. Scenario 2: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 126. Scenario 2: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 127. Scenario 2: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 128. Scenario 2: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
Figure 129. Scenario 2: Blind CO\textsubscript{2} model P25-5 results for timestep end of injection 01-01-2030
Figure 130. Scenario 2: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
Figure 131. Scenario 2: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
P50-6 at time step 01-01-2030 (End of injection)
Figure 132. Scenario 2: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
Figure 133. Scenario 2: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
Figure 134. Scenario 2: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 135. Scenario 2: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
Figure 136. Scenario 2: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 137. Scenario 2: Blind CO$_2$ model P25-5 results for timestep end of injection 01-01-2030
Figure 138. Scenario 2: Blind CO₂ model P25-5 results for timestep end of injection 01-01-2030
P75-14 at time step 01-01-2030 (End of injection)
Figure 139. Scenario 2: Blind CO₂ model P75-14 results for timestep end of injection 01-01-2030
Figure 140. Scenario 2: Blind CO$_2$ model P75-14 results for timestep end of injection 01-01-2030
Figure 141. Scenario 2: Blind CO₂ model P75-14 results for timestep end of injection 01-01-2030
Figure 142. Scenario 2: Blind CO₂ model P75-14 results for timestep end of injection 01-01-2030
Figure 143. Scenario 2: Blind CO$_2$ model P75-14 results for timestep end of injection 01-01-2030
Figure 144. Scenario 2: Blind CO$_2$ model P75-14 results for timestep end of injection 01-01-2030
Figure 145. Scenario 2: Blind CO$_2$ model P75-14 results for timestep end of injection 01-01-2030
Scenario #3
Realization P25-5
Layer #5

01-01-2021
p25-5: Layer-5 [R2= 0.45]

01-01-2025
p25-5: Layer-5 [R2= -0.19]

01-01-2030
p25-5: Layer-5 [R2= -0.37]

Figure 146. Scenario 3: Layer 5: Blind CO₂ model P25-5 results
Layer #10

Figure 147. Scenario 3: Layer 10: Blind CO$_2$ model P25-5 results
Layer #15

01-01-2021

p25-5: Layer-15  [R²= 0.0]

01-01-2025

p25-5: Layer-15  [R²= -283.94]

01-01-2030

p25-5: Layer-15  [R²= -31.01]

Figure 148. Scenario 3: Layer 15: Blind CO₂ model P25-5 results
Figure 149. Scenario 3: Layer 22: Blind CO$_2$ model P25-5 results
Figure 150. Scenario 3: Layer 30: Blind CO$_2$ model P25-5 results
Realization P50-6
Layer #5

01-01-2021
p50-6: Layer-5 [R2= 0.6]

01-01-2025
p50-6: Layer-5 [R2= -3.22]

01-01-2030
p50-6: Layer-5 [R2= -2.52]

Figure 151. Scenario 3: Layer 5: Blind CO₂ model P50-6 results
Layer #10

Figure 152. Scenario 3: Layer 10: Blind CO₂ model P50-6 results
Figure 153. Scenario 3: Layer 15: Blind CO₂ model P50-6 results
Figure 154. Scenario 3: Layer 22: Blind CO$_2$ model P50-6 results
Layer #30

Figure 155. Scenario 3: Layer 30: Blind CO$_2$ model P50-6 results
Realization P75-14
Layer #5

01-01-2021
p75-14: Layer-5 [R^2 = 0.36]

01-01-2025
p75-14: Layer-5 [R^2 = -7.24]

01-01-2030
p75-14: Layer-5 [R^2 = -4.95]

Figure 156. Scenario 3: Layer 5: Blind CO\textsubscript{2} model P75-14 results
Figure 157. Scenario 3: Layer 10: Blind CO$_2$ model P75-14 results
Layer #15

01-01-2021
p75-14: Layer-15 [R2= 0.46]

01-01-2025
p75-14: Layer-15 [R2= -16.8]

01-01-2030
p75-14: Layer-15 [R2= -12.23]

Figure 158. Scenario 3: Layer 15: Blind CO₂ model P75-14 results
Layer #22

01-01-2021
p75-14: Layer-22 [R² = 0.71]

01-01-2025
p75-14: Layer-22 [R² = -4.42]

01-01-2030
p75-14: Layer-22 [R² = -1.54]

Figure 159. Scenario 3: Layer 22: Blind CO₂ model P75-14 results
Layer #30

01-01-2021
p75-14: Layer-30 [R2 = -2.17]

01-01-2025
p75-14: Layer-30 [R2 = -1254.84]

01-01-2030
p75-14: Layer-30 [R2 = -1414.68]

Figure 160. Scenario 3: Layer 30: Blind CO2 model P75-14 results
CHAPTER 6: SUMMARY, CONCLUSION AND FUTURE RESEARCH

6.1 Summary:
Any CO₂ sequestration project must monitor the breadth of the CO₂ plume as well as pressure distribution across the reservoir at any location. During injection, the Smart Proxy Model produced a CO₂ saturation or plume that varied spatially and temporally. It's worth noting that although the pressure smart proxy model could replicate the outcome of the numerical reservoir simulation to a high degree of accuracy, however, the CO₂ model struggled to replicate CO₂ at early time steps due to a lack of data on CO₂ specifically originating from the first timestep. As a result, the Smart Proxy's accuracy throughout the initial steps was not as good as the late timesteps. This, along with the limitation of number of training records available to train the neural network, resulted in accumulation of the error and therefore, poor prediction across forthcoming timesteps in the saturation model.

6.2 Conclusion:
The main conclusions drawn from this study can be summarized as below:

• Several smart proxy models were constructed using artificial intelligence and machine learning technologies to replicate pressure and extend of the CO₂ plume from the numerical reservoir simulation.
• About 300 static and dynamic features were generated and engineered to account for the inputs to the smart proxy models.
• The Pressure model results showed that this technology can replicate the CMG model with high accuracy.
• For the CO₂ Model, in general, the CO₂ plumes were significantly small and that affected the amount of data that can be used to train the smart proxy model. Therefore, the predictions were not as accurate as the pressure model in static scenario.
• Dynamic scenarios of CO₂ model were tested, and the results seems promising (Also the cascading dynamic scenario had been previously tested by our research group, and the results were good).
• As mentioned, the objective of the study was to test smart proxy technology on a brand-new dataset to replicate pressure and CO₂ distribution results of the numerical simulation. Such dataset even though was geologically unrealistic, was used as input to the smart proxy models, and several approaches were systematically developed and investigated to improve the quality of the data fed into the neural network. However, in majority of these approaches, the smart proxy model showed its capabilities in a reasonable manner. It is believed that if the dataset was realistic in the first place, the accuracy of the smart proxy model specially in the CO₂ predictions were significantly higher than what was presented in this study.
• It must be emphasized again that the objective of this study was to replicate the extent of the CO₂ plume (and not necessarily the grid-based values across the reservoir grid cells) which was different
than the objective set for the Pressure model. Therefore, if one were to look at the CO₂ results with this objective in mind, then the results appear to be satisfactory given the small number of records available to train the neural network. As a matter of fact, the Smart Proxy Model exceeded the initial expectations.

- The design of the input geological parameters to the smart proxy must make sense in terms of CO₂ sequestration objectives. We were aware of such shortcomings in the dataset, however, we wanted to push the limits of the smart proxy technology when it comes to the amount of data provided as main features to the model.

- In addition to the above-mentioned conclusions, this study also provides us with an important conformation about the size and quality of the data fed into the neural network and smart proxy models.

- The main challenge in building the CO₂ model was dealing with significantly small number of cells with presence of CO₂ (1% CO₂) and therefore the extent of the CO₂ plumes. This lowered the amount of usable training data from around 53 million records to around 0.25 million records compared to pressure data.

- The CO₂ model results of blind cases show that the Smart Proxy Model in scenario #1, that is when the model was both trained and deployed only on the plume, the results were satisfactory at the timestep which it was tested (end of injection).

- In Scenario #2, where the model was trained using 10 yearly-timesteps [injection period from 2021 to 2030], and deployed on the end of injection timestep, the results were better than expected since we used the actual t-1 CO₂ values from the numerical reservoir.

- In Scenario #3, where the model was trained using 10 yearly-timesteps [injection period from 2021 to 2030], and deployed on all the 10 yearly-timesteps, the results were not as expected since we used the predicted t-1 CO₂ values from the cascading deployment.

- Among the three scenarios, the cascading scenario (#3) is more practical, and would most likely give better results if the size of the plume (number of cells affected with >1% CO₂ values) was significantly larger. This approach has been tested in our research group on similar projects and the results were comparable to the numerical reservoir simulation CO₂ outputs.

6.3 Future Work Recommendation:
This study yielded the following recommendations:

- Neural network work best when there is enough data to be trained with. The main input data (porosity, and permeability) distribution used in this project were estimated based on unrealistic geological models. Therefore, the size of the CO₂ plumes controlled by these main inputs was limited. It is recommended to include more realistic input data, especially the main porosity and permeability distribution.

- Training a neural network when applied to a highly complex and specialized domain such as reservoir engineering requires teaching the neural network the essence of fluid flow in porous media, and all the interactions between injection and production at every grid cells. Although we made our best efforts to come up with new features to capture such behavior and relationship, it might be the fact that there are more important features that if they are defined, could help the neural network better, even with this small amount of data in this study. Therefore, we recommend future projects to focus more on creating advanced features and exploring their impact on the accuracy of the neural network.
CHAPTER 7: APPENDIX

---------------------------------------- Train Results for Pressure----------------------------------------

Appendix 7.1: Result for Training Run P25-3 pressures Second Time Step 01-01-2025 for Several Selected reservoir Layers:

For consistency purposes, the selected layers are kept constants for all the models and timesteps in the Appendix section and the layers are: #5, #10, #15, #22, and #30. For cases where the timesteps are varied, the selected timesteps are also kept constant for consistency purposes and are 01-01-2021, 01-01-2025, 01-01-2030, and 01-01-2100.
Appendix 7.2: Result for Training Run P25-3 pressures for Forth Time Step 01-01-2100 for Several Selected reservoir Layers:
Appendix 7.3: Result for Training Run P50-2 pressures First Time Step 01-01-2021 for Several Selected reservoir Layers:
Appendix 7.4: Result for Training Run P50-2 pressures Second Time Step 01-01-2025 for Several Selected reservoir Layers:
Appendix 7.5: Result for **Training Run P50-2** pressures Third Time Step *01-01-2030* for Several Selected reservoir Layers:
Appendix 7.6: Result for **Training Run P50-2** pressures Forth Time Step **01-01-2100** for Several Selected reservoir Layers:
Appendix 7.7: Result for **Training Run P75-8 pressures First Time Step 01-01-2021** for Several Selected reservoir Layers:
Appendix 7.8: Result for Training Run P75-8 pressures Second Time Step 01-01-2025 for Several Selected reservoir Layers:
Appendix 7.9: Result for Training Run P75-8 pressures Third Time Step 01-01-2030 for Several Selected reservoir Layers:
Appendix 7.10: Result for Training Run P75-8 pressures Forth Time Step 01-01-2100 for Several Selected reservoir Layers:
Blind Results for Pressure

Appendix 7.11: Result for Blind Run P25-5 pressures Second Time Step 01-01-2025 for Several Selected reservoir Layers:
Appendix 7.12: Result for **Blind Run P25-5** pressures Forth Time Step **01-01-2100** for Several Selected reservoir Layers:
Appendix 7.13: Result for **Blind Run P25-15** pressures First Time Step **01-01-2021** for Several Selected reservoir Layers:
Appendix 7.14: Result for **Blind Run P25-15** pressures Second Time Step **01-01-2025** for Several Selected reservoir Layers:
Appendix 7.15: Result for **Blind Run P25-15** pressures Third Time Step **01-01-2030** for Several Selected reservoir Layers:
Appendix 7.16: Result for **Blind Run P25-15** pressures Forth Time Step **01-01-2100** for Several Selected reservoir Layers:
Appendix 7.17: Result for Blind Run P50-6 pressures Second Time Step 01-01-2025 for Several Selected reservoir Layers:
Appendix 7.18: Result for **Blind Run P50-6** pressures Forth Time Step 01-01-2100 for Several Selected reservoir Layers:
Appendix 7.19: Result for **Blind Run P50-14** pressures First Time Step 01-01-2021 for Several Selected reservoir Layers:
Appendix 7.20: Result for Blind Run P50-14 pressures Second Time Step 01-01-2025 for Several Selected reservoir Layers:
Appendix 7.21: Result for **Blind Run P50-14** pressures Third Time Step **01-01-2030** for Several Selected reservoir Layers:
Appendix 7.22: Result for **Blind Run P50-14** pressures Forth Time Step **01-01-2100** for Several Selected reservoir Layers:
Appendix 7.23: Result for **Blind Run P75-1** pressures First Time Step 01-01-2021 for Several Selected reservoir Layers:
Appendix 7.24: Result for **Blind Run P75-1** pressures Second Time Step **01-01-2025** for Several Selected reservoir Layers:
Appendix 7.25: Result for Blind Run P75-1 pressures Third Time Step 01-01-2030 for Several Selected reservoir Layers:
Appendix 7.26: Result for **Blind Run P75-1** pressures Forth Time Step **01-01-2100** for Several Selected reservoir Layers:
Appendix 7.2: Result for **Blind Run P75-14** pressures Second Time Step **01-01-2025** for Several Selected reservoir Layers:
Appendix 7.28: Result for Blind Run P75-14 pressures Forth Time Step 01-01-2100 for Several Selected reservoir Layers:
Appendix 7.29: Result for Training Run P50-2 CO₂ Saturation for Several Selected reservoir Layers and Timesteps:

**Scenario #1**

![Graphs showing Actual, Predicted, and Error for CO₂ Saturation for different layers (Layer 5, Layer 10, Layer 15).]
Scenario #2

p50-2: Layer-5 [R2= 0.99]
Scenario #3

Layer #5

01-01-2021

p50-2: Layer-5 [R^2 = 0.44]

01-01-2025

p50-2: Layer-5 [R^2 = -0.41]
Layer #10

01-01-2030

p50-2: Layer-5 [R2 = -0.91]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2021

p50-2: Layer-10 [R2 = 0.09]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2025

p50-2: Layer-10 [R2 = -40.13]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold
Layer #15

01-01-2030
p50-2: Layer-10 [R² = -43.51]

01-01-2021
p50-2: Layer-15 [R² = 0.86]

01-01-2025
p50-2: Layer-15 [R² = -0.71]
Layer #22

01-01-2030
p50-2: Layer-15  [R2= -0.36]

01-01-2021
p50-2: Layer-22  [R2= 0.69]

01-01-2025
p50-2: Layer-22  [R2= -2.4]
Layer #30

01-01-2030
p50-2: Layer-22 [R2= -1.53]

01-01-2021
p50-2: Layer-30 [R2= 0.46]

01-01-2025
p50-2: Layer-30 [R2= -51.09]
Appendix 7.3: Result for Training Run P75-8 CO₂ Saturation for Several Selected reservoir Layers:

**Scenario #1**
Scenario #2

p75-8: Layer-5 \([R^2 = 0.99]\)

- CMG Model, 10 % Threshold
- Smart Proxy, 10 % Threshold
- Binary Error, 10 % Threshold

p75-8: Layer-10 \([R^2 = 1.0]\)

- CMG Model, 10 % Threshold
- Smart Proxy, 10 % Threshold
- Binary Error, 10 % Threshold

p75-8: Layer-15 \([R^2 = 0.99]\)

- CMG Model, 10 % Threshold
- Smart Proxy, 10 % Threshold
- Binary Error, 10 % Threshold
Scenario #3

Layer #5

01-01-2021

p75-8: Layer-22 [R2= 0.99]

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

CMG Model, 10 % Threshold

p75-8: Layer-30 [R2= 0.94]

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

CMG Model, 10 % Threshold

p75-8: Layer-5 [R2= 0.31]

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

CMG Model, 10 % Threshold
Layer #10
01-01-2025
p75-8: Layer-10 [R² = -10.17]

01-01-2030
p75-8: Layer-10 [R² = -8.59]

Layer #15

01-01-2021
p75-8: Layer-15 [R² = 0.85]

01-01-2025
Layer #22
Layer #30

01-01-2025
p75-8: Layer-22 [R2= -4.89]

01-01-2030
p75-8: Layer-22 [R2= -3.25]

01-01-2021
p75-8: Layer-30 [R2= 0.29]
Appendix 7.31: Result for **Blind Run P25-5** CO₂ Saturation for Several Selected reservoir Layers and Timesteps:

**Scenario #1**
Scenario #2

p25-5: Layer-5  [R² = 0.99]

p25-5: Layer-10  [R² = 0.98]
Scenario #3

Layer #5

01-01-2021
p25-5: Layer-5 [R2= 0.45]

01-01-2025
p25-5: Layer-5 [R2= -0.19]

01-01-2030
p25-5: Layer-5 [R2= -0.37]
Layer #10

01-01-2021
p25-5: Layer-10 [R2= 0.05]

01-01-2025
p25-5: Layer-10 [R2= -4.98]

01-01-2030
p25-5: Layer-10 [R2= -2.88]
Layer #15

01-01-2021
p25-5: Layer-15 [R2= 0.0]

01-01-2025
p25-5: Layer-15 [R2= -283.94]

01-01-2030
p25-5: Layer-15 [R2= -31.01]
Layer #22

01-01-2021
p25-5: Layer-22 [R2= 0.61]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2025
p25-5: Layer-22 [R2= -0.07]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2030
p25-5: Layer-22 [R2= -0.11]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold
Layer #30

01-01-2021
p25-5: Layer-30 [R2= 0.14]

01-01-2025
p25-5: Layer-30 [R2= -9.81]

01-01-2030
p25-5: Layer-30 [R2= -8.57]
Appendix 7.32: Result for **Blind Run P25-15** CO₂ Saturation for Several Selected reservoir Layers and Timesteps:

**Scenario #1**

![Diagram of actual, predicted, and error values for CO₂ saturation in Layers 5, 10, and 15.](image-url)

- **Layer 5**: Actual vs. Predicted vs. Error for CO₂ saturation.
- **Layer 10**: Actual vs. Predicted vs. Error for CO₂ saturation.
- **Layer 15**: Actual vs. Predicted vs. Error for CO₂ saturation.
Scenario #2

p25-15: Layer-5 [R^2= 0.99]
Scenario #3

Layer #5

01-01-2021
p25-15: Layer-5 [R2= 0.34]

01-01-2025
p25-15: Layer-5 [R2= 0.38]
Layer #10
Layer #15
Layer #22

01-01-2030

p25-15: Layer-15  [R2= -1.45]

01-01-2021

p25-15: Layer-22  [R2= 0.66]

01-01-2025

p25-15: Layer-22  [R2= -2.11]
Layer #30

01-01-2030
p25-15: Layer-22 [R2= -1.19]

01-01-2021
p25-15: Layer-30 [R2= 0.06]

01-01-2025
p25-15: Layer-30 [R2= -32.18]
Appendix 7.33: Result for **Blind Run P50-14** CO$_2$ Saturation for Several Selected reservoir Layers and Timesteps:

**Scenario #1**
Scenario #2

p50-14: Layer-5  [R2= 0.99]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

p50-14: Layer-10  [R2= 1.0]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold
Scenario #3

Layer #5

01-01-2021
p50-14: Layer-5 [R2 = 0.32]

01-01-2025
p50-14: Layer-5 [R2 = -0.32]

01-01-2030
p50-14: Layer-5 [R2 = -0.15]
Layer #10

01-01-2021
p50-14: Layer-10 [R2= 0.79]

01-01-2025
p50-14: Layer-10 [R2= -3.11]

01-01-2030
p50-14: Layer-10 [R2= -3.72]
Layer #15

01-01-2021
p50-14: Layer-15 [R2 = -0.26]

CMG Model, 10 % Threshold

01-01-2025
p50-14: Layer-15 [R2 = -1.18]

CMG Model, 10 % Threshold

01-01-2030
p50-14: Layer-15 [R2 = -0.67]

CMG Model, 10 % Threshold
Layer #22

01-01-2021
p50-14: Layer-22  [R2= 0.17]

01-01-2025
p50-14: Layer-22  [R2= -1.61]

01-01-2030
p50-14: Layer-22  [R2= -0.54]
Layer #30

01-01-2021
p50-14: Layer-30  [R2= 0.81]

CMG Model, 10 % Threshold
Smart Proxy, 10 % Threshold
Binary Error, 10 % Threshold

01-01-2025
p50-14: Layer-30  [R2= -4.88]

CMG Model, 10 % Threshold
Smart Proxy, 10 % Threshold
Binary Error, 10 % Threshold

01-01-2030
p50-14: Layer-30  [R2= -5.31]

CMG Model, 10 % Threshold
Smart Proxy, 10 % Threshold
Binary Error, 10 % Threshold

Appendix 7.35: Result for Blind Run P75-1 CO₂ Saturation for Several Selected reservoir Layers and Timesteps:
Scenario #1
Scenario #2

p75-1: Layer-5  [R2= 0.97]
Scenario #3

Layer #5

01-01-2021
p75-1: Layer-5 [R2= 0.1]

01-01-2025
p75-1: Layer-5 [R2= -5.78]
Layer #10

01-01-2030
p75-1: Layer-5 [R^2 = -5.89]

01-01-2021
p75-1: Layer-10 [R^2 = 0.05]

01-01-2025
p75-1: Layer-10 [R^2 = -2.28]
Layer #15

01-01-2030
p75-1: Layer-10 [R2= -2.05]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2021
p75-1: Layer-15 [R2= 0.38]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2025
p75-1: Layer-15 [R2= -22.83]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold
Layer #22

01-01-2030
p75-1: Layer-15 [R² = -21.13]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2021
p75-1: Layer-22 [R² = 0.66]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2025
p75-1: Layer-22 [R² = -2.1]
Layer #30

01-01-2030
p75-1: Layer-22 [R2 = -1.37]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2021
p75-1: Layer-30 [R2 = 0.15]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold

01-01-2025
p75-1: Layer-30 [R2 = -598.37]

CMG Model, 10 % Threshold

Smart Proxy, 10 % Threshold

Binary Error, 10 % Threshold
01-01-2030

p75-1: Layer-30 [R2 = -1021.09]
CHAPTER 7: REFERENCES


Alabboodi, Maher Jasim, "Dynamic Data-Driven Smart Proxy Modeling For Numerical Reservoir Simulation" (2021). Graduate Theses, Dissertations, and Problem Reports. 10235.


Mohaghegh, S. Smart proxy modeling for numerical reservoir simulations - big data analytic in e and p, 2015.


