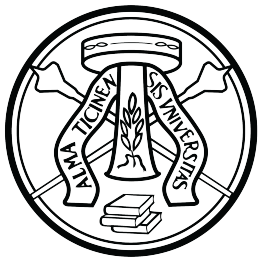

UNIVERSITÀ
DI PAVIA

UNIVERSITÀ DI PAVIA

FACULTY OF ENGINEERING

DEPARTMENT OF CIVIL ENGINEERING

MASTER'S DEGREE IN ENVIRONMENTAL ENGINEERING

MASTER THESIS

TITLE

Modeling CO_2 flow in Injection Wells for Storage Projects

Modellizzazione del Flusso di CO_2 in Pozzi di Iniezione per Progetti di Stoccaggio

Candidate: Hadis Yousefi

Supervisor: Prof. Sauro Manenti

Co-supervisor: Emanuele Vignati

A.Y. 2026/2025

Abstract

Carbon Capture and Storage (CCS) is one of the technology to mitigate the climate change, since greenhouse gases play a significant role in changes of the climate. CCS operations rely on the injection wells as the primary conduit for delivering CO_2 into underground storage formations, where it will be safely trapped in underground formations for geological timescales. The performance and integrity of these wells directly depend on robust injection strategies that account for the appropriate management of the pressure and temperature in the wellbore.

Therefore, monitoring the well is a key factor to ensure the reliability and safety of the entire CCS project, as excessive pressure may cause phase transition of the fluid and damage equipment or reduce injectivity. Effective monitoring of the CO_2 injection is essential for operational safety, regulatory compliance, and storage verification. However, traditional monitoring approaches face significant challenges, including harsh downhole environments that would compromise gauge reliability and limitations in conventional simulation tools.

This thesis investigates on the application of the usage of Machine Learning in order to enhance the Monitoring of the CCS projects, by focusing on having accurate prediction of flowing bottom hole pressure using surface measurements. In this work an evaluation of industry standard tools for multiphase simulation tools like Prosper ™ (Petroleum Expert) and OLGA™ (slb) has been done that demonstrates their respective limitations for CO_2 applications, particularly when it comes to capturing transient behavior and phase transitions of the fluid. Through analysis of the real time data, the study establishes that physics-based simulation can provide valuable training data gather from OLGA simulations for machine learning models when they have been properly calibrated against field measurements.

A comprehensive machine learning methodology has been developed, involving robust data preprocessing using Median Absolute Deviation (MAD) techniques and feature engineering to capture physical phenomena. and a comparison analysis of other algorithms. Among the tested algorithms, Random Forest demonstrated the highest predictive accuracy, surpassing Support Vector Regression, Linear Regression, and Gradient Boosting. Moreover, the implementation of an ensemble approach, integrating multiple algorithms, provided further improvements in the model's predictive performance.

The research extends beyond theoretical performance to address practical implementation challenges, documenting a successful real-time virtual sensing system. A lengthy operational evaluation confirmed its consistent and reliable performance.

The methodology and results offer a framework for implementation at other CCS operations,

with recommendations for data requirements, model selection, operational integration, and continuous improvement processes. The thesis concludes by identifying promising future directions, such as bottom hole pressure prediction, anomaly detection for well integrity monitoring, and integration with reservoir-scale models. This work advances CCS monitoring technology by establishing machine learning-based virtual sensing as a valuable complement to physical instrumentation, improving system redundancy and operational visibility.

Keywords: Carbon Capture and Storage, Machine Learning, Virtual Sensing, Support Vector Regression, Random Forest, CO_2 Monitoring, Ensemble Models

Acknowledgments

I would like to thank all the people who supported me in completing this Master's thesis.

First and foremost, I am deeply grateful to Eni S.p.A. for giving me the incredible opportunity of carrying out this research within the framework of the LM+ (Lauree Magistrali Plus) program. This collaboration between Eni and the University of Pavia has given me the chance to work on cutting-edge Carbon Capture and Storage technology that has been developed for real industrial applications. Access to operational data from the Ravenna CCS Phase 1 project and state-of-the-art simulation tools has been very valuable in the course of this work.

I would like to take this opportunity to thank my supervisor, Professor Sauro Manenti, for his technical support, constant encouragement, and valuable advice throughout this research experience. His expertise in process engineering and willingness to explore novel applications of machine learning in environmental engineering have been invaluable to the completion of this study.

I am especially thankful to my co-supervisors, Emanuele Vignati and Fabrizio Ursini, without whom this research would never have been possible. Their exceptional guidance, patience, and dedication have been nothing less than life-altering for my career and academic development. Emanuele has been an extraordinary mentor for the entire research process, providing phenomenal guidance in the machine learning aspect of this research and instructing me in state-of-the-art data science techniques with immense patience. His expertise in computational methods, algorithm development, and validation has been absolutely invaluable, and his ability to guide me through the subtleties of machine learning application in engineering has greatly enhanced the quality and rigor of this work. His willingness to always be available for consultation, troubleshooting, and guidance throughout every phase of this work has been nothing short of phenomenal. Fabrizio has provided excellent expertise in multiphase flow simulation and modelling techniques. His deep insight into CCS operations and advanced simulation software has been instrumental in developing the physics-based framework supporting this research. His guidance on the complexities of co2 injection systems and support with model calibration and validation have been essential in ensuring the technical validity of this work. They have both gone beyond the call of supervisory responsibilities, becoming mentor figures who have touched not only this thesis but my overall approach to engineering problem solving. Their collective knowledge and unwavering encouragement have made this challenging interdisciplinary research both feasible and pleasant.

I would also like to thank the University of Pavia, Faculty of Engineering, Department of Civil Engineering for the excellent academic atmosphere and the facilities necessary to carry out this research. The interdisciplinarity generated by the department has been

fundamental for managing the complex issues presented by co2 injection systems.

I would also like to thank the Eni technical teams working on the Ravenna CCS project for their collaboration and for their openness in sharing their operating experience. Their insight into the day-to-day issues of co2 injection operations has been invaluable in maintaining the relevance and applicability of this research.

I would also like to thank my classmates in the LM+ program for the stimulating discussions and supportive learning atmosphere that made my learning process more enjoyable along the way.

Finally, I would like to express my most sincere gratitude to my family and friends for their unwavering support, patience, and understanding as I completed this thesis. Their belief in me has been a steady source of motivation.

This is not only an academic achievement but also a step towards the advancement of sustainable technologies to combat climate change. I am honored to have been part of this important endeavor.

The work upon which this thesis is founded was conducted at Eni S.p.A. within the context of the LM + program (Lauree Magistrali Plus), a collaboration of Eni and the University of Pavia for the advancement of innovation in energy and environmental technologies.

Contents

List of Figures

Chapter 1

Introduction to Carbon Capture and Storage

1.1 Background of Carbon Capture and Storage

1.1.1 The Climate Imperative

Climate change stands as the most severe contemporary crisis that has widespread effects on ecological, social and economic systems throughout the world. The scientific findings from Berrang-Ford and his team show that human activities have created an undeniable global climate change. [3] The world has seen a surface temperature rise of about 1.5°C since the past hundred years because of industrial activities and fossil fuel burning and forest clearing.

The world experiences more extreme weather events and faster glacier melting which leads to higher sea levels and major environmental destruction because of this warming pattern. The effects reach past environmental issues because they endanger food security and water resources and human health and economic stability. Climate change demands urgent attention because its effects grow worse each day so we need both deep knowledge and fast action for prevention and adjustment measures.

1.1.2 Global Carbon Emissions Context and Trajectory

Human activities have created a dangerous pattern of rising greenhouse gas emissions which alters the natural composition of the atmosphere. The GOSAT-2 satellite observations reveal a continuous growth of atmospheric carbon dioxide which started at about 375 parts per million in 2005 and reached beyond 420 parts per million in 2023 indicating a 12 percent rise during the last twenty years. The yearly growth rate showed variations between 1.5 and 3.0 ppm annually while 2023 experienced a major increase. This trend continued by reaching 422.5 ppm in 2024.

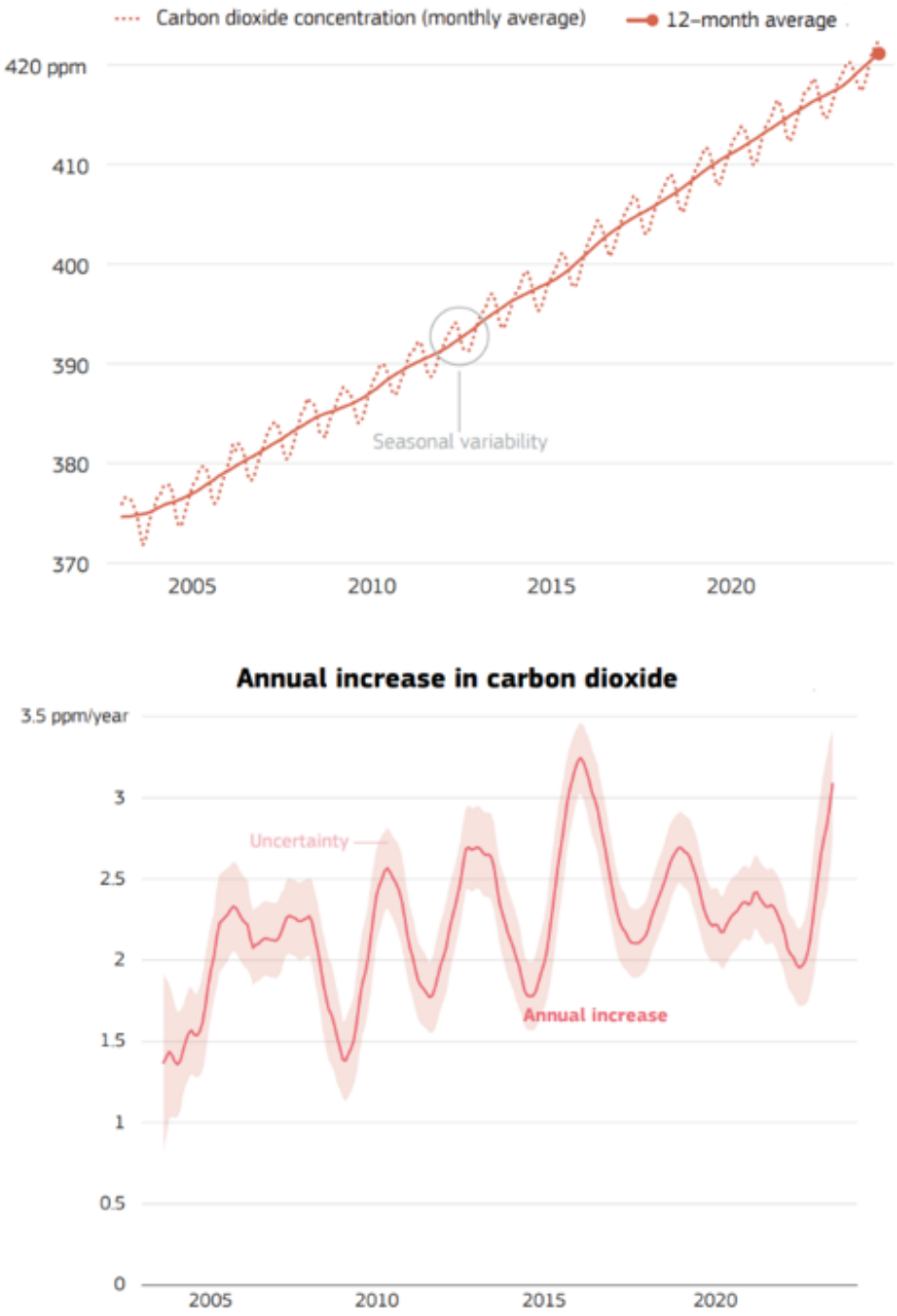

Figure 1.1: Data from C3S/Obs4MIPs (v4.5) consolidated (2003–2022) and CAMS preliminary near real-time column-averaged data (2023) GOSAT-2 records • Credit: C3S/CAMS/ECMWF/University of Bremen/SRON

As shown in Figure 1.1, the satellite data reveals an accelerating growth pattern with the steepest increases occurring in recent years. The plot demonstrates no signs of leveling off, with the curve showing annual growth rates consistently exceeding natural variability, indicating that anthropogenic emissions continue to overwhelm natural carbon sinks.

Methane concentrations similarly had concerning trends, from 1750 ppb in 2005 to over

1900 ppb by 2023—a 9% increase. The annual increase rate jumped significantly since 2020 and is now increased to about 18 ppb per year. This is troubling enough because methane is a potent greenhouse gas, with global warming potential of about 28-34 times that of CO_2 in a 100-year time horizon.

Insight into the patterns in growth of emissions today, particularly for CO_2 emissions by fuel and region, comes from International Energy Agency analysis. In 2024, natural gas contributes most to emissions growth with a 2.5% increase (180 Mt CO_2); coal emissions rise by 0.9% (135 Mt CO_2). Regionally, most emissions growth at 1.5% (375 Mt CO_2) is in the emerging market and developing economy sector while advanced economies decreased emissions by 1.1% (120 Mt CO_2).

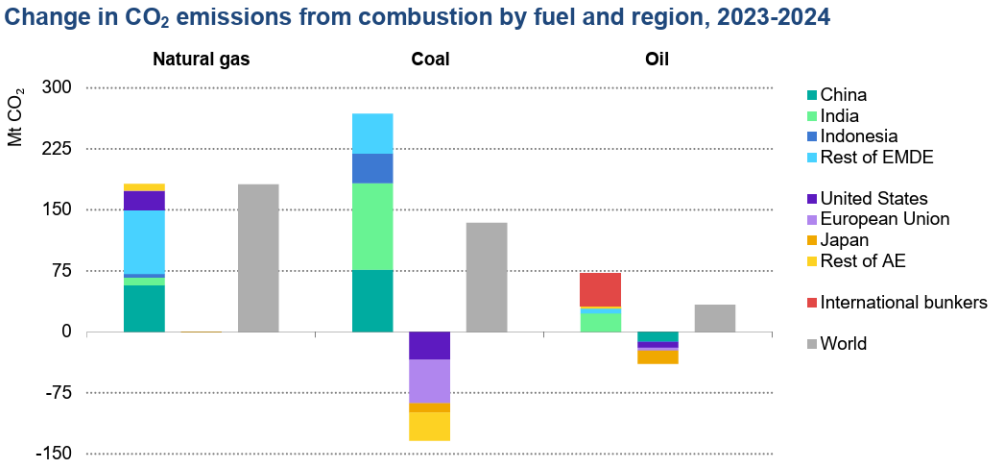

Figure 1.2: co2 emissions by fuel and region. Notes: AE = advanced economies; EMDE = emerging market and developing economies

Figure 1.2 illustrates that natural gas dominates emission growth at 180 Mt CO_2, primarily from power generation and industrial use in emerging economies. The plot clearly shows the stark contrast between declining emissions in advanced economies and rising emissions in developing regions, highlighting the global challenge of balancing economic development with climate goals. Despite troubling emission patterns, clean energy technologies are starting to show real effects. The period from 2019-2024 saw impressive growth of solar PV, wind power, nuclear capacity, electric vehicles, and heat pumps. These technologies currently represent an estimated 2.6 Gt net annual co2 reductions or about 7% of global energy-related emissions.

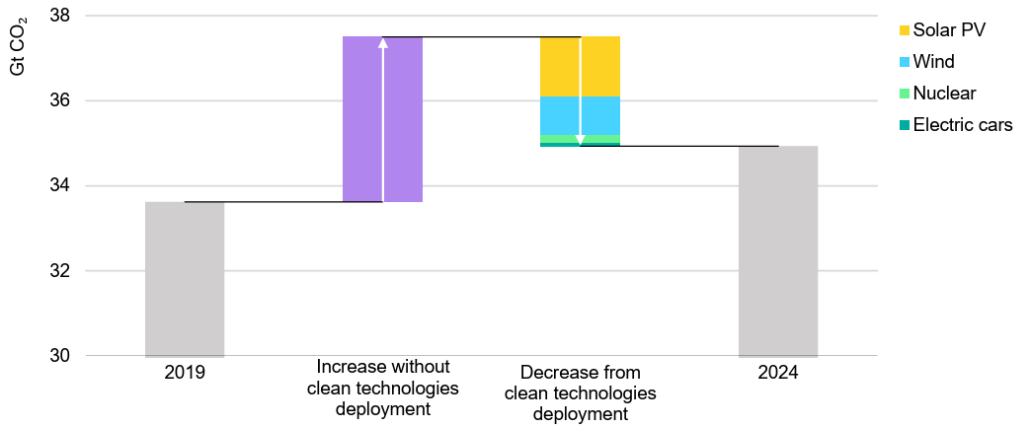

Figure 1.3: Change in co2 emissions from fuel combustion and avoided emissions from deployment of selected clean technologies, 2019-2024

Figure 1.3 demonstrates that solar PV and wind power contribute the largest emission reductions, followed by electric vehicles and heat pumps. The plot reveals that despite this progress, the avoided emissions remain small compared to total global emissions, indicating that clean technology deployment has not yet reached the scale needed to reverse global emission trends.

1.1.3 Paris Agreement Target and Gap Analysis

The Paris Agreement which came into existence in 2015[4], established a bold target to reduce global warming below two degrees Celsius past pre-industrial temperatures while working toward keeping it under 1.5 degrees Celsius.The IEA's Global Energy Review [5] shows a severe discrepancy between stated climate targets and current warming patterns because global temperatures reached over 1.5°C above pre-industrial times during 2024.

The emissions that are originated from climate are originated in various regions from 2000-2024 reveal diverging trajectories that more complexly shape global climate governance. While the European Union and United States have reduced emissions, the same is not true for China and India. The per capita analysis shows continued disparities that need to be addressed in order to mount an effective climate response.

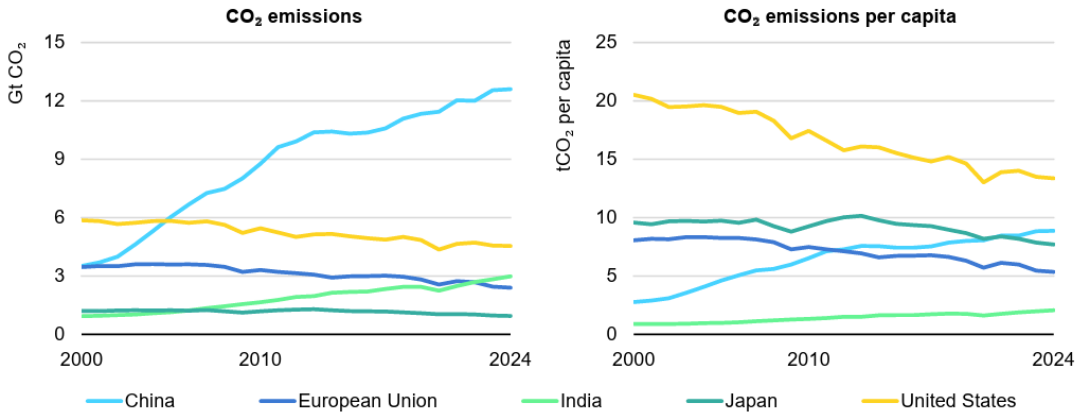

Figure 1.4: co2 total emissions and per capita emissions by region, 2000-2024

Figure 1.4 reveals that while China's total emissions have grown substantially, its per capita emissions remain below those of advanced economies. The plot shows the US and EU with declining trends in both total and per capita emissions, while India maintains the lowest per capita emissions despite economic growth, reflecting different stages of industrial development and energy system maturation.

The increased gap between Paris Agreement targets and actual action is due to several factors: increased carbon emissions, extreme weather events, stalled energy efficiency gains, and socioeconomic inequity between regions. As a result, action needs to increase, not only on the existing technologies available but also developing new methods, such as Carbon Capture and Storage, particularly with emission sources that cannot be eliminated through only electrification or renewable energy alone.

1.1.4 Role of Carbon Capture and Storage in Climate Mitigation

Carbon Capture and Storage functions as a key solution which supports the achievement of worldwide climate targets. The International Energy Agency Special Report on Carbon Capture, Utilization and Storage states that CCUS stands as the only technology set which reduces emissions across major economic sectors while removing CO_2 to offset unpreventable emissions. This unique dual role qualifies CCS as a key element in any net-zero emissions pathway.

CCS technologies provide a comprehensive approach to emissions reduction by addressing all three categories of organizational emissions:

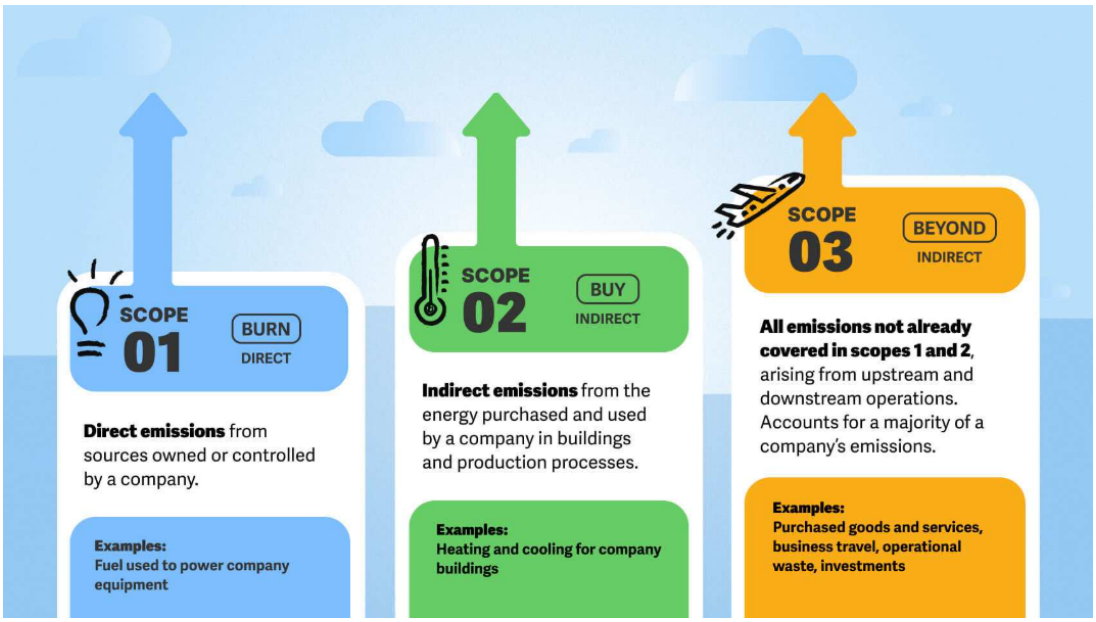

Figure 1.5: The three scopes of greenhouse gas emissions: Scope 1 (direct emissions), Scope 2 (indirect emissions from purchased energy), and Scope 3 (value chain emissions)

As illustrated in Figure 1.5, emissions cascade through organizational boundaries, with Scope 1 representing direct operational control, Scope 2 reflecting energy purchasing decisions, and Scope 3 encompassing the broader value chain. The diagram shows how

CCS can address emissions across all three scopes, making it uniquely comprehensive for industrial decarbonization strategies.

1.1.4.1 Scope 1 (Direct emissions)

The CCS system operates by extracting CO_2 emissions from industrial operations which include steel production in blast furnaces to stop pollution from reaching the atmosphere while allowing storage in underground rock formations.

1.1.4.2 Scope 2 (Indirect emissions from purchased energy)

Power plants that use CCS technology generate low-carbon electricity which helps decrease the environmental impact of industrial operations that need power. Studies predict that adding carbon capture systems to coal power plants can reduce their emissions by 90%.

1.1.4.3 Scope 3 (Value chain emissions)

The adoption of CCS technologies by suppliers will receive influence from companies who operate within their value chains because this practice holds key importance for cement manufacturing businesses since they emit about 60% of their total emissions from process sources.

1.1.5 Addressing Emissions from Existing Infrastructure

The worldwide energy system maintains a collection of current assets which hold considerable operational value for future use. Emerging Asian nations operate their coal power plants at less than 13-20 years of age even though these plants should last between 40-50 years. The existing primary steel production facilities which make up 40% of the current infrastructure will continue to function until 2050. These facilities will produce more than 600 $GtCO_2$ if no action is taken which equals the amount of current annual emissions for almost twenty years.

CCS provides the only practical solution to avoid premature shutdown of these facilities because it allows them to operate with lower emissions while keeping their current structure and supply networks and staff levels intact.

1.1.6 Solution for Hard-to-Abate Sectors

Heavy industry produces about 20% of worldwide CO_2 emissions while certain industrial sectors face difficult decarbonization obstacles which make CCS an effective solution.

1.1.6.1 Cement Production

The majority of emissions stem from the calcination process (chemical reaction) rather than fuel combustion, making CCS virtually the only technology option for deep emissions reductions. The fundamental chemical process $CaCO_3 \rightarrow CaO + CO_2$ is essential to cement production and cannot be eliminated through fuel switching or efficiency improvements.

1.1.6.2 Steel Manufacturing

Requires carbon as both a reductant and energy source. The CCUS systems operate at a cost that increases production expenses by 8-9% while hydrogen-based alternatives result in price hikes between 35-70%. The reduction of iron ($Fe_2O_3 + 3CO \rightarrow 2Fe + 3CO_2$) depends on carbon as an essential component which no other material can substitute in current blast furnace operations.

1.1.6.3 Chemical Production

The elimination of process-related emissions proves to be challenging through alternative methods. The production of numerous chemical processes depends on fossil fuels as raw materials instead of power sources which leads to unavoidable co2 emissions that require capture and storage solutions.

1.1.6.4 Natural Gas Processing

Removing CO_2 stands as a requirement because pipeline specifications need it regardless of climate change concerns which leads to capture and storage systems becoming a natural development of existing operations.

The IEA [5]report shows CCUS will achieve 60% of cement industry emission reductions under the sustainable development scenario which proves its essential value for industrial decarbonization.

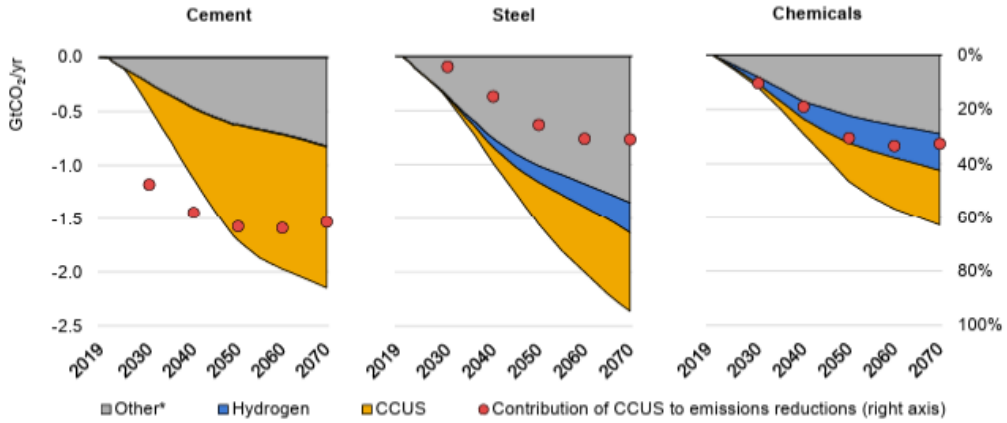

Figure 1.6: Global co2 emissions reductions by abatement measure in heavy industry in the Sustainable Development Scenario relative to the Stated Policies Scenario

Figure 1.6 clearly shows CCUS contributing the largest emission reduction wedge across multiple industrial sectors, significantly outpacing material efficiency, recycling, and fuel switching measures. The plot demonstrates that in cement production specifically, CCUS represents nearly twice the contribution of all other abatement measures combined.

1.1.7 Platform for Low-Carbon Hydrogen Production

Hydrogen-based CCS involves co2 capture from fossil fuel-based hydrogen production processes. The hydrogen industry generates 75 million tons each year through natural gas which makes up 76 percent of production and coal which accounts for 23 percent of

production while producing over 800 Mtco2 emissions that match the combined energy sector emissions of Indonesia and the United Kingdom. The integration of carbon capture technology with steam methane reforming systems allows hydrogen production to continue while achieving major reductions in environmental emissions.

CCS technology enables clean hydrogen production from natural gas or coal at prices that tend to be around 50% of electrolytic hydrogen costs in various areas which makes it the most affordable method to expand hydrogen production during the upcoming years.

The study [6],explains that fossil fuel-based hydrogen production with CCUS will stay the most affordable method because domestic coal and natural gas prices remain low and co2 storage facilities exist. The Sustainable Development Scenario produces 40% of its low-carbon hydrogen through fossil-based methods which use CCS technology because these methods work well in places where fossil fuels cost less and there is space for co2 storage.

1.1.8 Removing Carbon from the Atmosphere

The IPCC [7]Special Report on 1.5°C showed that carbon removal technologies must exist for successful climate action because 88 of 90 IPCC scenarios used net-negative emissions to achieve 1.5°C temperature limits.

CCS provides the technological foundation for carbon removal methods:

Bioenergy with CCS (BECCS)

The system extracts carbon dioxide from energy systems that generate power through renewable biomass resources.

Direct Air Capture with Storage (DACS)

The system extracts carbon dioxide from the surrounding air.

These solutions help organizations handle emissions from sectors that struggle with reduction and they provide protection against possible setbacks in emerging decarbonization technologies.

1.1.9 Importance of CCS in Decarbonization Efforts

Carbon Capture and Storage functions as a key climate change mitigation system which addresses emissions that prove challenging or impossible to reduce through conventional methods. The IEA analysis shows that CCS functions as a part of other climate solutions instead of operating as an independent technology.

Within the mitigation hierarchy, CCS serves crucial functions:

1.1.9.1 Primary mitigation option for hard-to-abate sectors

The main solution for industries which generate process emissions during production such as cement and steel manufacturing and specific chemical operations is CCS.

1.1.9.2 Transitional technology for existing infrastructure

Fossil fuel power plants together with industrial facilities use CCS technology to decrease emissions while continuing their regular operations without interruption.

1.1.9.3 Negative emissions enabler

The combination of BECCS and DACS enables CCS to extract co2 from the atmosphere which fulfills the key requirement of negative emissions for reaching advanced climate targets.

1.1.9.4 Hydrogen production pathway

The production of clean hydrogen through CCS serves as a transitional solution to create a hydrogen economy until green hydrogen achieves cost parity for broad industrial use. Figure 1.7 projects that gross CO_2 emissions will decline but remain substantial

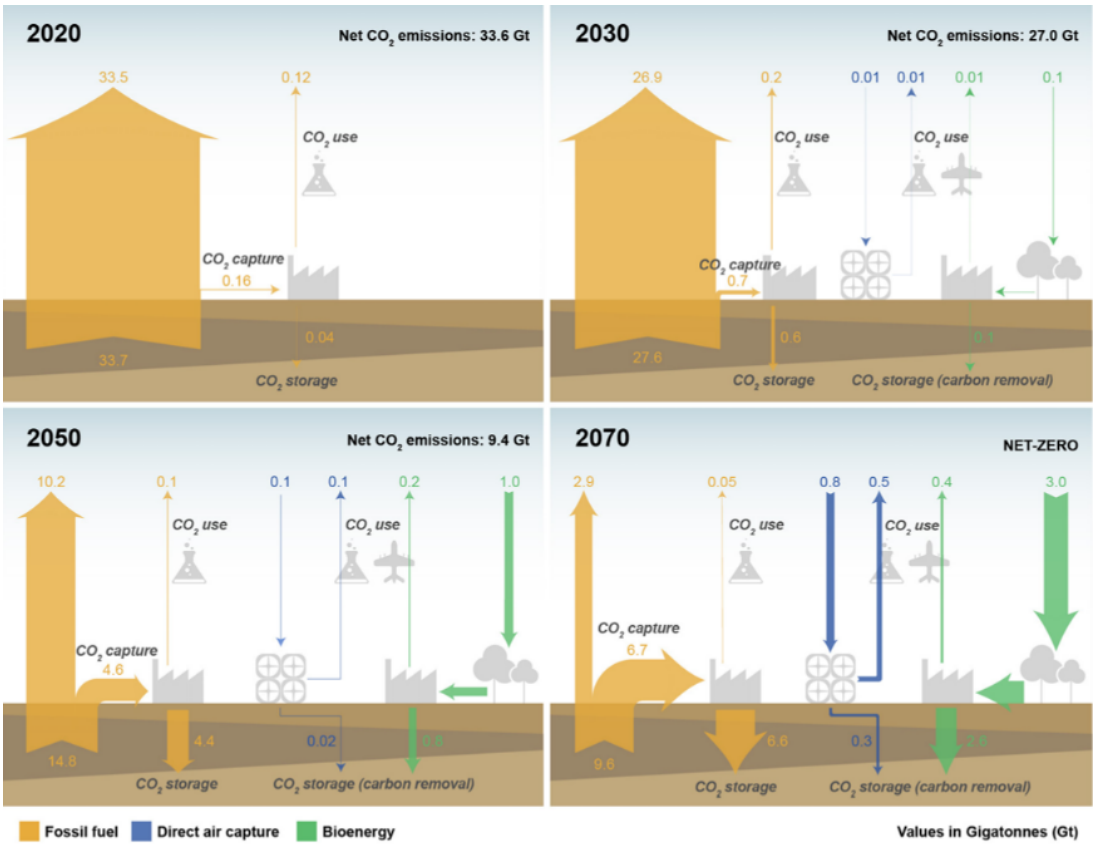

Figure 1.7: co2 emissions, capture and removal in the Sustainable Development Scenario

through 2070, while carbon capture and removal requirements grow exponentially. The plot indicates that by 2050, captured and removed CO_2 must reach approximately 10 Gt annually, scaling to over 15 Gt by 2070, demonstrating the massive scale of carbon management infrastructure required for net-zero targets.

Renewable energy systems together with electrification and efficiency improvements have shown notable advancement yet these methods alone do not reach the required emission reduction targets. CCS functions as a key element which supports these technologies to help achieve stronger climate action results. The IEA's Sustainable Development Scenario

predicts that CCUS will achieve almost 15 percent emission reduction targets through 2070 because of technological advancements and reduced expenses and limited environmental options.

1.2 Carbon Capture and Storage Fundamentals

1.2.1 Definition and Technical Components

The Carbon Capture and Storage (CCS) system operates to decrease industrial carbon dioxide emissions which stem from cement manufacturing and steel production and hydrogen creation. The technology provides major emission reductions for difficult-to-treat industrial sectors while producing negative emissions when used with bioenergy systems or direct air capture.

The CCS chain operates through four fundamental stages.

1.2.1.1 Capture

The capture phase represents the most technically complex and expensive

component of the CCS chain.The system requires the removal of carbon dioxide from industrial waste streams and combustion emissions before they enter the atmosphere. The three main strategies exist.

Post-combustion capture: Employs chemical absorption methods like amine scrubbing.The technology stands as the most developed method which shows commercial readiness according to Technology Readiness Level (TRL) assessments.

Pre-combustion capture: Typically used in gasification-based processes where coal is gasified to produce CO and H_2, then steam converts CO to co2 for separation from the hydrogen stream.The method faces difficulties in implementation because the Kemper County IGCC project stopped operations due to technical and financial issues.

Oxy-fuel combustion: Burns fuel in pure oxygen instead of air, producing concentrated co2 flue gas.The technology has reached demonstration-scale development which shows promising commercial potential in the near future.

Scientists have developed new carbon capture methods which include solid sorbent adsorption and polymeric and ceramic membranes and metal oxide oxygen carriers in chemical looping combustion (CLC).

1.2.1.2 Transport

The process starts when captured co2 undergoes compression to reach supercritical state which exists at around 100-150 bar pressure and 30-40°C temperature. The phase transition generates a major reduction in co2 volume which allows for better storage capacity and cost-saving pipeline transportation. Engineering design requires knowledge about fluid phases and pressure control and water and H2S and O2 presence and suitable materials to stop corrosion and hydrate formation.

Alternative transport methods include shipping for long-distance and overseas transport (co2 stored as liquid at -50°C and 7 bar) and trucking/rail for small volumes and

demonstration projects.

1.2.1.3 Injection

Specialized wells operate to inject compressed captured co2 into deep geological storage sites located underground. The design of wells for supercritical co2 injection requires them to function under corrosive environments and fluctuating pressure conditions and possible geological reactions.

1. **Casing materials:** Corrosion-resistant alloys (CRA) or internally coated carbon steel
2. **Cement sheaths:** Low permeability formulations resistant to carbonation
3. **Well completion:** Multiple packers for zone isolation and fiber-optic cables for distributed temperature and acoustic sensing

   (DTS/DAS)

The systems collect co2 data which lets operators detect leaks and determine pressure status. The injection method needs exact control of pressure and temperature because it stops caprock damage and deals with pressure changes between wells and geochemical effects that reduce injection capacity.

1.2.1.4 Storage

The success of CCS in climate change mitigation depends on having safe underground storage facilities which can last for extended periods. The process of geological storage sends co2 deep into porous rock structures which contain the gas for lengthy periods ranging from thousands to millions of years while preventing any major escapes into water sources or air.

Geological Criteria for Storage Formations:

- High porosity (15-30%) for sufficient pore volume
- High permeability (10-1000 millidarcies) for effective injectivity
- Deep burial (>800m) to maintain supercritical phase
- Thick, laterally continuous caprock for sealing
- Geological stability without compromising faults or fractures
- Hydraulic isolation from freshwater aquifers

Principal Storage Formation Types:

Depleted Oil and Gas Reservoirs: Offer attractive storage solutions due to proven containment capability over geological timescales.The benefits section provides detailed geological information together with pre-existing infrastructure for reduced costs and verified storage systems and the possibility to perform Enhanced Oil Recovery operations.

1.2.2 Supercritical co2 Properties

Supercritical co2 develops when substances undergo heating together with compression beyond critical temperature point of 31.1°C and critical pressure value of 7.38 MPa which produces properties that combine characteristics of gases and liquids (National Institute of Standards and Technology, 2008).

Density: The density falls between 600 and 800 kg/m^3 which allows it to hold 300-400 times more mass than standard gaseous co2 making it suitable for underground storage.

Viscosity: The substance maintains a viscosity range of 0.05-0.08 mPa·s which surpasses the density of water and oil thus enabling it to move through porous materials with minimal pressure requirements while enhancing its injection capacity.

Diffusivity: The substance maintains moderate diffusivity between liquid and gas states which allows it to flow through tiny openings while contacting formation fluids and staying mobile for environmental changes.

Surface Tension: The lower surface tension between co2 and formation water compared to gaseous co2 results in improved displacement performance and changes to how residual trapping operates.

Solvent Properties: Functions as powerful non-polar solvent,

1.3 Monitoring Challenges in CCS Operations

1.3.1 Regulatory Requirements and Pressure Monitoring Mandates

The regulatory systems for carbon capture and storage demand complete monitoring systems to verify both safety and effectiveness of co2 storage. These requirements establish foundations for monitoring protocol development and technology selection, directly that directly influences the need for advanced pressure prediction capabilities.

1.3.1.1 International Monitoring Framework

Large-scale CCS implementation involves storing millions of tons of co2 annually in extensive geological formations, creating health, environmental, financial, and property risks requiring robust monitoring frameworks. The International Risk Governance Council (IRGC) recognized the need for regulatory systems to address co2 storage as an urgent priority [8].

Multiple international frameworks determine the monitoring standards which CCS operations must follow.

The Kyoto Protocol recognized CCS as a climate change solution which led to its official inclusion within the Clean Development Mechanism (CDM) during 2011 The system demands thorough monitoring and reporting and verification (MRV) processes to confirm that the stored co2 remains permanent for carbon credit eligibility.

1.3.1.2 Regional Regulatory Variations

Multiple regions maintain various standards for CCS monitoring systems.

EU: The EU Directive 2009/31/EC mandates that operators continuously monitor the co2 volumetric flow rate, injection wellhead temperature and pressure, and the temperature and pressure in the reservoir [9].

US: The EPA's Underground Injection Control (UIC) Class VI rule sets requirements for the tracking of the co2 plume and pressure front, monitoring of injection well testing, and groundwater monitoring for the different zones [10]

1.3.1.3 Monitoring Parameters and Frequency

The majority of regulatory systems include equivalent pressure-related standards.

Injection parameters: co2 injection pressure and temperature, real-time injection rate monitoring

Reservoir parameters: Formation pressure and temperature, pressure front migration, spatial distribution of co2 plume

The monitoring frequency depends on specific parameters and project stages yet most regulations require operators to monitor injection parameters continuously which creates a need for instant pressure prediction under changing underground conditions.

1.3.2 Critical Parameters for CCS Pressure Prediction

1.3.2.1 Bottom Hole Pressure Monitoring and Measurement Importance

Bottom hole pressure serves as the core parameter for CCS projects because it allows direct measurement of reservoir conditions instead of relying on surface-based estimates. Physical phase and behavior of compressed co2 within reservoirs depend highly on pressure conditions, directly impacting storage capacity, injectivity, and containment security.

The co2 phase diagram shows why bottom hole pressure monitoring stands as essential because reservoir conditions often place co2 near phase transition points which makes co2 properties extremely sensitive to minor pressure changes.

Figure 1.8 clearly illustrates why precise pressure monitoring is critical for CCS operations. The diagram shows that within the typical CCS pressure range (highlighted in red), CO_2 exists near the supercritical boundary where small pressure variations can cause significant phase transitions. The plot demonstrates that reservoir conditions often place CO_2 in regions where density and viscosity properties change dramatically with minor pressure fluctuations, making accurate bottom hole pressure measurement essential for predicting storage behavior and ensuring containment integrity.

Bottom hole pressure monitoring functions as a key tool which supports various essential operations in the field.

1. The system allows users to view real-time data about reservoir pressure levels.
2. The system detects pressure surges at injection sites which might lead to damage of the caprock seal.

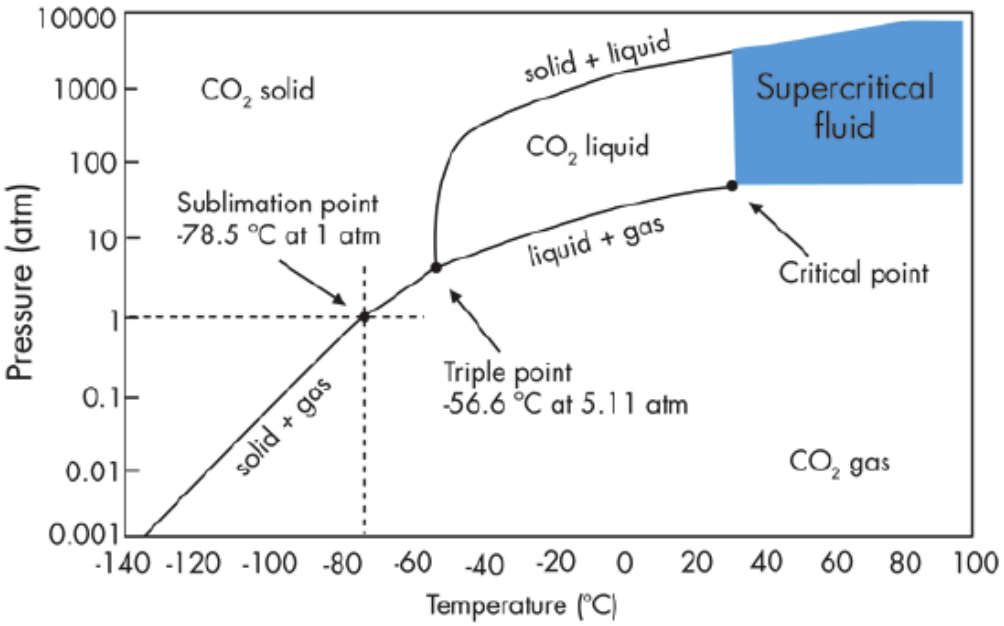

Figure 1.8: Phase diagram for pure (100%) co2 (adapted from IPCC, 2005) [1]. The red box indicates approximate temperature and pressure ranges for CCS

3. The system identifies when pressure changes occur between different wells and storage areas.
4. The system detects pipeline leaks when it observes unusual pressure fluctuations in the network.
5. The process aligns reservoir simulation models with actual formation pressure data that has been gathered.

The EU Storage Directive mandates storage operators to prove predicted co2 behavior through their storage facilities by using exact bottom hole pressure measurements according to [9].

1.3.2.2 Temperature Measurements and Significance

The method of surface temperature measurement lacks the fundamental information which bottom hole temperature monitoring provides to establish correct thermal storage conditions for co2 behavior analysis. co2 viscosity shows major thermal changes during the storage temperature range of 30-60°C.

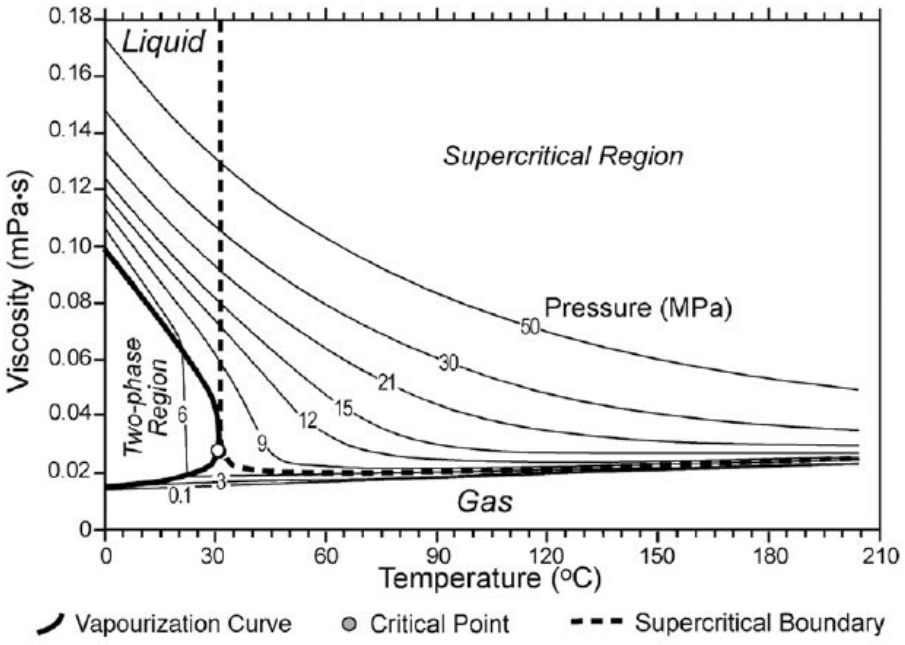

Figure 1.9: Viscosity as function of temperature and pressure (IPCC, 2005) [1]

Figure 1.9 demonstrates the complex relationship between CO_2 viscosity, temperature, and pressure within CCS operating ranges. The plot shows that viscosity varies by more than an order of magnitude across typical storage conditions, with particularly steep gradients near phase boundaries. This dramatic variation illustrated in the figure explains why surface temperature measurements are inadequate for CCS operations, as accurate reservoir temperature data is essential for predicting injection rates, flow patterns, and long-term storage behavior.

Bottom hole temperature monitoring is critical because:

- The stored co2 undergoes phase changes and density variations because the reservoir temperature determines its physical state.
- The process of injecting heat into the reservoir creates major alterations in the stress distribution of reservoir rocks.
- The way heat spreads through reservoirs determines how co2 moves and where it settles.
- Temperature irregularities serve as indicators of unexpected fluid flow or chemical interactions.
- Thermal monitoring systems detect co2 migration through their monitoring well detection systems.

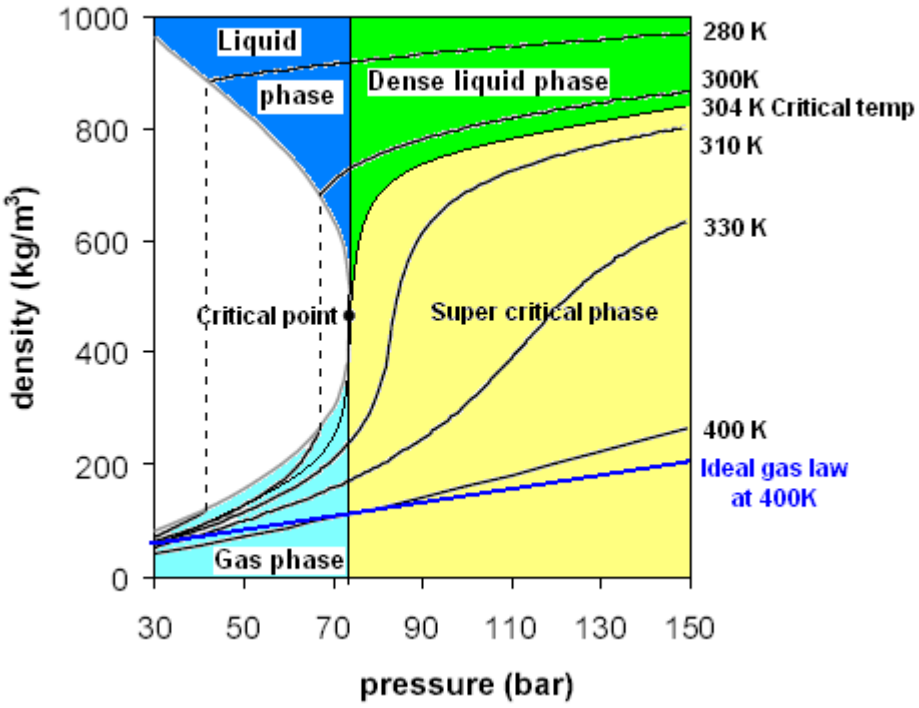

Figure 1.10: Density diagram for pure co2 between 30 and 150 bar at temperatures between 280K (7°C) and 400K (127°C)

Figure 1.10 reveals the extreme sensitivity of CO_2 density to temperature variations, particularly in the supercritical region relevant to CCS operations. The plot shows density changes of several hundred kg/m^3 across the typical CCS temperature range of 30-60°C. This steep density gradient demonstrated in the figure directly impacts storage capacity calculations and injection volume predictions, highlighting why precise bottom hole temperature monitoring is crucial for accurate reservoir management and capacity assessment.

The supercritical area demonstrates extreme CO_2density sensitivity to temperature variations which requires precise temperature monitoring to determine injection volume and storage capacity.

1.3.3 Environmental Challenges and Monitoring Limitations

1.3.3.1 Effects of High Pressure/Temperature on Equipment

The bottom hole monitoring equipment needs to function under extreme conditions which include pressures that exceed 200 bar and temperatures between 40 and 100 degrees Celsius. The equipment faces major reliability issues because of these environmental conditions which impact its performance over extended periods.

The process of observing CO_2 phase behavior encounters disruptions because nitrogen gas contains oxygen impurities. N_2 modifies the phase boundaries of CO_2 which leads to direct effects on the calibration of the equipment and the accuracy of the measurement.

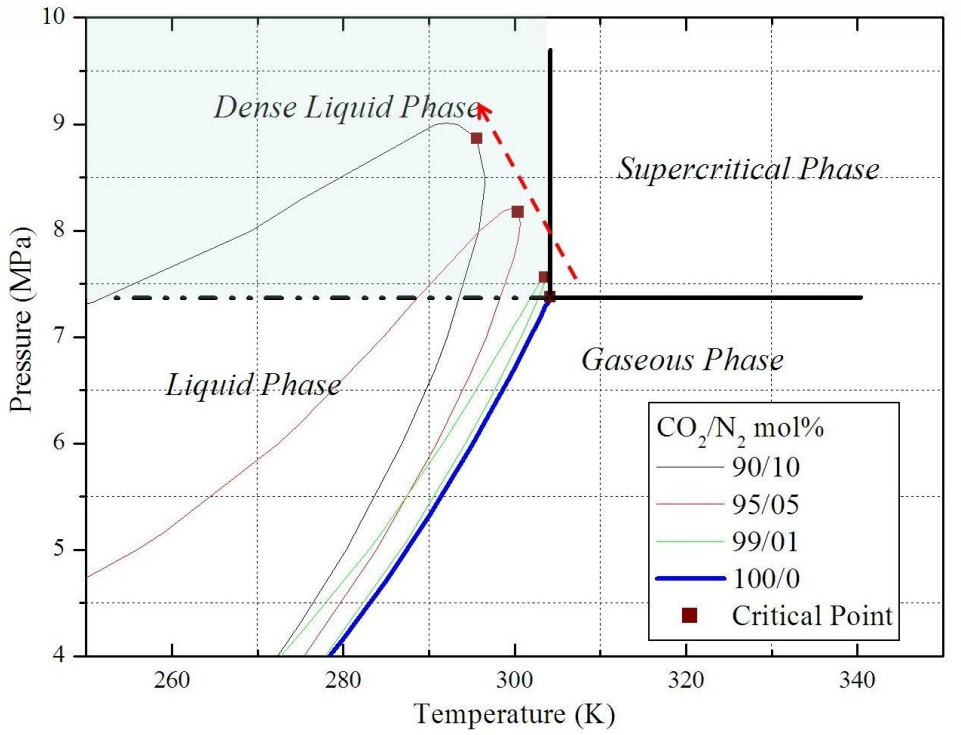

Figure 1.11: Influence of N_2 on the physical state of co2 (Pipich & Schwahn, 2020) [2]

Figure 1.11 illustrates how nitrogen contamination significantly alters CO_2 phase behavior, shifting phase boundaries and creating measurement uncertainties. The plot demonstrates that even small concentrations of N_2 can modify the pressure-temperature conditions at which phase transitions occur, directly affecting equipment calibration and measurement accuracy. This contamination effect shown in the figure represents a major challenge for bottom hole monitoring systems, as it introduces systematic errors that must be accounted for in CCS operations and reservoir modeling.

Bottom hole instruments encounter multiple operational difficulties which affect their ability to perform correctly.

- Excessive pressure levels threatening sensor elements and calibration precision
- Electronic component degradation through accelerated temperature exposure
- Shifting co2 states influencing sensor functionality
- Pressure surges during injection and shut-in operations
- Temperature cycling creating damaging effects on instrument components

1.3.3.2 co2 Corrosivity Factors

The combination of supercritical co2 with formation brines and contaminants creates an extremely corrosive environment. Principal corrosivity factors include:

- Carbonic acid formation from co2 dissolution in formation water

- Chloride presence in formation brines accelerating corrosion
- Possible elemental sulfur or hydrogen sulfide in certain formations
- Galvanic corrosion between dissimilar metals
- Erosion-corrosion from high-velocity injection flow
- High-temperature, high-pressure stress corrosion cracking

The EU Storage Directive demands operators to accept only co2 streams which have undergone analysis for corrosive substances and risk assessment [9].

1.3.3.3 Depth and Accessibility Limitations

The tracking of bottom hole operations becomes extremely difficult when working at depths between 1 and 3 kilometers.

- Extremely limited equipment retrieval or replacement opportunities
- Intervention operations costing over $1 million per operation
- Specialist equipment requirements (workover rigs, coiled tubing units)
- Well control loss risks during intervention operations
- Design limitations to solutions deployable during initial completion

Offshore projects encounter these problems at a higher level because they face restricted access to intervention equipment and short operational periods due to weather and costly transportation and limited platform space.

1.3.3.4 Essential Requirement for Predictive Modeling

The reduced availability of bottom hole gauges creates an urgent need for precise predictive models which need to:

1. Complete data gaps when monitoring systems fail before replacement
2. Validate measurements and identify faulty readings from deteriorating instruments
3. Extend coverage to non-directly gauged reservoir areas
4. Forecast future reservoir behavior from existing trends
5. Optimize intervention decisions for gauge replacement timing

1.3.4 Long-Term Reliability Requirements

CCS facilities demand continuous monitoring throughout their operational life and during their post-closure phase which could last for many decades or even centuries. The EU regulatory framework establishes distinct monitoring stages which include:

- Pre-injection monitoring: Baseline condition establishment
- Operational monitoring: Active injection phase

- Post-closure monitoring: After injection cessation
- Post-transfer monitoring: After responsibility transfer to authorities

The prolonged time periods create various difficulties which include:

- Equipment requiring year-to-year operation without maintenance
- Continued monitoring responsibility after storage site abandonment
- Monitoring continuation after responsibility transfer for leakage detection
- Operator cost responsibility for monitoring at least 30 years post-transfer
- Standardized quantification methods with known uncertainties

The Storage Directive mandates operators to keep monitoring responsibilities active after site closure while they must maintain detection capabilities for leaks and major irregularities at reduced levels of surveillance [9].

The monitoring difficulties in CCS operations require scientists to create new predictive models which produce accurate forecasts of pressure and temperature since direct measurement becomes unfeasible.

The direct need for virtual sensing technology development drives this research to establish machine learning-based virtual sensing systems as its main contribution.

Chapter 2

Simulation Tools for CCS Applications

Computational modelling tools are very critical for designing, optimizing, and the prediction of the behavior of the carbon capture and storage systems. These tools allow engineers to analyze the fluid dynamics, phase behavior and the operational risks in different conditions. There are two categories of simulators that are used:

1. Steady-state simulators
2. Dynamic transient simulators

Each of these simulators address an aspect of the CCS workflows.

In this section we evaluate the application of **Prosper** for **steady-state** modeling, and **OLGA** for **transient** modeling in the process of the transport and storage of carbon dioxide along with each of the limitations that these simulations have.

2.1 Steady-State Simulators (PROSPER)

Fixed-time equilibrium systems are provided with advanced simulation by steady-state simulators mimicking constant flow rates, pressures, and temperatures. The Prosper computer program performs the nodal analysis to build steady-state models for co2 injection wells and individual pipeline-wellbore systems and surface facilities. Using formalized methods, nodal analysis divides the flow system into nodes and then calculates the pressure differences between nodes using empirical correlations developed from experimental data, such as Hagedorn & Brown [11] and Beggs & Brill [12]. PROSPER relies on these empirical correlations rather than mechanistic models, applying them to estimate pressure drops and predict multiphase flow behavior, including Vertical Lift Performance (VLP) correlations used to evaluate flowline and conduit pressure losses.

In CCS operation, Prosper facilitates:

- co2 injection system design: Calculation of steady-state pressure and temperature profile for isolated pipeline-wellbore systems
- Sensitivity analyses: Computation of parameter effects like injection rate, wellhead pressure, or fluid composition on system performance

Prosper employs thermodynamic models like Prosper's Peng-Robinson equation of state to simulate co2 phase behavior at the critical point at which pressure and temperature changes control density and viscosity. Its steady-state assumption excludes it from simulating dynamic processes like transient pressure waves or operation change in startup-shutdown sequences in pipelines.

2.2 Transient Simulators (OLGA)

OLGA is an advanced multiphase flow simulator that uses computational fluid dynamics equations to model high-performance pipeline, wellbore, and storage reservoir behavior. The numerical methods and thermodynamic models of OLGA provide full capabilities for co2 system modeling, which are subject to phase transitions and heat effects that control their behavior.

Transient simulators are required to investigate time-dependent behavior of the system in CCS dynamic operations with changing operational conditions over time. OLGA demonstrates good performance in modeling transient conditions based on:

1. Transient flow in pipelines: Simulation of the phase transition of co2 from supercritical to liquid and gas phases during transportation under conditions of terrain-caused pressure and temperature variations.
2. Integrity failure analysis: Sensitivity analysis of the co2 leakage rate after pipeline or wellbore integrity failure.
3. Time-dependent injection operations: Time-dependent injection well simulation is coupled simulation accounting for pipeline, wellbore, and near-wellbore reservoir properties under transient conditions.

OLGA accurately models multiphase flow behavior like slugging, transient heat transfer, and compositional gradients. The simulator will project peak pipeline pressure when starting up and model contaminant dispersion in co2 transport streams for various time increments. The computational demands and complexity of OLGA require careful consideration of simulation scope and objectives when applying it to CCS system analysis.

2.2.1 Numerical Approach: Navier-Stokes Equations

OLGA uses a simplified one-dimensional representation to model unsteady flow characteristics which follow the Navier-Stokes equations. The mathematical representation of mass and momentum and energy conservation laws is applied to the geometries of pipelines and wellbores through a discretization procedure.

2.2.1.1 Mass Conservation

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho u)}{\partial x} = 0 \tag{2.1}$$

Where:

- ρ is fluid density
- u is velocity

- x is the spatial coordinate

2.2.1.2 Momentum Conservation

$$\frac{\partial(\rho u)}{\partial t} + \frac{\partial(\rho u^2)}{\partial x} = -\frac{\partial P}{\partial x} + \rho g \sin\theta - \frac{f \rho u^2}{2D} \tag{2.2}$$

Where:

- P is pressure
- g is gravitational acceleration
- θ is pipe inclination
- f is the Darcy-Weisbach friction factor
- D is pipe diameter

2.2.1.3 Energy Conservation

$$\frac{\partial(\rho h)}{\partial t} + \frac{\partial(\rho u h)}{\partial x} = \frac{\partial P}{\partial t} + q - \frac{4U}{D}(T - T_{amb}) \tag{2.3}$$

Where:

- h is specific enthalpy
- q is heat transfer rate
- U is the overall heat transfer coefficient
- T_{amb} is ambient temperature

OLGA solves the equations mentioned by a finite volume numerical model, with the pressure-velocity coupling stabilized in a staggered grid [13]. For multiphase flow (e.g., co2 with contaminants), it employs a two-fluid model where phases are treated as separate continua and interphase momentum and energy transfer are assumed.

2.2.2 Joule-Thomson Effect in co2 Systems

The JT effect is the change of temperature during adiabatic expansion/compression when a fluid is caused to pass adiabatically through a restricted passage or a porous plug under such conditions that its kinetic energy of flow on either side of the plug is negligible [14]. The Joule-Thomson effect is critical for CCS operations, as quick pressure drops, for example, across valves or when there is a leakage, can induce the phase transition of carbon dioxide or thermal stresses. The limiting ratio of the temperature change Δt to the pressure change Δp as the latter approaches zero is variously known as the Joule-Thomson effect, the Joule-Kelvin effect, or simply the porous plug effect. This ratio, here designated by μ, is analytically defined by the relation $\mu = (\partial t/\partial p)_h$ (Burnett, 1923). OLGA incorporates this effect through the energy equation and its thermodynamic package:

2.2.2.1 Thermodynamic Modelling

- OLGA uses equations of state (e.g., Span and Wagner EOS) to compute co2 properties (ρ, h, C_p)
- The JT coefficient ($\mu_{JT} = \left(\frac{\partial T}{\partial P}\right)_h$) is derived implicitly from the EOS, and it enables accurate temperature prediction during throttling processes

2.2.2.2 Transient Scenarios

- During pipe depressurization (e.g., emergency shutdown), OLGA simulates local cooling due to μ_{JT}, which can cause co2 to transition from a supercritical state to liquid/gas phases, resulting in the potential for hydrate formation or ductile fracture
- Temperature changes within wellbores from JT effects are simulated for injection wells to assess thermal stresses on casing materials, including co2 heating downhole due to fluid compression during injection operations, as experienced in field applications

2.2.2.3 Applications in CCS

OLGA's physics-driven approach enables simulation of:

Flow assurance studies

- Pipeline transient flow: Phase change (supercritical $\leftrightarrow$ liquid/gas) due to terrain-induced pressure drops or valve closure
- Leakage dynamics: JT cooling during co2 blowdown, affecting leak rates and dispersion behavior
- Startup/shutdown sequences: Thermally induced stresses in wellbores and pipelines

Chapter 3

Research Gap Analysis

Even though there have been major advances in machine learning applications for carbon capture and storage, several critical research gaps limit the widespread deployment of virtual sensing technologies for real-time downhole condition monitoring. The most major of these challenges is the need for systematic training procedures that effectively integrate physics-based simulation outputs with real-time operational data to create robust predictive models for CCS monitoring applications.

3.1 Limited Deployment in Operational Environments

While ML models have showen promising performances in simulations, the real-world implementation in CCS plants remains unsure. The U.S. Department of Energy's SMART Initiative, a comprehensive ten-year program funded by DOE's Carbon Storage and Upstream Oil and Gas Program, is one of the most significant efforts to fill this gap through science-informed ML applications in subsurface operations [15]. However, extensive field validation is still needed to showcase the scalability, robustness, and flexibility of ML algorithms under various operating conditions, where changing the process parameters can highly influence model performance.

The above gap is more evident in downhole monitoring, where challenging environmental conditions, as well as multiphase flow dynamics, are the main hurdles in the process. In this regard, the demanding environment associated with CCS operations, including high-pressure (above 150 bars), high-temperature (40-100 degrees Celsius), and corrosive CO_2, poses considerable challenges to sensor deployment and ML application reliability. [9]. Training models solely based on simulation data does not account for all the complexities of an environment, and training using only small amounts of real-time data does not offer the holistic physical comprehension needed for accurate performance.

Recent studies involving ML for prediction of pressure in petroleum engineering applications have yielded encouraging results [15, 16], although such techniques have yet to be successfully used for the thermodynamic behavior of CO_2 in supercritical phases.

3.2 Machine Learning Applications in CCS and Petroleum Engineering

3.2.1 Current Applications and Limitations

Machine learning in carbon capture, utilization, storage, and transportation technologies has proved to be an effective means for improving systems efficiency and reducing operational expenses [17]. The current uses of machine learning technologies include several aspects of CCS technology:

Reservoir characterization and site selection: Machine learning models process huge amounts of data obtained via seismic analysis, well logging, geological surveys, etc., allowing for selecting the best CO_2 storage sites, predicting reservoir porosity, permeability, and capacity for CO_2 storage [17]. It has demonstrated considerable potential in addressing uncertainties in site selection process.

Monitoring and optimization: After CO_2 has been injected into the formation, the machine learning algorithm analyzes the data provided by the sensors measuring pressure, temperature, and geophysical parameters to track CO_2 migration, leaks, or other anomalies [18]. However, at present, this kind of machine learning technology is mostly used for reservoir-scale monitoring instead of individual wells.

Process optimization: ML technologies optimize the injection rate and pressure in order to prevent formation damage and improve the dispersion of CO_2 underground [19]. Even though this technology exists, its combination with downhole data has not been addressed yet.

3.2.2 Gaps in Virtual Sensing Applications

Despite the effectiveness of ML in traditional petroleum engineering through pressure prediction and monitoring of reservoirs [15, 16], there are some vital shortcomings preventing the immediate use of this technique in CO_2 sequestration and storage systems:

- **Thermodynamically Complex Nature:** Traditional ML techniques developed for petroleum engineering mainly consider conventional systems far from critical conditions, whereas CO_2 injection systems are thermodynamically close to critical conditions.
- **Scarce Operational Information:** As the history of petroleum industry is rich in terms of operational information, CO_2 sequestration and storage do not possess enough historical data.
- **Strict Regulatory Standards:** CO_2 sequestration and storage systems have strict regulations compared to traditional petroleum engineering.

3.2.3 Virtual Sensing as Real-Time Simulation for CCS Applications

The recent literature by Mustafee et al. [20] offers a theoretical basis for hybrid modeling that combines real-time data with simulation. They classify models into traditional simulation (only historical data), Real-time Simulation (RtS) (both historical and limited

real-time data), and Digital Twins (full real-time data). This theory directly describes the problem of CCS virtual sensing.

Virtual sensing for predicting the pressure of CCS falls under RtS, where *information fusion* is necessary to combine simulation history data with limited operational data. This approach bridges the inherent gap between comprehensive simulation ability and limited sensor availability in CCS systems.

Their four-dimensional model explains why existing methods are insufficient in CCS operations: (1) **Modeling Objective** involves making real-time decisions instead of strategic planning, (2) **Data Requirement** requires information fusion because sensors are limited, (3) **Implementation** aims at "situational awareness", and (4) **Experimentation** requires faster-than-real-time predictions for decision-making.

This theory offers theoretical background on why simulation alone cannot be used effectively to train AI for CCS applications and supports the need for using hybrid training techniques that involve incorporating limited field data into extensive simulated data.

3.3 Challenges of Data Quality and Availability

The availability of high-quality and varied datasets is crucial for training machine learning models effectively. However, obtaining such datasets from real-life CCS processes is quite challenging since most of the datasets remain proprietary and regulated, and there are few ongoing CCS projects in the world. In addition, the availability of operational datasets is quite limited for downhole applications because field data can only cover the range of operation during the entire life cycle of a CCS project.

The ongoing CCS projects include the Sleipner project in Norway [21], and the upcoming CCS projects include the Ravenna CCS Phase 1 project. However, the varied geologies, operating conditions, and composition of CO_2 in other CCS projects makes training generalized machine learning models very challenging.

For example, physics-based simulations, such as those performed using software like OLGA, can produce large sets of data over a wide range of operating regimes, including regimes that have not yet been experienced in practice but that are still physically feasible. However, there are few methods available in the literature for integrating these two data sources into valid physical and operational datasets.

3.4 Shortage of Standardized Benchmarking and Datasets

Lack of standard benchmarking techniques and datasets is another challenge facing the use of machine learning in downhole operations in CCS. The majority of studies in this area depend on proprietary or project-specific datasets, which makes it hard to directly compare different models or develop industrial standard procedures. This challenge becomes worse when we consider downhole applications since there is a lack of data.

There have been advances in developing standard techniques for some CCS applications under the SMART initiative [22]. Unfortunately, there are no standard benchmarking datasets for downhole pressure estimation. The use of physics-based simulation outputs

and measurements to develop a more comprehensive dataset for virtual sensing applications is a viable solution. Still, standard techniques for this purpose are yet to be developed in the literature.

Some recent studies in machine learning applications for carbon dioxide storage [17] emphasize the importance of having standard datasets that can be used to evaluate various machine learning approaches in different CCS applications. These studies do not address the specific application of down-hole pressure prediction in injection wells.

3.5 Explainability and Trust in ML Predictions

The problem of making sense of decision-making algorithms based on machine learning (ML) in safety-critical CCS operations is an important issue for practical implementation. Most ML models operate as "black boxes," which leads to a trust issue for engineers and operators because they need to be able to comprehend the underlying prediction rationale in order to make decisions.

According to the EU CCS Directive, operators are obligated to show the "expected CO_2 behavior" using monitoring systems [9]. Therefore, there is a need for prediction methods that can provide a sufficient level of explanation and trustworthiness. While current ML applications in CCS operations seem promising regarding their performance, they still lack explainability.

The idea behind using the outputs of a well-known physics-based tool (e.g., OLGA) as the training datasets of ML models in combination with real-time operational data sets offers a starting point for developing a framework for explainable AI.

3.6 Limited Integration of Physics-Based Knowledge with Real-Time Data

There is currently a problem inherent to all machine learning techniques used in subsurface monitoring: the use of physics-based models or real-time operational data in an ad-hoc manner. While the former can provide detailed simulations for flow dynamics and thermodynamics, the latter provides accurate, operational information that cannot be captured via modeling. On the contrary, models based only on operational data are unable to capture enough physics in the process to predict anything outside of what was seen in the data collected.

It is especially challenging in situations where the physics of multiphase flows and pressure dynamics is highly complex, but the amount of data available is small or of poor quality. The works on CO_2 solubility prediction [19] and reservoir monitoring for geological storage [18] showcase how this issue could be addressed via physics-informed machine learning; yet a framework for merging simulations with operational data to train a model for predicting downhole pressures has not been established.

The lack of structured procedures that combine the physical credibility of traditional simulation tools with the adaptability of real-time operational data represents a critical limitation in current virtual sensing research. Systematic integration of OLGA simulation

outputs with real-time gauge measurements could overcome this limitation by generating training datasets that combine comprehensive scenario coverage from physics-based simulations with operational relevance from real sensor observations.

3.7 Limited Real-Time Performance Validation

While several studies have demonstrated ML performance using historical data or simulation-based validation, comprehensive real-time performance validation in operational CCS environments remains limited. The regulatory requirements for CCS monitoring [9] demand continuous, reliable performance under dynamic operational conditions, but most ML applications have not been validated under these stringent requirements.

The temporal stability of ML models under varying operational conditions represents an underexplored area in CCS applications. Unlike laboratory or simulation environments, operational CCS facilities experience equipment drift, measurement noise, changing environmental conditions, and operational procedures that can significantly affect model performance over time.

Chapter 4

Methodology

4.1 Research Problem and Approach Overview

This study presents a virtual sensing solution using machine learning techniques for CO_2 injection pressure prediction by integrating physics-based simulation results with operational measurements. Virtual sensing provides an innovative way of overcoming the existing dilemma associated with highly accurate but highly computational traditional simulators and down-hole measurements required in CCS operations.

Virtual sensing can be defined as a new approach to down-hole monitoring through the use of machine learning techniques to predict "virtual sensors." As opposed to physics-based simulators that involve solving differential equations describing multiphase flow phenomena, virtual sensing relies on data-driven correlations to produce quick estimates of key down-hole variables while maintaining physical correctness.

The four consecutive phases of the proposed method are as follows:

Phase 1: Simulation Tool Evaluation - Systematic assessment of conventional steady-state (PROSPER) and transient (OLGA) simulators for CO_2 injection modeling, establishing the foundation for understanding simulator capabilities and limitations.

Phase 2: Enhanced Dataset Generation - Development of calibrated OLGA models specifically optimized for machine learning training data generation, covering comprehensive operational scenarios while maintaining thermodynamic consistency.

Phase 3: Physics-Informed Preprocessing - Implementation of specialized outlier detection methods that preserve thermodynamic relationships while ensuring high-quality training datasets through systematic data validation procedures.

Phase 4: Hybrid Machine Learning Implementation - Integration of simulation-based and operational data through systematic model development, ensemble methods, and comprehensive validation against real-world CCS operational data.

It is guaranteed that the developed solution will possess physical consistency and achieve the needed efficiency for real-time CCS monitoring applications. Physics are still respected while gaining the operational feasibility of CCS plant monitoring.

4.2 Comparative Analysis of Simulation Tools

The existing petroleum engineering simulators cannot be applied straightforwardly to CO_2 injection systems because of specific thermodynamic properties of supercritical CO_2. This comparative analysis will provide us with the technical basis to justify the development shift to more advanced physics-based solutions followed by ML algorithms usage.

4.2.1 PROSPER Evaluation and Fundamental Limitations

The multiphase flow correlations of PROSPER have significant restrictions while applying to CO_2 injection wells. The validation process started with implementing the Beggs and Brill correlation for the water-methane system to identify the behavior of the correlation and ensure mathematical correctness (full mathematical derivation in Appendix A).

The independent mathematical derivation of 34 cases representing three different pressure zones (low: 25-50 bara; medium: 100-120 bara; high: 160-180 bara) produced satisfactory results compared to PROSPER's Beggs and Brill correlation, verifying the accuracy of the mathematical implementation. Comparison to other PROSPER correlations (Petroleum Expert 2, 3, 5, Mukherjee-Brill) produced results consistent with expectations.

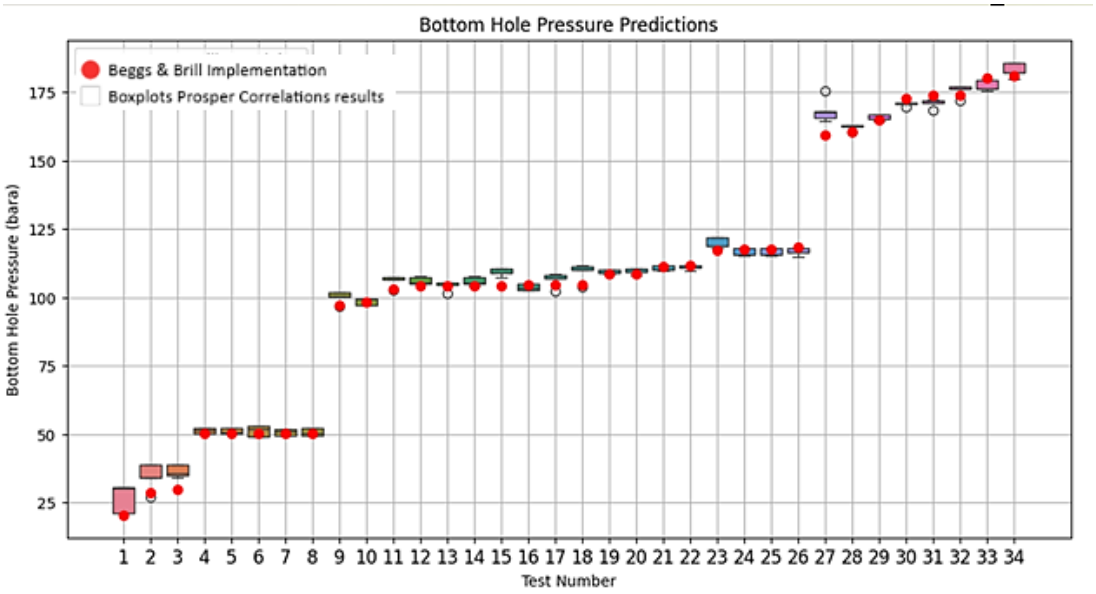

Figure 4.1: Comparison of bottom hole pressure predictions across 34 test cases. Red dots represent results from the implemented Beggs & Brill correlation, while box plots show the distribution of predictions from five different correlations in PROSPER software. Test cases are grouped into three distinct pressure regimes: low pressure (Tests 1-8, ∼25-50 bara), medium pressure (Tests 9-24, ∼100-120 bara), and high pressure (Tests 25-34, ∼160-180 bara).

The application of the methodology on the traditional multiphase flow correlations is shown in Figure 4.1. It can be observed that the red dots representing the Beggs & Brill correlation show consistency in the correlation results generated by the program throughout various pressure levels. Box plot comparison shows the variation among several PROSPER correlations with smaller variation at medium pressures and larger variation

at extremely high and low pressures, which suggests the degree of uncertainty associated with the use of correlations in fluid systems.

However, the analysis using real CO_2 injection field data from the Ravenna CCS Phase 1 project showed the following three major problems, which make the software incompatible with supercritical CO_2 systems:

Extensive Pressure Over-Prediction: PROSPER predicted the flowing bottom hole pressure (FBHP) to be 73% higher than the actual value as measured by downhole gauges. This significant over-prediction is a prediction failure that would prevent the proper calculation of the optimal injection rate, inaccurate wellhead pressure values, and potential hazards due to incorrect pressure calculations.

Discrepancies in Temperature Gradients: The temperature gradients generated by the PROSPER model were inconsistent with those recorded by Distributed Temperature Sensing (DTS). It is evident that the model underestimated the temperature gradient compared to downhole conditions, signifying some inherent weaknesses in the heat transfer model used to simulate CO_2 processes. The inability of the model to accurately predict the temperatures will affect its ability to predict other properties such as pressure and density.

Instability in VLP: The VLP curve showed non-continuous pressure-rate relationships, which are necessary when conducting simulations. The instability of the VLP will make it difficult to predict the appropriate injection rate and pressure required.

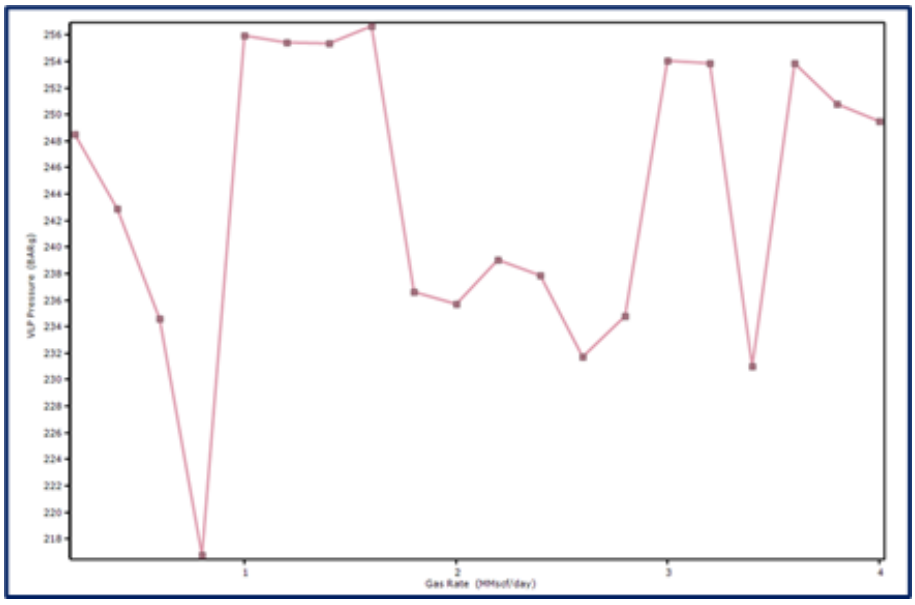

Figure 4.2: VLP curve instability in PROSPER for CO_2 injection systems. Instead of following a smooth trajectory, the curve progressed in a broken manner and reached high pressure over short time intervals, making the simulator unsuitable for operational forecasting.

As shown by Figure 4.2, it is evident how the basic premise of correlation-based models breaks down when applied to CO_2 injection processes. The figure depicts the erratic, unpredictable behavior in the pressure vs. rate relationship instead of the steady curves required for operations. These irregular discontinuities in the pressure-rate curve highlight

how the mathematical singularities in the correlations used by PROSPER prevent the simulator from accurately modeling supercritical CO_2, rendering the tool incapable of being used for injection well optimization purposes.

From analyzing the failed results, it becomes evident that the cause of the failure is due to fundamental mathematical weaknesses in the use of correlations in supercritical CO_2. Multiphase flow correlations make the assumption that fluid properties will exhibit typical behaviors found in standard oilfield environments. However, CO_2 injection wells operate at critical pressures and temperatures ($T_c = 31.1°C$, $P_c = 73.8$ bar). As such, these systems can cause singularity problems within the mathematics used within the correlations, as explained in Appendix A.

Such systematic inconsistencies illustrate how steady-state simulators meant for hydrocarbon applications are inherently inconsistent with CO_2 injection projects, thus requiring a different kind of simulator to properly handle supercritical CO_2 thermodynamics. The calibration study against the Ravenna field data provided quantitative insight into the limitations of PROSPER in CO_2 injection scenarios through systematic comparisons between predictions and observations in terms of pressure and temperature.

The calibration study against the Ravenna field data provided quantitative insight into the limitations of PROSPER in CO_2 injection scenarios through systematic comparisons between predictions and observations in terms of pressure and temperature.

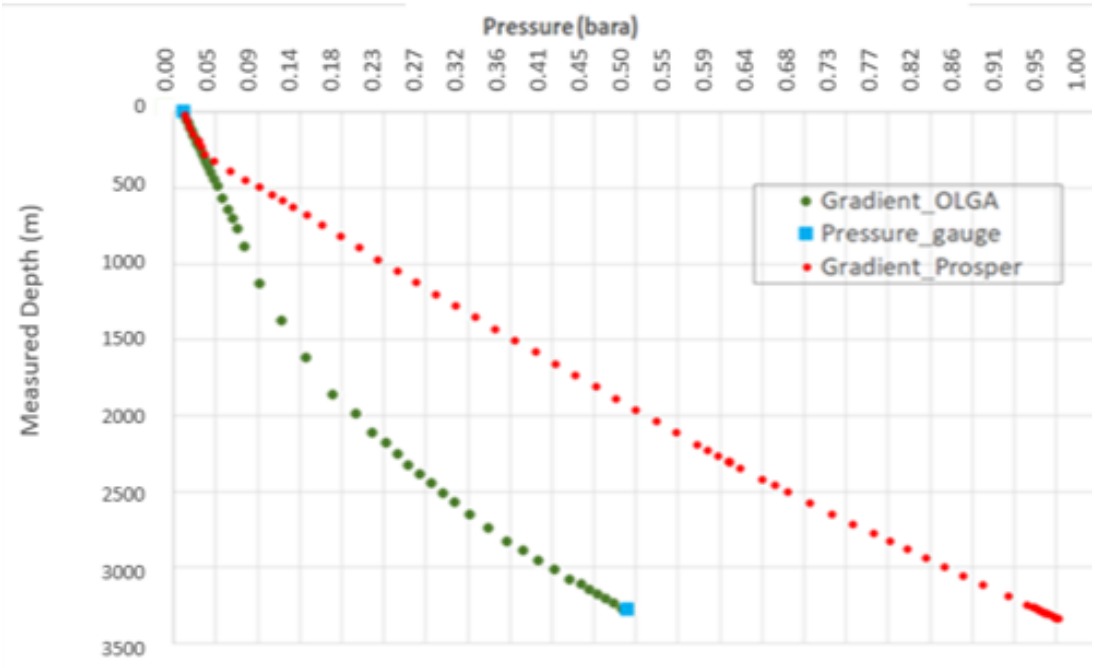

Figure 4.3: PROSPER pressure gradient calibration results against Ravenna CCS Phase 1 field data. The comparison reveals >75% overestimation of FBHP by PROSPER compared to downhole gauge measurements, with systematic deviation from both OLGA simulations and actual field measurements throughout the wellbore depth.

Figure 4.3 quantifies the severe limitations of PROSPER for CO_2 injection modeling through systematic field data comparison. The plot reveals that PROSPER consistently

overestimates pressure throughout the wellbore depth by more than 75%, with the deviation increasing significantly compared to both OLGA simulations and actual downhole gauge measurements. The systematic nature of this error, evident across the entire pressure profile, indicates fundamental thermodynamic modeling inadequacies rather than simple calibration issues.

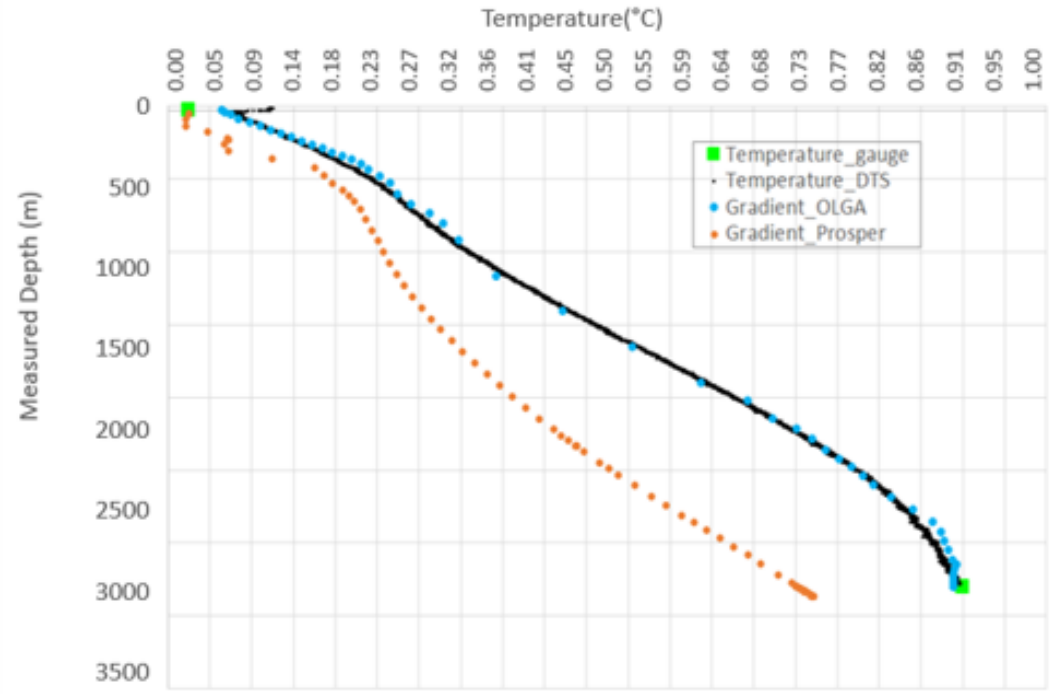

Figure 4.4: PROSPER temperature gradient calibration results showing significant deviation from DTS measurements. PROSPER consistently predicts lower temperatures than actual downhole conditions, indicating fundamental problems in heat transfer modeling for CO_2 systems with cascading effects on density calculations and overall system behavior representation.

Further evidence of modeling deficiencies in PROSPER's core algorithm is shown in Fig. 4.4, based on the temperature gradient comparison. Similar to the pressure prediction deviations, wellbore temperatures are consistently underestimated in the model versus Distributed Temperature Sensing (DTS). These deviations occur systematically in relation to DTS measurements and illustrate how modeling deficiencies can influence other aspects of the calculation, such as density, and thus affect the overall system behavior.

Fig. 4.3 and 4.4, provides evidence of the systematic character of PROSPER's deficiencies across multiple parameters. The first graph illustrates significant deviation of predicted pressures from physics-based OLGA simulations and down-hole gauge readings in the field. As seen, PROSPER tends to overestimate the required formation pressures for successful CO_2 injection operations. In addition, Fig. 4.4 illustrates the temperature gradient deviations from DTS measurements.

The presence of such systematic errors with respect to both pressure and temperature shows that the deficiencies of PROSPER go beyond mere calibrations but point towards a lack of fundamental compatibility with the CO_2 systems altogether. The consistent nature of errors in all of these different scenarios proves that conventional multiphase flow

correlations cannot be adjusted to work properly with supercritical CO_2 processes via mere calibration. In addition to its academic significance, such discrepancies will also have important implications on practical engineering applications in terms of 75

This evidence suggests a compelling need for a transition from steady-state simulation based on empirical correlations to transient simulations governed by physical laws.

4.2.2 OLGA Advanced Modeling Capabilities

The physics-based model of OLGA based on fundamental conservation equations has greater capacity to simulate supercritical CO_2 systems than the correlation-based simulation model. The advanced mathematical model of the simulator has been developed to overcome the limitations existing in traditional models by means of thermodynamic analysis and numerics.

Fundamental Physics-Based Model: The fundamental conservation equations for mass, momentum, and energy in one-dimensional form are used by OLGA. The solution of the set of differential equations allows modeling thermodynamic effects associated with supercritical CO_2 systems without mathematical restrictions of correlation-based models.

The governing equations consist of the mass conservation equation (2.1), the momentum conservation equation (2.2), and the energy conservation equation (2.3), as detailed in the numerical approach section.

Complex Thermodynamic Description: In the current version of OLGA, the thermodynamic description uses the well-known equation of state by Span and Wagner [23] specially designed for CO_2 systems. This allows performing a reliable computation of thermodynamic properties of the fluids under any conditions corresponding to the entire scope of CCS applications.

Joule-Thomson Effects Implementation: The representation of the Joule-Thomson effects for the systems of CO_2 is one of the key aspects of OLGA. In particular, the Joule-Thomson effects are essential for describing the processes that happen in CO_2 wells due to considerable temperature variations associated with pressure changes.

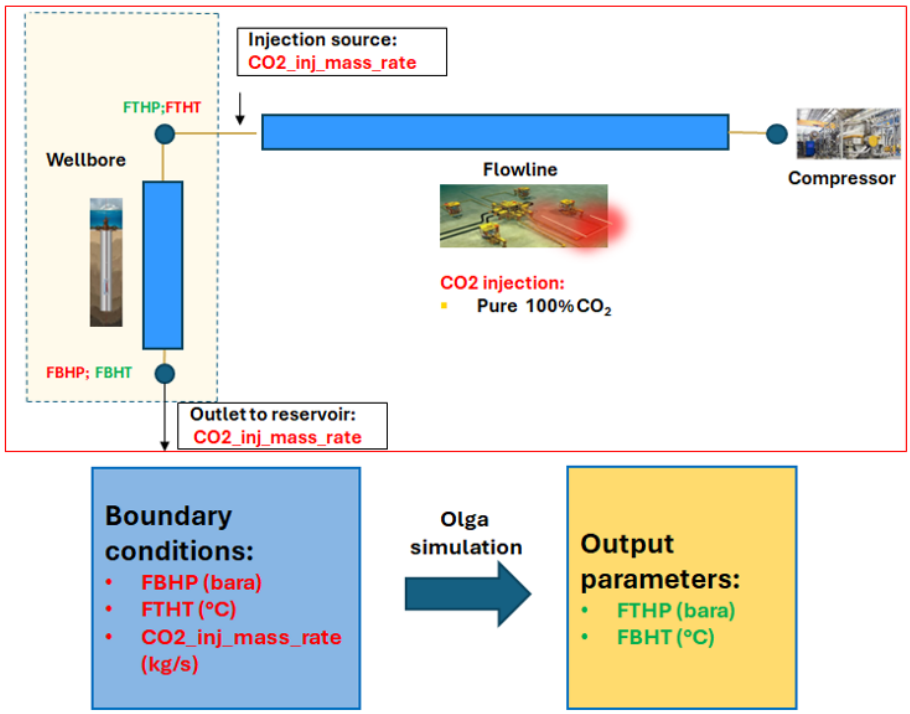

Figure 20: OLGA configuration used to simulate steady state conditions

Figure 4.5: OLGA configuration used to simulate steady-state conditions for CO_2 injection systems. The model incorporates detailed wellbore geometry, completion components, and thermal boundary conditions.

Figure 4.5 illustrates the comprehensive modeling approach implemented in OLGA for CO_2 injection simulation. The configuration incorporates detailed wellbore geometry, completion components, and thermal boundary conditions necessary for accurate supercritical CO_2 behavior modeling. This physics-based approach, utilizing fundamental conservation equations rather than empirical correlations, enables accurate representation of the complex thermodynamic processes characteristic of CO_2 injection systems.

4.3 Enhanced OLGA Model Development

The development of an enhanced OLGA model specifically calibrated for CO_2 injection applications represents a critical component of the methodology, providing the physics-based foundation for subsequent machine learning model training. This section details the systematic calibration protocol, validation procedures, and optimization strategies developed to create high-quality training datasets.

4.3.1 Ravenna CCS Phase 1 Calibration Protocol

The documented failures of PROSPER's correlation-based approach necessitated transitioning to physics-based simulation methodologies capable of accurately representing supercritical CO_2 behavior. OLGA's advanced mathematical framework, employing fundamental conservation equations rather than empirical correlations, provides the sophisticated thermodynamic modeling required for CO_2 injection systems operating near critical conditions.

However, in order to adopt OLGA for machine learning purposes, a specific calibration approach had to be adopted that was specifically designed for the generation of training data rather than for simulation precision. It was important to note that whereas machine learning algorithms are able to correct for model biases, systematic deviations are needed to ensure proper learning rather than random fluctuations. Therefore, the calibration process aimed at achieving consistent, predictable relations among all possible scenarios while maintaining thermodynamic consistency.

OLGA was calibrated using the operational data obtained from the Ravenna CCS Phase 1, which marked the first industrial-scale CO_2 injection system implemented in the Mediterranean region. The project involved the CO_2 injection into the Porto Corsini Mare Ovest (PCMW) depleted gas field situated within the Adriatic Sea, whereby the injection took place via the well up to 3000m.

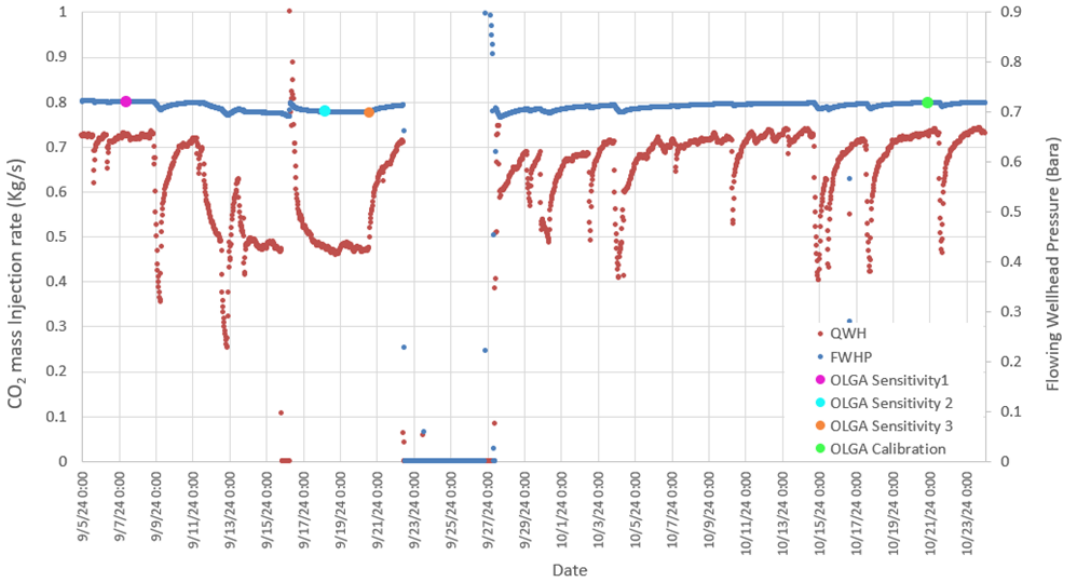

Figure 4.6: Selected sensitivity analysis points during CO_2 injection operation showing pressure validation across multiple operational scenarios. The system operates with sustained injection, facilitating stable conditions with steady wellhead pressures.

Figure 4.6 demonstrates the systematic validation approach applied across multiple operational scenarios from Ravenna CCS operations. The plot shows sustained injection periods with stable wellhead pressures, providing diverse validation points for model calibration. The consistent pressure levels across different operational periods validate the enhanced OLGA model's ability to maintain stable convergence under varying injection conditions, essential for generating reliable training datasets.

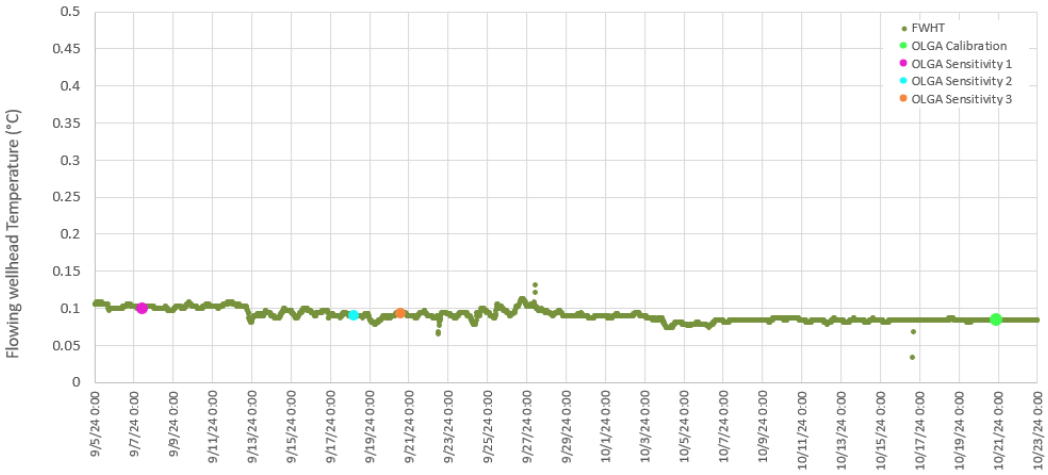

Figure 4.7: Selected sensitivity analysis points for wellhead temperature validation. The flowing wellhead temperature remains constant with minimal variations across different operational periods.

Figure 4.7 shows the wellhead temperature validation data used for model calibration, revealing relatively constant flowing temperatures with minimal variations across operational periods. The stable temperature profile demonstrated in the plot provides reliable reference conditions for thermal modeling validation, supporting the systematic calibration approach that prioritizes consistent, learnable patterns over absolute accuracy for individual validation cases.

Systematic Deviation Strategy: The calibration approach prioritizes establishing consistent, predictable deviation patterns across all operational scenarios rather than achieving perfect accuracy for individual validation cases. This methodology addresses a fundamental challenge in machine learning applications where systematic model bias can be identified and compensated for by ML algorithms through pattern recognition, whereas inconsistent simulation deviations create unpredictable training data that prevents effective learning.

4.3.2 Density Coefficient Optimization

The calibration process prioritized density coefficient optimization over detailed temperature calibration, reflecting both the physical importance of accurate density predictions and the relative complexity of each calibration process. In CO_2 injection wells, pressure drop consists of hydrostatic and frictional components, where the hydrostatic component directly depends on fluid density distribution along the wellbore.

Quantitative Calibration Results:

Table 4.1: Density Coefficient Calibration Performance

Validation Case	Before Modification	After Modification
Calibration Point	0.12	0.38
Sensitivity 1	-0.01	0.00
Sensitivity 2	1.00	0.20
Sensitivity 3	0.80	0.15
Range Reduction	**73% Improvement**	

The density coefficient modification achieved substantial improvement in system consistency, reducing the range of deviations by approximately 73%.

4.3.3 Thermal Conductivity Parameter Optimization

Thermal modeling calibration addressed systematic temperature prediction differences between Distributed Temperature Sensing (DTS) measurements and OLGA simulations across the entire wellbore depth.

Table 4.2: Thermal Modeling Performance Improvement

Validation Case	Before (°C RMSE)	After (°C RMSE)	Improvement
Calibration Case	1.105	0.55	50%
Sensitivity 1	1.28	0.57	55%
Sensitivity 2	1.06	0.55	48%
Sensitivity 3	0.64	0.42	34%
Average Improvement		**47%**	

4.3.4 Calibration Results and Validation

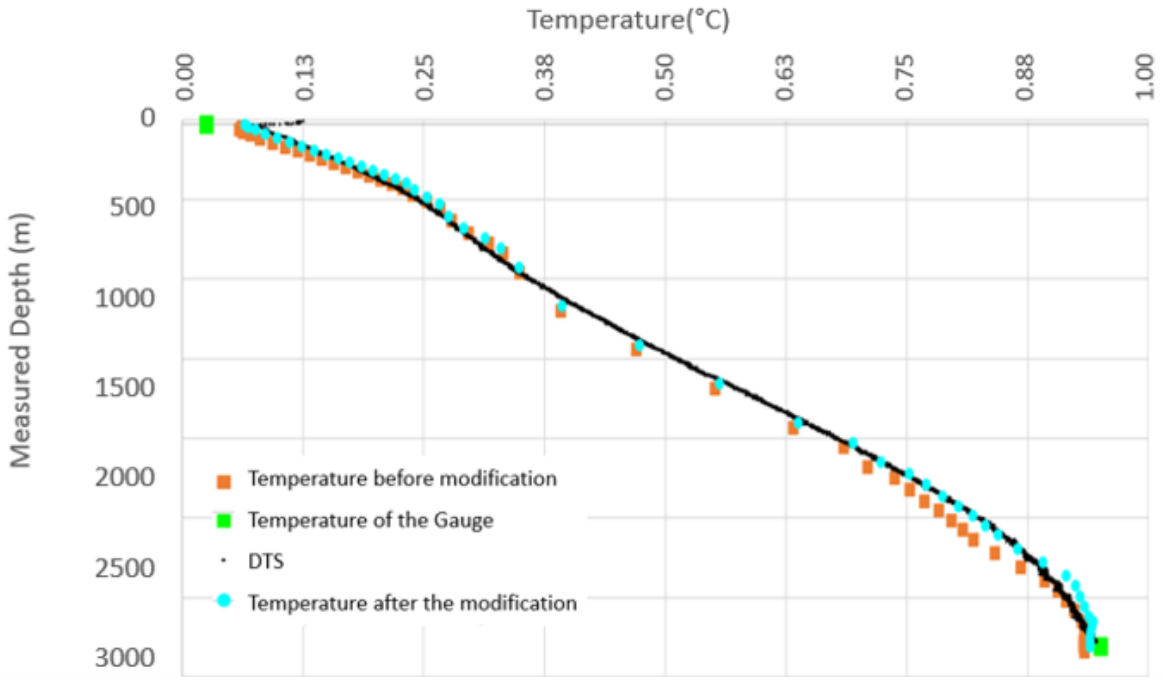

Figure 4.8: OLGA model calibration methodology results showing systematic improvement in temperature prediction accuracy. The comparison displays temperature profiles before modification (orange), after thermal conductivity optimization (blue), actual gauge measurements (green), and DTS measurements (black dots). The calibration achieved 50% RMSE improvement in temperature prediction while maintaining consistent, learnable patterns essential for machine learning training data generation. *Temperature values are normalized to [0,1] range for confidentiality purposes.

Figure 4.8 demonstrates the effectiveness of the systematic calibration approach developed specifically for machine learning applications. The thermal conductivity parameter optimization achieved substantial quantitative improvements in temperature prediction accuracy, reducing RMSE from the original OLGA predictions to within 0.5°C of DTS measurements throughout the wellbore depth.

Calibration Achievement Summary:

- **Density Calibration:** Systematic bias correction achieved 73% improvement in pressure prediction consistency across validation scenarios
- **Temperature Optimization:** 50% RMSE improvement in temperature gradient predictions compared to DTS measurements
- **Field Validation:** Consistent performance across multiple operational scenarios from Ravenna Phase 1 operations
- **Training Data Quality:** Generated consistent, learnable patterns suitable for robust machine learning model development

The successful implementation of the calibrated model in reproducing the complex physics associated with CO_2 injection allows us to keep the systematic errors which are necessary

for machine learning training. The excellent match between OLGA predictions and real data obtained during the whole well depth confirms the ability of the improved model to describe the operational physics.

The successful calibration of the OLGA model is the basis for obtaining an appropriate set of data for efficient training of ML models based on physics of the problem. In other words, the developed capability solves the primary problem formulated by PROSPER Evaluation Team as the need for simulations able to capture supercritical CO_2 physics.

4.3.5 Comprehensive Dataset Generation

Following successful calibration and validation, the enhanced OLGA model generated a comprehensive training dataset covering realistic operational scenarios for CO_2 injection systems. This approach resulted in 2,100 individual simulation cases executed using automated batch processing procedures with consistent convergence criteria.

4.4 Physics-Informed Data Preprocessing

Data preprocessing ensured training dataset quality through specialized outlier detection and validation procedures designed specifically for CO_2 injection systems. This phase combines traditional statistical approaches with physics-informed validation techniques to ensure high-quality training data while preserving essential thermodynamic relationships.

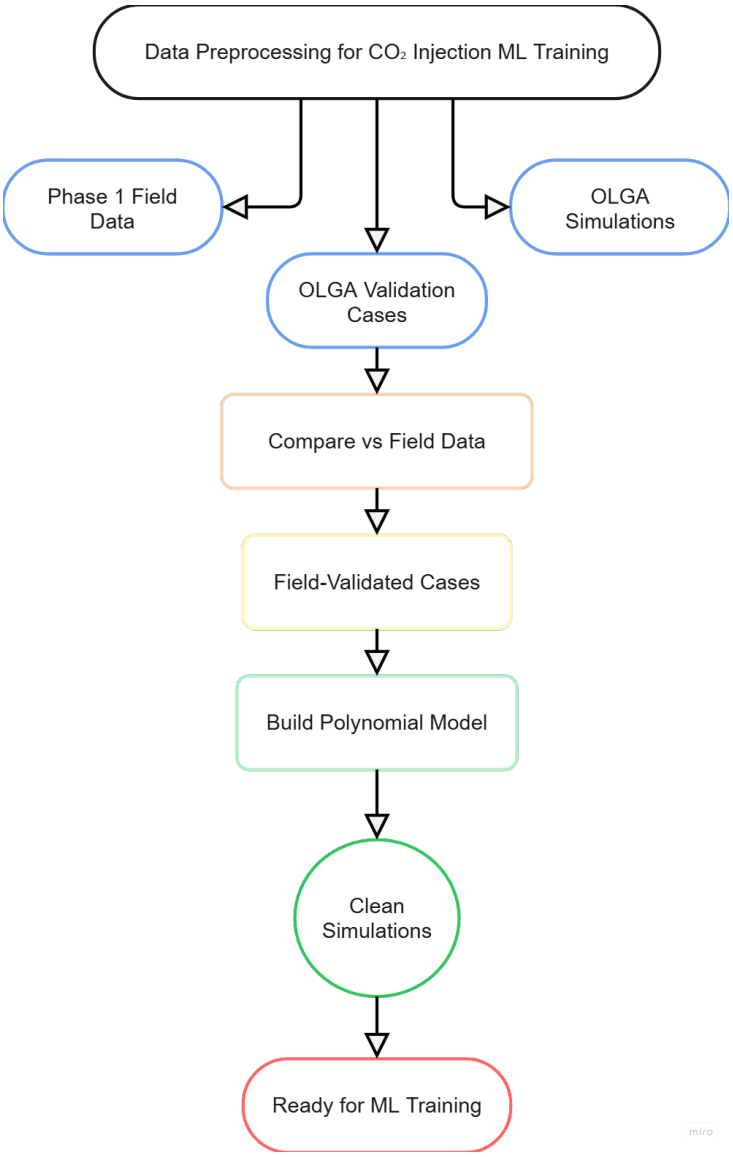

Figure 4.9: Workflow for preprocessing CO_2 injection data for machine learning training. The process includes systematic data loading, outlier detection, physics-based validation, and final dataset preparation.

Figure 4.9 outlines the systematic approach developed for ensuring training dataset quality in CO_2 injection applications. The workflow integrates data loading, physics-based outlier detection, thermodynamic validation, and final dataset preparation in a structured process specifically designed for engineering simulation data. This comprehensive preprocessing framework ensures that machine learning models train on physically consistent, high-quality data while preserving essential thermodynamic relationships.

4.4.1 Two-Stage Outlier Detection Methodology

Standard statistical outlier detection methods prove inadequate for engineering simulation data where physical consistency is paramount. This research developed a specialized preprocessing framework that systematically combines thermodynamic validation with statistical robustness, ensuring that outlier removal preserves essential physical relationships while eliminating problematic data points.

4.4.2 Stage 1: Physics-Based Validation

The physics-based validation stage employed a 20-point OLGA validation dataset representing specific operational conditions from Ravenna Phase 1 field operations. Four validation points were identified as outliers based on their systematic deviations from expected physical behavior:

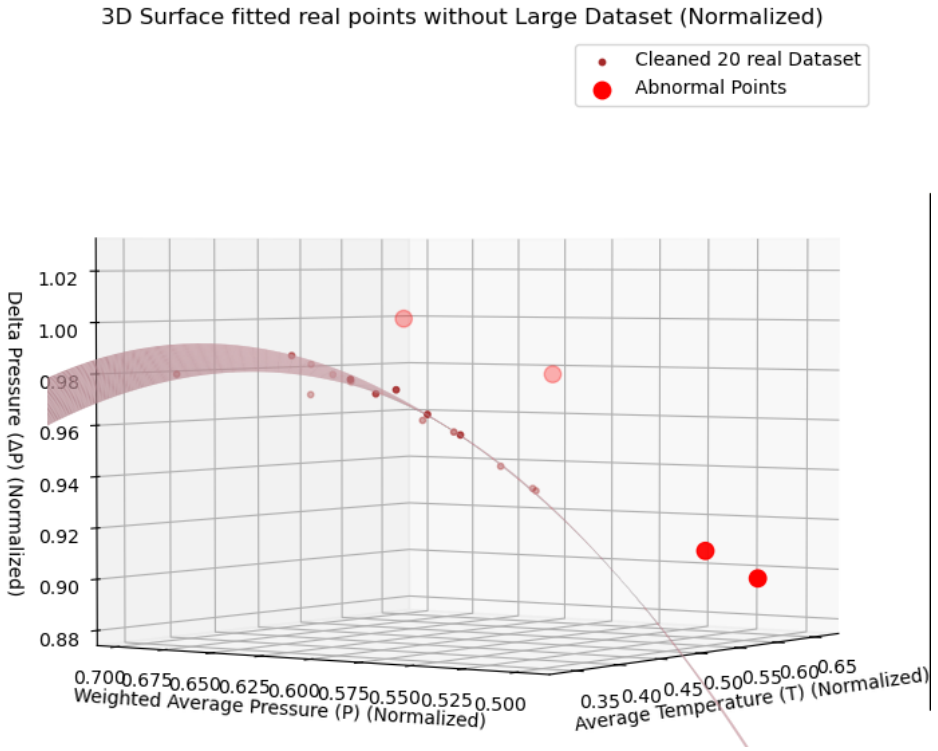

Figure 4.10: 3D plot showing outliers identified through comprehensive thermodynamic analysis. The visualization plots weighted average pressure, average temperature, and delta pressure, revealing outliers that exhibit inconsistent thermodynamic behavior when evaluated across the complete operational parameter space.

Figure 4.10 visualizes the physics-based outlier identification methodology through three-dimensional thermodynamic analysis. The plot reveals four outlier cases that exhibit inconsistent behavior when evaluated across weighted average pressure, average temperature, and delta pressure dimensions. The clustering of normal validation points demonstrates the expected thermodynamic relationships, while the isolated outlier points indicate systematic deviations from established physical behavior that could compromise machine learning model training effectiveness.

Issue	**Explanation**
Starter cooling	OLGA simulation cooler than expected
Material imbalance	High RMSE makes model unpredictable
Temperature deviation	OLGA simulation shows cooling effects

Table 4.3: Summary of simulation issues

4.4.3 Stage 2: Surface-Based Analysis for Large Dataset

With 16 validated OLGA points established as a quality reference, the challenge became cleaning the full 2,100-point simulation dataset. A physics-informed approach was developed using the validated points as a quality standard for identifying problematic simulations.

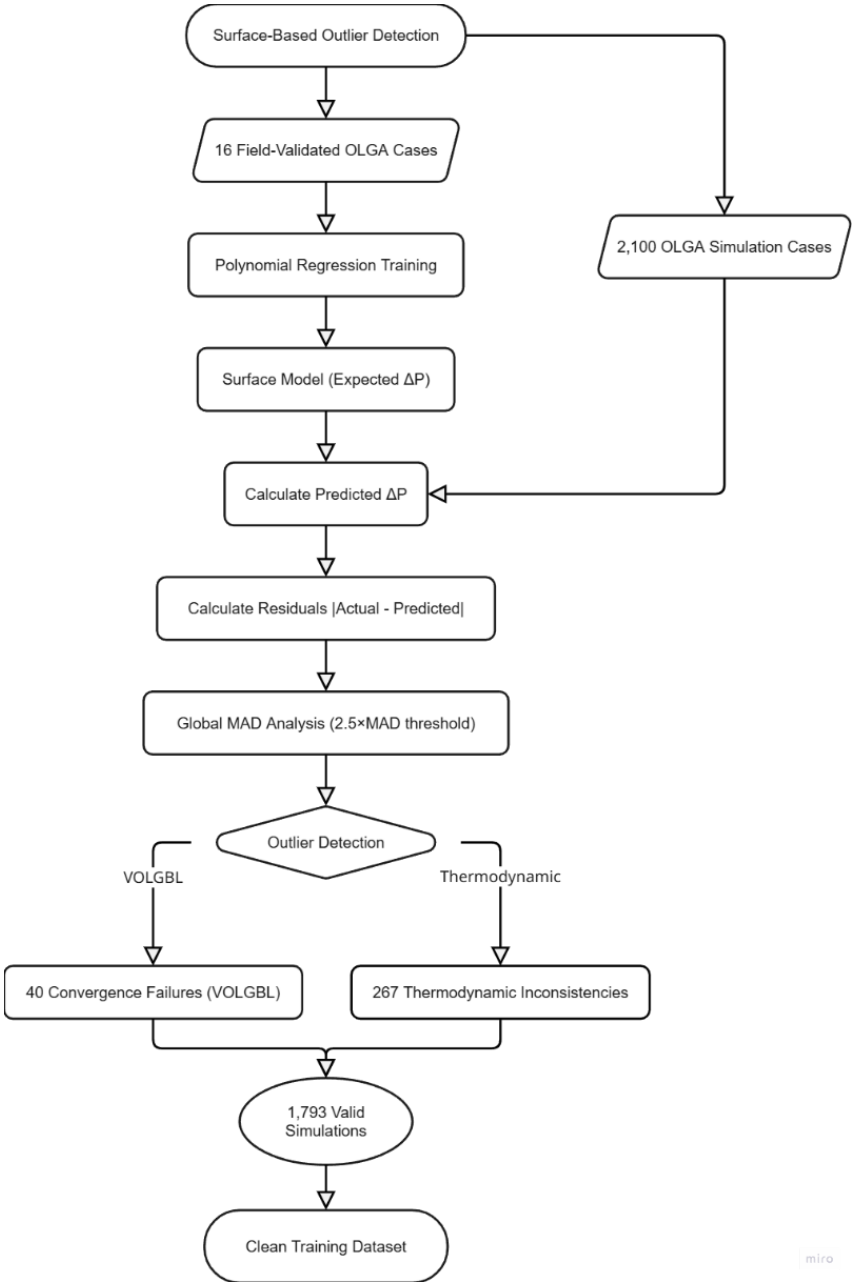

Figure 4.11: Workflow for surface-based outlier detection in reservoir simulation data. The process uses validated OLGA points to establish a quality standard for identifying problematic simulations in the large dataset through polynomial regression surface fitting and residual analysis.

Figure 4.11 details the systematic approach for extending physics-based validation to large simulation datasets. The workflow utilizes validated OLGA points as quality standards for establishing expected thermodynamic relationships, then applies polynomial regression

surface fitting to identify problematic simulations in the complete 2,100-point dataset. This scalable methodology enables efficient quality control while maintaining physics-based validation principles essential for CCS applications.

4.4.4 Polynomial Regression Surface Development

Using the 16 validated OLGA simulation cases accurately representing Ravenna Phase 1 field physics, a second-degree polynomial regression model was developed establishing the fundamental thermodynamic relationship:

$$\Delta P = f(\text{Weighted Average Temperature}, \text{Weighted Average Pressure}) \tag{4.1}$$

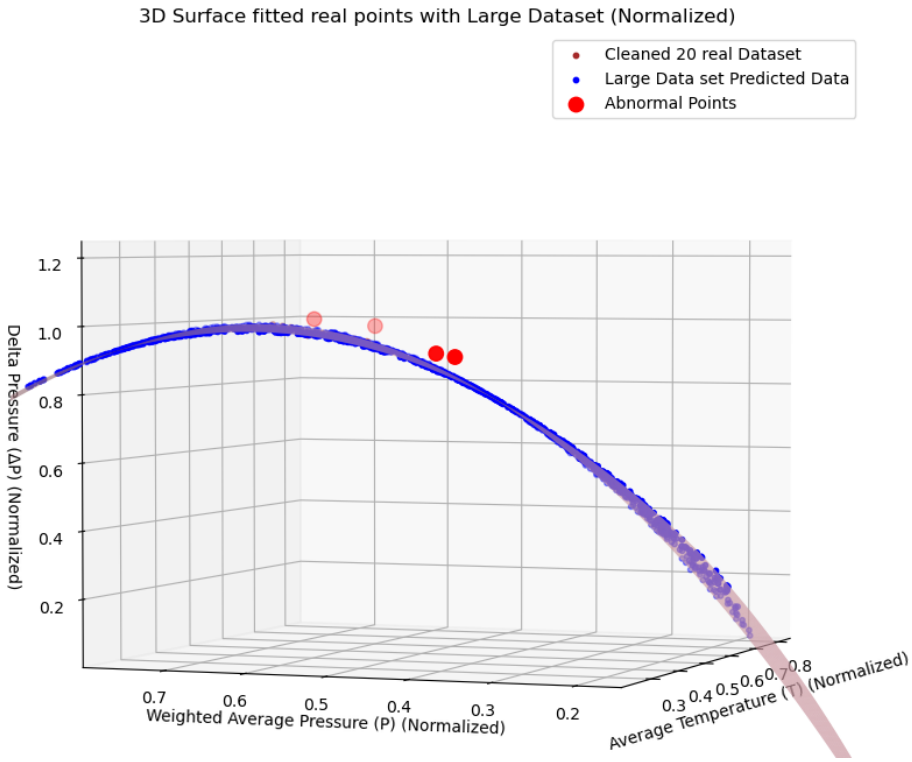

Figure 4.12: 3D surface fit demonstrating outlier detection methodology applied to full OLGA simulation dataset. The polynomial regression model establishes expected thermodynamic relationships based on field-validated simulation points, enabling identification of problematic simulations that deviate from established physical behavior.

Figure 4.12 demonstrates the polynomial regression surface methodology applied to the complete OLGA simulation dataset. The plot shows how field-validated simulation points establish expected thermodynamic relationships between weighted average temperature, weighted average pressure, and pressure drop. Points deviating significantly from this physics-based surface represent problematic simulations that exhibit behavior inconsistent with established field validation, enabling systematic identification and removal of unsuitable training data.

4.4.5 Final Training Dataset Characteristics

After comprehensive preprocessing, 1,793 simulations remained (85.4 retention rate), ensuring physics-based consistency, numerical stability, operational representativeness, and machine learning suitability.

4.5 Theoretical Foundations of Machine Learning in CCS Applications

Machine Learning represents a fundamental shift from traditional engineering systems that rely on explicitly programmed rules and formulas. Instead of programming for each situation, the system learns to generalize the data and apply the learned rules to new unseen cases.

CO_2 injection systems are characterized by problems that lend themselves well to Machine Learning approaches:

- Near-critical CO_2 exhibits extremely nonlinear relations between pressure, temperature, and density which are hard to predict using correlations
- Decision making must take place rapidly (within seconds to minutes), while physics-based simulations need to be done slowly (hours to days)
- Several variables act together and must be considered at the same time to understand the behavior of the system
- Conditions vary constantly while operation is taking place

4.5.1 The Learning Process in CCS Applications

The machine learning process for CCS pressure prediction follows a systematic approach tailored for engineering applications. Data collection combines historical operational data from surface and downhole measurements with physics-based simulation outputs from calibrated OLGA models. This hybrid approach leverages both empirical operational knowledge and theoretical understanding of CO_2 thermodynamics.

Feature engineering transforms raw measurements into meaningful predictors that capture the underlying physics of supercritical CO_2 systems. This process involves generating new features reflecting important relationships (e.g., pressure-temperature interactions) and scaling features to ensure equal contribution to model training.

4.5.2 Hyperparameter Optimization and Model Selection

Hyperparameters are configuration settings that control how machine learning algorithms learn, but cannot be learned automatically from data. These settings must be specified prior to training and significantly influence model performance, complexity, and generalization capability.

The Bias-Variance Tradeoff: Hyperparameter selection controls the fundamental bias-variance tradeoff in machine learning:

- **High bias (underfitting):** Models too simple to capture patterns in data, yielding poor predictions on both training and test sets
- **High variance (overfitting):** Models too complex, memorizing training data noise and achieving perfect training performance but poor testing performance
- **Optimal balance:** Achieves appropriate model complexity for underlying relationships, performing well on both training and test data

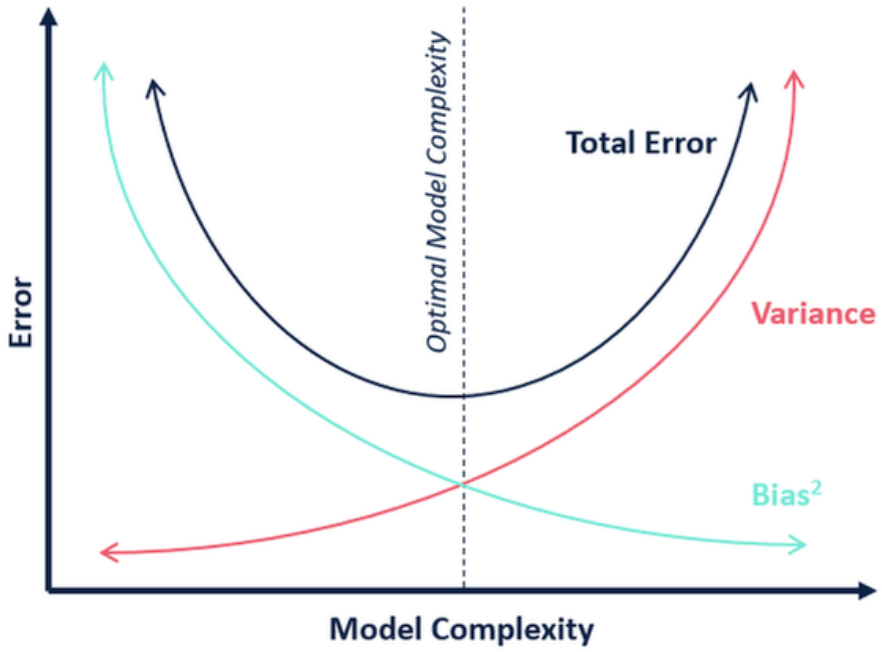

Figure 4.13: Bias-variance tradeoff illustration showing underfitting, optimal fit, and overfitting scenarios for CCS pressure prediction. The optimal model balances complexity to capture essential patterns without memorizing noise.

Grid Search Hyperparameter Optimization: Grid search systematically evaluates all possible combinations of hyperparameter values within predefined ranges. The process:

1. Define parameter grids containing all hyperparameters to tune and their candidate values
2. Systematically evaluate every parameter combination using cross-validation
3. Select optimal parameters based on validation performance metrics
4. Apply best parameters to final model training

Cross-Validation Implementation: The methodology employs 5-fold cross-validation, dividing training data into five equal sections. Four folds train the model while one fold validates performance, rotating through all combinations. This provides statistically robust performance estimates while preventing overfitting to specific data subsets.

4.6 Machine Learning Implementation

The machine learning solution takes into account the highly nonlinear dependencies that exist between various parameters involved in CO_2 injection processes through proper selection of algorithms, their optimization, and forming ensembles. This part describes how such models can be developed in full detail specifically designed to be used within CCS operations.

4.6.1 Problem Formulation and Algorithms Selection

CO_2 injection pressure prediction is stated as a supervised learning problem of regression predicting continuous values of two physical quantities – FBHP and FBHT. Regression algorithms are a perfect choice for pressure prediction as target classes belong to the set of numerical values and need to be predicted numerically.

Algorithms Selected and Reasons Specific to CCS Problem:

Support Vector Regression (SVR): Selected due to its kernel-based nonlinearity processing abilities, especially RBF type of kernel to model nonlinear dependency of properties of CO_2. SVR is good at working with high dimensional spaces of features, which corresponds well to supercritical CO_2.

Random Forest: Selected due to intrinsic overfitting prevention by bootstrap aggregation, as well as ability to perform feature importance ranking, which can provide useful insight about physical relations of features.

Gradient Boosting: Utilized due to its sequential learning method, which facilitates the correction of residuals in errors, especially beneficial for detecting fine nuances in the behavior of CO_2 injections that individual models may overlook.

XGBoost: Chosen for sophisticated gradient boosting techniques featuring L1/L2 regularization, enhancing generalization skills critical for CCS tasks, where models need to operate effectively under various conditions.

4.6.2 Hyperparameters Optimization Methodology

The hyperparameters optimization methodology centers on the key parameters for each algorithm, determined through an exhaustive literature survey of machine learning applications in engineering and petroleum systems. Originally, the theoretical range of possible values for hyperparameters included more than 23,000 different combinations in total, considering full factorial grid search. After literature-based hyperparameter selection and initial experiments, this vast parameter space was narrowed down to only 150 unique combinations centered around hyperparameters' optimal ranges.

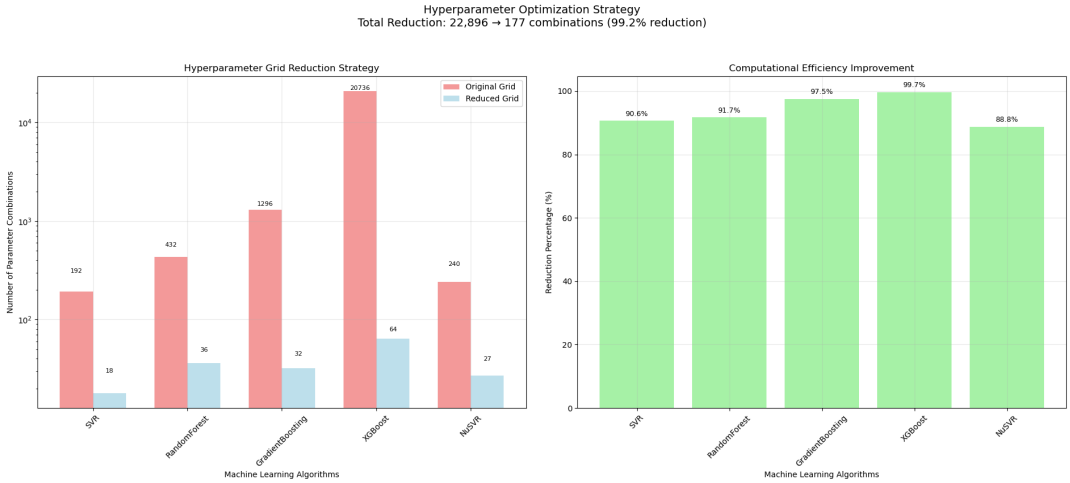

Figure 4.14: Hyperparameter optimization grid showing parameter space exploration for SVR, Random Forest, and XGBoost algorithms. The systematic reduction from 23,000 to 150 combinations maintains optimization effectiveness while achieving computational efficiency.

Figure 4.14 visualizes the systematic reduction of parameter space exploration from approximately 23,000 to 150 carefully selected combinations. The plot demonstrates how literature-based parameter selection focuses optimization efforts on hyperparameter ranges most likely to yield optimal performance for engineering applications. This strategic approach maintains optimization effectiveness while achieving computational efficiency essential for practical CCS model development.

4.6.3 Performance Evaluation Framework

For the evaluation of performance, the following three metrics will be used, as they provide an assessment of both accuracy and reliability of predictive models:

Mean Absolute Error (MAE): The main evaluation metric, chosen due to high transparency and ability to interpret the result in the engineering scale despite possible failures in CCS operation.

Root Mean Squared Error (RMSE): The average value of prediction error, which is highly sensitive to outliers in order to prevent models from making large prediction errors, endangering CCS operation.

Coefficient of Determination (R^2): This coefficient shows what percentage of the variation of target variables is explained by the model predictions.

4.6.4 Machine Learning Algorithms for CCS Pressure Prediction

4.6.4.1 Linear Models for Pressure Prediction

Linear regression assumes relationships between input variables and targets can be represented as linear combinations:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{4.2}$$

where β represents model coefficients and ε denotes the error term.

Polynomial regression extends linear regression by adding polynomial terms to capture nonlinear relationships:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \ldots + \varepsilon \tag{4.3}$$

Lasso (Least Absolute Shrinkage and Selection Operator) regression adds L1 penalty for automatic feature selection:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \alpha \sum_{j=1}^{p} |\beta_j| \tag{4.4}$$

4.6.4.2 Support Vector Regression: Advanced Nonlinear Modeling

Support Vector Regression extends support vector machines to regression problems, finding functions that predict target variables while balancing model complexity and prediction accuracy. The core innovation lies in the kernel trick, transforming input space into higher-dimensional space where complex nonlinear relationships become linear.

The SVR optimization problem:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{4.5}$$

subject to:

$$y_i - w^T \phi(x_i) - b \leq \varepsilon + \xi_i \tag{4.6}$$

$$w^T \phi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \tag{4.7}$$

$$\xi_i, \xi_i^* \geq 0 \tag{4.8}$$

where $\phi(x_i)$ represents the kernel transformation, C controls regularization strength, and ε defines the insensitive zone width.

Radial Basis Function (RBF) Kernel:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{4.9}$$

RBF kernels prove particularly effective for CO_2 systems because they can capture exponential density changes around critical points, subtle pressure-temperature coupling in supercritical regime, and nonlinear flow behavior under varied conditions.

4.6.4.3 Tree-Based Ensemble Methods

Random Forest builds multiple decision trees using bootstrap sampling and random feature selection, combining predictions through averaging:

$$\hat{y}_{RF} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}) \tag{4.10}$$

where B represents the number of trees and T_b denotes individual tree predictions.

Gradient boosting builds ensembles sequentially, with each new model correcting errors made by previous models:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}) \tag{4.11}$$

XGBoost enhances gradient boosting with advanced regularization and optimization:

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4.12}$$

where $\Omega(f_t)$ represents regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4.13}$$

4.6.4.4 Ensemble Methods: Combining Multiple Models

Voting ensembles combine multiple base models through simple averaging:

$$\hat{y}_{vote} = \frac{1}{M}\sum_{m=1}^{M} \hat{y}_m \tag{4.14}$$

where M represents the number of base models and $\hat{y}_m$ denotes individual model predictions.

Stacking employs two-level architecture where meta-learners learn optimal combination strategies:

$$\hat{y}_{stack} = g(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M) \tag{4.15}$$

where g represents the meta-learning function trained on base model predictions.

4.6.5 Advanced Feature Engineering for CCS Applications

Feature engineering incorporates CCS domain knowledge to create meaningful predictors capturing underlying physics of supercritical CO_2 systems. This process recognizes that CO_2 injection operations involve complex thermodynamic behaviors differing significantly from conventional hydrocarbon systems.

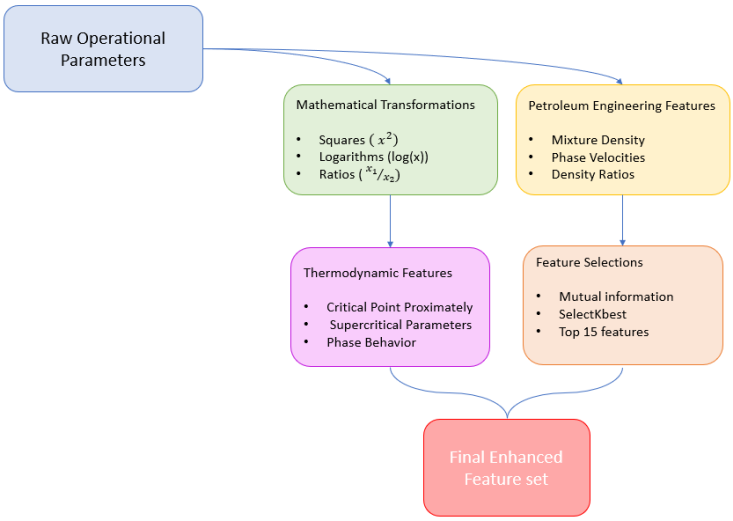

Figure 4.15: Feature engineering process flowchart showing transformation of raw operational parameters into CCS-specific features. The process includes thermodynamic features, phase behavior indicators, and mathematical transformations.

Figure 4.15 outlines the comprehensive transformation process for converting raw operational parameters into meaningful predictors for CCS applications. The flowchart shows how thermodynamic features, phase behavior indicators, and mathematical transformations systematically capture the complex physics of supercritical CO_2 systems. This domain-informed approach ensures that machine learning models have access to features that reflect the underlying physical processes governing CO_2 injection behavior.

4.6.5.1 CCS-Specific Feature Development

Thermodynamic Features: Focus on density variations significantly impacting pressure drops throughout injection systems:

Density-Based Features:

$$\rho_{effective} = f(P_{avg}, T_{avg}, \text{CO}_2 \text{ composition}) \tag{4.16}$$

Critical Point Proximity:

$$\Delta P_{critical} = |P_{system} - P_{critical}|, \quad \Delta T_{critical} = |T_{system} - T_{critical}| \tag{4.17}$$

where $P_{critical} = 73.8$ bar and $T_{critical} = 31.1$řC for CO_2.

Phase Behavior Features: Target CO_2 transitions in supercritical conditions:

Supercritical Parameter:

$$SC_{parameter} = \frac{P \cdot T}{P_{critical} \cdot T_{critical}} \tag{4.18}$$

Reduced Properties:

$$P_r = \frac{P}{P_{critical}} \tag{4.19}$$

$$T_r = \frac{T}{T_{critical}} \tag{4.20}$$

Flow Dynamics Features: Capture complex behavior characteristic of supercritical CO_2:

Reynolds Number:

$$Re = \frac{\rho v D}{\mu} \tag{4.21}$$

Modified Froude Number:

$$Fr_{modified} = \frac{v^2}{gD \cdot \frac{\Delta\rho}{\rho}} \tag{4.22}$$

4.6.5.2 Mathematical Transformations

Mathematical transformations enhance feature spaces by creating variables better representing underlying physical relationships:

Polynomial Expansions:

$$\mathbf{x}_{poly} = [x_1, x_2, x_1^2, x_2^2, x_1 x_2, x_1^3, \ldots] \tag{4.23}$$

Logarithmic Transformations:

$$x_{log} = \log(|x| + 1), \quad x_{log-ratio} = \log\left(\frac{x_1}{x_2 + \epsilon}\right) \tag{4.24}$$

Pressure-Temperature Interactions:

$$PTI = P \cdot T \cdot \exp\left(-\frac{|P - P_{critical}|}{P_{critical}}\right) \tag{4.25}$$

4.6.5.3 Feature Selection Strategy

Feature selection employs mutual information-based approaches:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{4.26}$$

The SelectKBest algorithm identifies features with highest mutual information scores:

$$\text{Selected Features} = \arg\max_k \{MI(X_i;Y)\}_{i=1}^n \tag{4.27}$$

Typically, 15 features provide optimal balance between model complexity and performance for CCS applications.

4.6.6 Ensemble Method Development

The systematic ensemble development process implements multiple ensemble variations to thoroughly explore potential benefits of combining different modeling approaches for CCS pressure prediction.

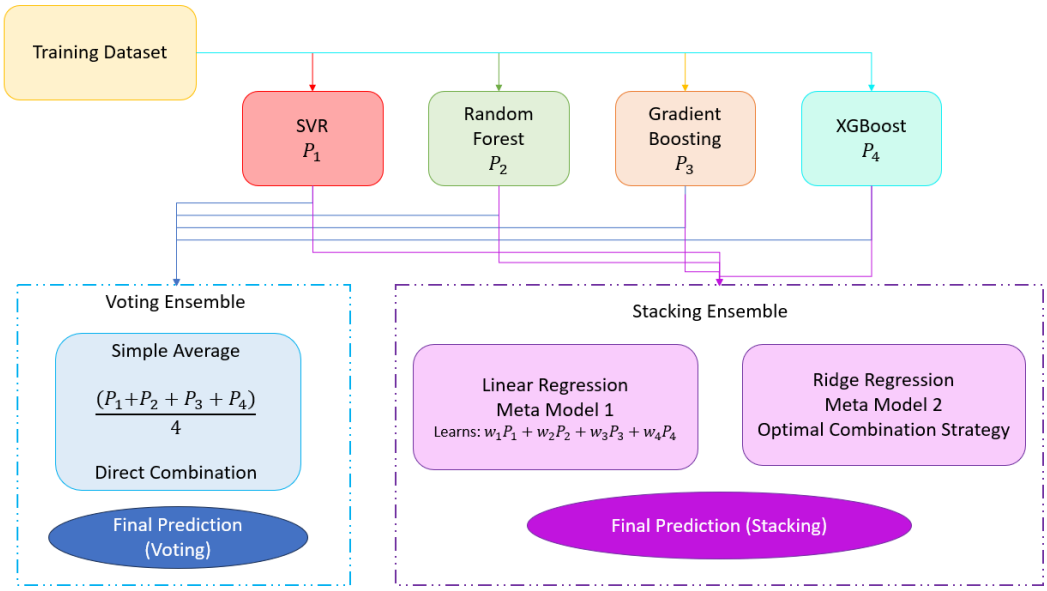

Figure 4.16: Ensemble method architecture showing voting and stacking approaches for combining multiple base models. Level 1 provides diverse predictions while Level 2 meta-learners optimize combination strategies.

The figure below highlights the two-level ensemble framework adopted for predicting CCS pressure. As can be observed, the level one base models are capable of generating predictions using diverse models with varying algorithms and hyperparameters. The meta learners at level two are optimized to combine the outputs generated from the base learners using voting and stacking techniques.

Voting Ensemble Variations:

- **Basic Voting:** Simply averaging out the outputs of the highest performing models by exploiting complementary strengths without being too computationally intensive.
- **Combined Voting:** Involves combining all the models that have been proven effective while taking into account that some of the less effective models may also bring important insights when used with more effective models.
- **Filtered Voting:** Excludes particular models from consideration to determine whether similar results can be obtained through particular algorithm combinations.

Stacking Ensemble Framework: Consists of a two-level approach where base models utilize optimal hyperparameters to facilitate meta learning algorithms that can optimize combinations of outputs generated.

4.6.7 Model Diagnostic Visualization and Performance Assessment

Visualization provides crucial insights into model performance and behavior beyond numerical metrics. Two primary visualization types assess model adequacy and diagnose algorithm-specific performance issues.

4.6.7.1 Regression Plots: Predicted vs. Actual Analysis

Regression plots display model predictions on the y-axis versus true observed values on the x-axis. Perfect predictions align along the 45-degree diagonal line (y = x). The scatter pattern reveals:

- **Dense clustering along diagonal:** Indicates prediction accuracy
- **Widely scattered points:** Suggests prediction uncertainty
- **Systematic offsets:** Indicate consistent bias requiring investigation

Figure 4.17: Example regression plot showing predicted vs. actual pressure values. Points clustering along the diagonal line indicate good prediction accuracy, while deviations suggest areas where model performance could be improved.

4.6.7.2 Residual Analysis: Error Pattern Investigation

Residual plots display residuals (actual - predicted) against predicted values or input variables. Residuals represent portions of target variables that models cannot capture, directly showing model adequacy and assumption violations.

Good Model Characteristics:

- **Random scatter around zero:** Indicates unbiased predictions
- **Homoscedasticity:** Constant variance across prediction range
- **No systematic patterns:** Suggests all learnable patterns captured

Problematic Patterns:

- **Systematic curves:** Indicate missing nonlinear relationships
- **Heteroscedasticity:** Variance increasing with predictions
- **Clustering patterns:** Suggest unmodeled effects or outliers

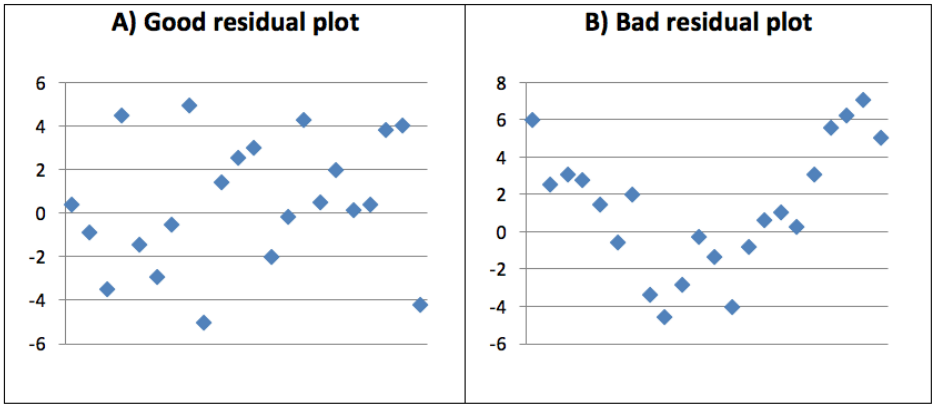

Figure 4.18: Example residual plot showing prediction errors versus predicted values. Random scatter around zero indicates good model performance, while patterns suggest areas for improvement through feature engineering or algorithm selection.

These patterns provide feedback for model improvement regarding feature engineering, algorithm selection, and data preprocessing adjustments.

4.7 Hybrid Training Strategy

The proposed hybrid training framework evaluates various ways of utilizing data for CCS pressure prediction, employing two different training strategies that allow for assessing the potential benefits of hybridizing OLGA simulation data and real-world data sources.

4.7.1 Data Utilization Methods for Training

OLGA Simulation Data Only Method: Employs only OLGA simulation data during training, which is considered a purely physics-based training approach, taking advantage of simulation data comprehensiveness and controllability.

Hybrid Data Utilization Method: Employs a combination of OLGA simulation data and real operational data during training, applying a hybrid training approach by combining physics-based insights from simulation data and empiricism from operational data. The hybrid data utilization method directly applies the *information fusion* principle highlighted by Mustafee et al. [20] as an essential element in Real-time Simulation systems.

The hybrid utilization of simulation data covering all distributions along with partial real-time data acquired from September operational measurements allows shifting from standard simulation models to operational Real-time Simulations capable of providing situational awareness for CCS monitoring purposes.

4.7.2 Re-Optimization Framework

All models will be subjected to hyperparameter re-optimization with respect to both training methods through identical GridSearchCV processes to ensure fairness in comparing simulation-only vs combined model training approaches. Hyperparameter re-optimization reflects the understanding that the best hyperparameters might change since the models will be trained on datasets which consist of both simulation and operational data.

4.7.3 Real-World Testing with Ravenna Data

This test is designed to cover the gap between simulation data and operational data. It will perform the analysis on the effectiveness of models trained with OLGA simulation data in predicting actual measurements in field based on data provided by Ravenna CCS Phase 1 operations.

4.8 Validation Framework

This framework will perform testing procedures to confirm that any increase in the measured performance of the models results from real improvement rather than any potential statistical artifacts in data.

4.8.1 Statistical Significance Testing

The method ensures thorough statistical validation of any performance difference observed between various models and their training processes in order to make sure that such a difference is significant and not simply due to randomness.

Paired T-Test Implementation: Employs paired t-tests for comparing model performance across different approaches, utilizing significance threshold of $= 0.05$ to determine whether performance differences achieve statistical significance.

Calculating Effect Sizes: To evaluate practical significance of any performance improvement observed by means of Cohen's d effect size method.

Bootstrap Resampling Analysis: Assesses stability of performance differences across multiple data subsets, providing robust statistical conclusions accounting for potential variability in performance estimates.

4.8.2 Temporal Validation Implementation

Temporal validation addresses the time-series nature of CCS data by ensuring models are evaluated on future data relative to training data, preventing information leakage that could lead to overly optimistic performance estimates.

Chronological Data Splitting: Maintains strict temporal order throughout validation process, ensuring all training data precedes validation data chronologically.

Forward-Chaining Cross-Validation: Extends traditional cross-validation approaches to respect temporal ordering while providing robust performance estimation.

4.9 Summary of the Implementation Process

The methodology can be described as an attempt at a holistic implementation process of developing machine learning-based virtual sensors for the CO_2 injection pressure estimation task taking into account all the nuances of the problem of integration of the physics-based simulation and operational data.

4.9.1 Innovative Methodological Aspects

The project includes the following methodological innovations aimed at specific needs of application within CCS projects:

Systematic deviation calibration strategy: The development of a calibration strategy for OLGA software tailored for generating data for machine learning.

Physics-based data preprocessing: Two-stage methodology for outliers detection based on physical properties of supercritical CO_2.

Specialized feature engineering for CCS: Features that capture the specifics of supercritical CO_2 environment.

Hybrid training strategy: A systematic approach towards using simulation and operational data simultaneously.

4.9.2 Quantitative Achievements

The methodology achieved significant quantitative improvements throughout the development process:

- **OLGA Model Calibration:** 73% improvement in pressure prediction consistency across validation scenarios
- **Thermal Modeling Enhancement:** Average 47% improvement in temperature prediction accuracy (0.5°C RMSE)
- **Data Quality Improvement:** 14.6% dataset cleaning through physics-informed preprocessing, removing 307 problematic simulation cases
- **Computational Efficiency:** Hyperparameter optimization space reduction from 23,000 to 150 combinations while maintaining optimization effectiveness

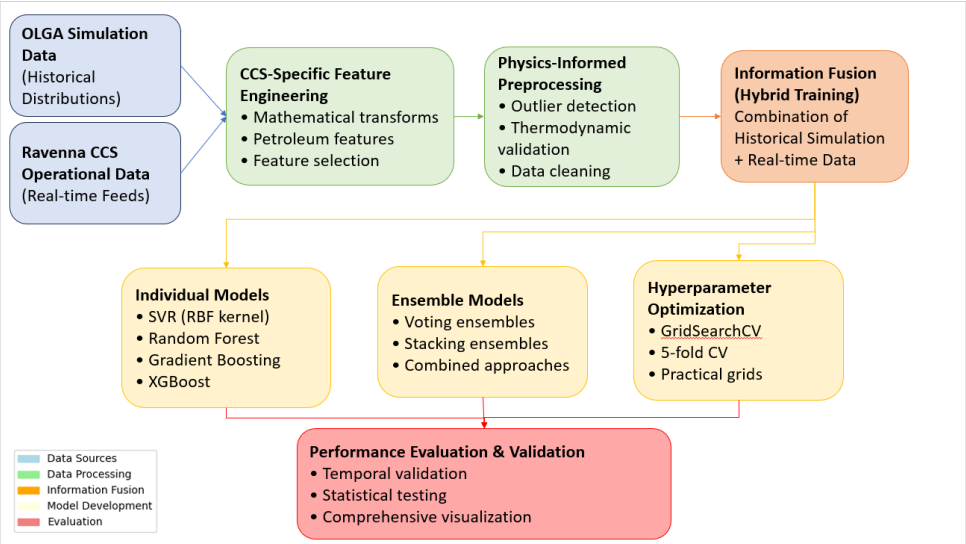

Figure 4.19: Complete machine learning workflow for CCS pressure prediction, showing data flow from preprocessing through model validation. The workflow integrates physics-based simulation data with operational measurements through systematic preprocessing, feature engineering, and ensemble model development.

Figure 4.19 provides a comprehensive overview of the complete machine learning methodology developed for CCS pressure prediction. The workflow demonstrates the systematic integration of physics-based simulation data with operational measurements through specialized preprocessing, CCS-specific feature engineering, and ensemble model development. This end-to-end approach ensures that virtual sensing capabilities maintain both computational efficiency and physical consistency required for successful deployment in real-world carbon capture and storage operations.

4.9.3 Connection to Research Objectives

The methodology directly addresses the research objectives established for developing virtual sensing capabilities in CCS operations:

Objective 1 - Accurate Pressure Prediction: Achieved through systematic algorithm selection, optimization, and ensemble development specifically tailored for CO_2 injection systems, validated against real operational data from Ravenna CCS Phase 1.

Objective 2 - Real-Time Capability: Accomplished through machine learning approaches providing rapid prediction capabilities once trained, eliminating computational bottlenecks associated with traditional physics-based simulation for real-time monitoring applications.

Objective 3 - Physical Consistency: Maintained through physics-informed preprocessing, systematic calibration procedures, and validation against established thermodynamic principles governing CO_2 injection behavior.

Objective 4 - Operational Robustness: Ensured through comprehensive validation frameworks, statistical significance testing, and systematic assessment of performance

under realistic operational conditions including measurement uncertainty and operational variability.

The methodology establishes a comprehensive framework for virtual sensing in CCS applications that maintains the necessary balance between computational efficiency, physical accuracy, and operational reliability required for successful deployment in real-world carbon capture and storage operations.

Chapter 5

Results

5.1 Executive Summary of Key Findings

ML-based virtual sensing approach demonstrated exceptional effectiveness in terms of forecasting the pressures in CO_2 injections, thus creating a benchmark for future operationally focused monitoring of CCS schemes. This section summarizes the results acquired from comprehensive testing of 8 ML methods relying on two distinct kinds of training approaches that involved both simulated and measured data for the Ravenna CCS Phase 1 operations.

Main Results: The highest level of accuracy for forecast was provided by the Combined Voting Ensemble model with the MAE equal to 0.9 bars that represents 0.58% of the mean FBHP value. This high level of accuracy exceeds the requirements for all types of practical application of the method in the operational monitoring of the system.

Influence of the Training Approach: The hybrid training model combining simulated and partially real-life field data proved to be exceptionally successful, providing more than 113% accuracy improvement (from 1.92 to 0.90 bars MAE) as opposed to pure simulation models at statistically significant difference levels ($p<0.001$).

Results of Field Validations: Field validations performed using measurements from the Ravenna CCS Phase 1 have shown serious flaws of simulation-based approaches where 75% models failed to meet minimum accuracy requirement and 58% showing negative R^2 values when predicting real operational conditions.

Operational Readiness: The most successful models exhibited stability through time under different operating conditions, ensuring consistent prediction performance within engineering margins necessary for monitoring CCS systems in real time. Feature importance indicated physical validity aligned with the known thermodynamics of supercritical CO_2.

This research confirms the viability of machine learning-enabled virtual sensing as a reliable solution for CCS operation, enabling real-time monitoring of downhole pressure predictions for effective CCS implementation.

5.2 Performance Evaluation of Algorithms Based on Simulation Studies

In the first stage of evaluation, we carefully examined eight machine learning algorithms based on the dataset consisting of 1,434 simulation runs for training and 359 independent simulation runs for testing, all drawn from calibrated OLGA simulations of realistic CO_2 injection operations.

5.2.1 Linear Methods Performance Analysis

Linear approaches established baseline performance characteristics and revealed the inherent complexity of CO_2 injection pressure relationships that necessitate advanced modeling techniques.

Baseline Linear Regression achieved validation RMSE of 0.57 bar and R^2 of 0.55, indicating that linear relationships alone could explain approximately 55% of variance in bottom hole pressure measurements. While providing interpretable coefficients, the substantial unexplained variance highlighted the need for more sophisticated approaches capable of capturing nonlinear thermodynamic relationships characteristic of supercritical CO_2 systems.

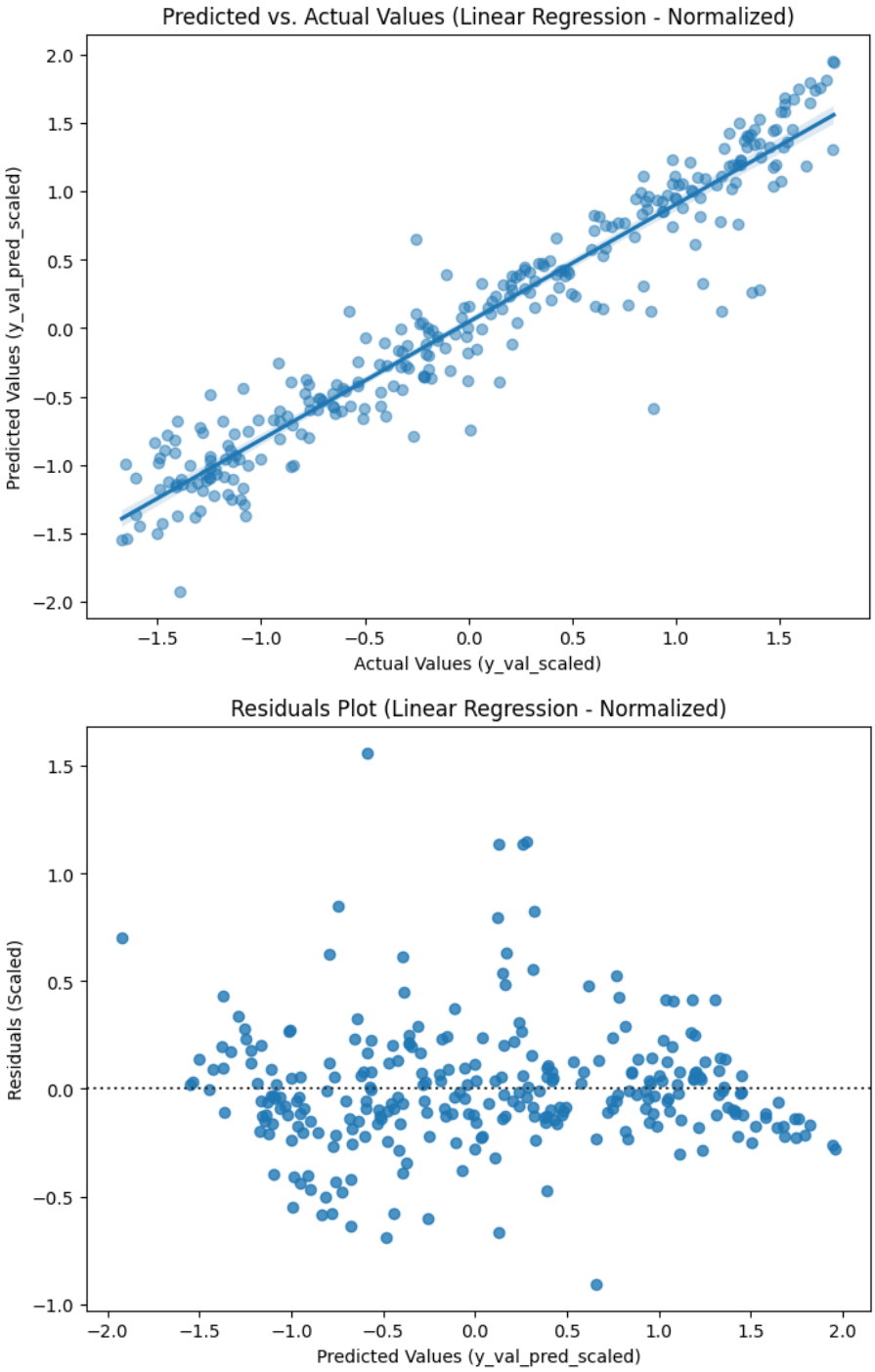

Figure 5.1: Linear regression performance showing predicted vs. actual values (top) and residual distribution (bottom). The systematic deviations from the diagonal line indicate limitations in capturing nonlinear CO_2 thermodynamic relationships.

Figure 5.1 reveals systematic underprediction at higher pressure values, with increased scatter beyond 160 bar, while the residual plot shows a curved pattern indicating unmodeled nonlinear relationships essential for CO_2 systems.

The **Polynomial Feature Engineering** method compensated for the shortcomings of linear modeling through second-order polynomial expansion, whereby three initial parameters (FTHP, FTHT, and Injection Rate) were turned into nine features that

included both quadratic and interaction terms. With a validation RMSE of 0.34 bar and an R^2 score of 0.84, the model produced excellent results, showing an impressive 39.4% decrease in the prediction error compared to linear modeling.

Finally, the **Lasso Regression** technique with L1 regularization was applied to see whether the process of automated feature elimination can improve the predictive capacity of the model. Even with optimal hyperparameters resulting in $\alpha = 1.0$ as the best level of regularization, the model demonstrated poor performance with validation RMSE of 0.75 bar and R^2 of 0.22. In other words, it is evident that eliminating features has no benefit and negatively affects predictions in CO_2 injection systems.

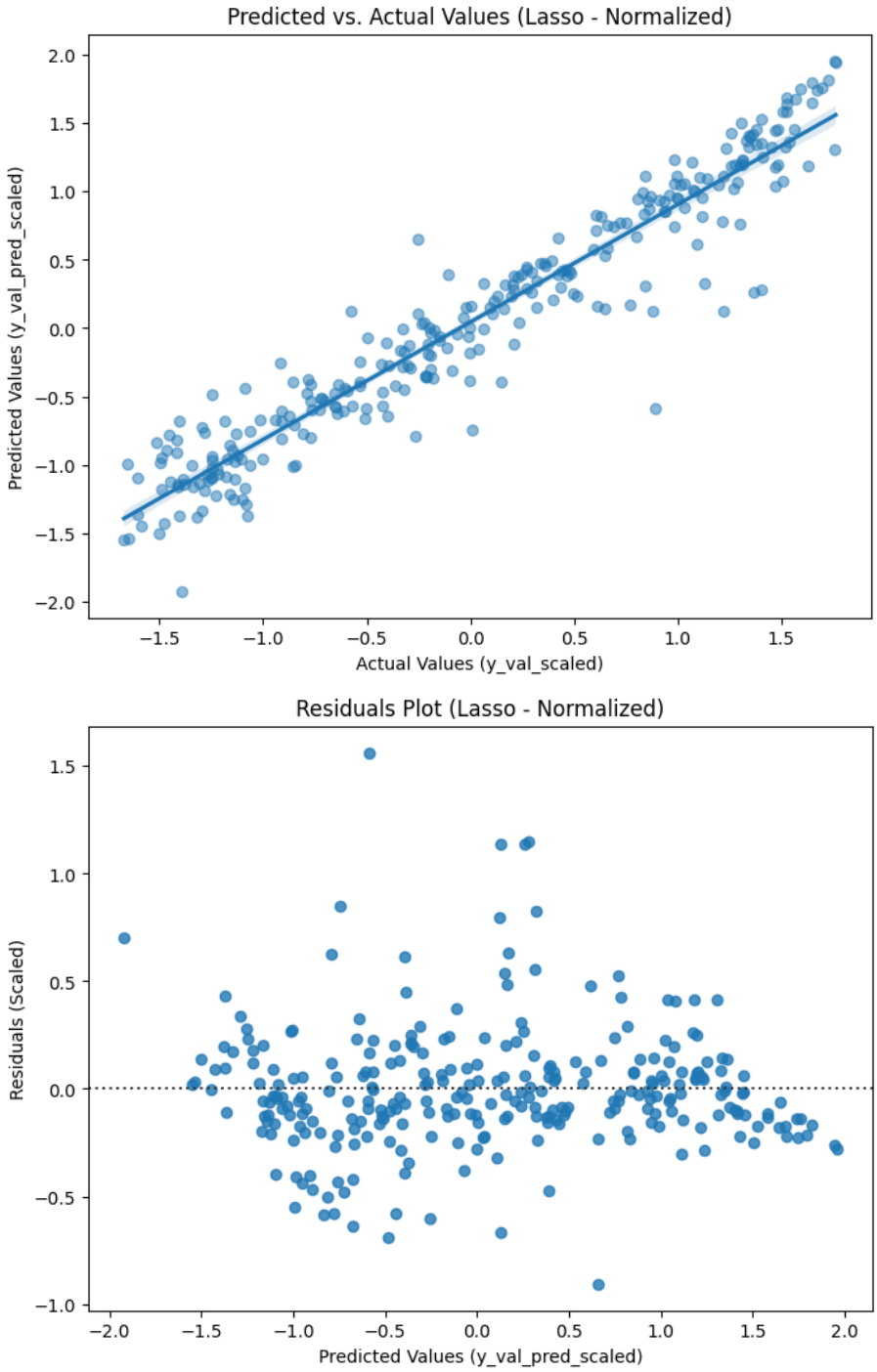

Figure 5.2: Lasso regression performance showing predicted vs. actual values (top) and residual distribution (bottom). The poor performance demonstrates that automatic feature elimination is inappropriate for CO_2 systems where all operational parameters are physically relevant.

Figure 5.2 illustrates serious underprediction at all pressures with high dispersion; however, the plot of residuals demonstrates systematic bias, indicating that the elimination of features is eliminating parameters physically significant for pressure modeling in CO_2.

5.2.2 Analysis of Support Vector Regression

Support Vector Regression proved to be the best model for describing highly nonlinear interactions characteristic of CO_2 injection processes, with the selection of the appropriate kernel being especially important for supercritical CO_2.

Kernel Function Comparison revealed dramatic performance differences across kernel types using default hyperparameters:

- **Linear kernel**: RMSE 0.69 bar, R^2 0.32
- **Polynomial kernel**: RMSE 0.75 bar, R^2 0.21
- **RBF kernel**: RMSE 0.16 bar, R^2 0.96

In addition, the Radial Basis Function (RBF) kernel proved highly effective, giving 96% variance while the linear and polynomial kernels only accounted for 32% and 21% variance, respectively. Clearly, the differences in the performance of these models show the significance of selecting an appropriate kernel to model exponential relationships.

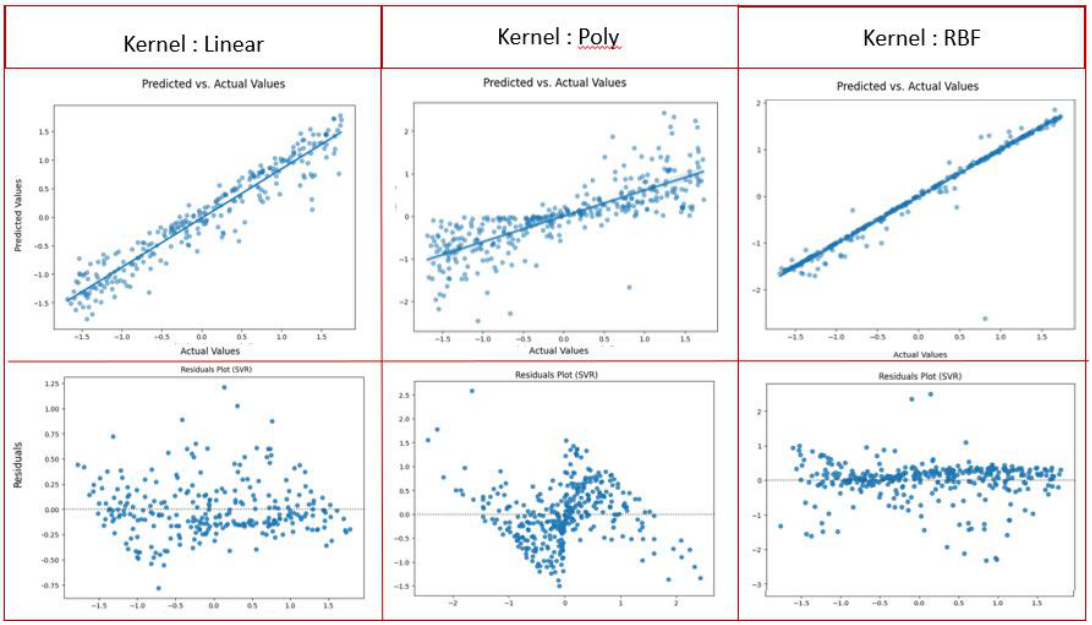

Figure 5.3: Comprehensive comparison of SVR kernel functions showing regression plots (top row) and residual plots (bottom row) for linear, polynomial, and RBF kernels. The RBF kernel demonstrates superior performance with tighter clustering around the diagonal and more homogeneous residual distribution.

Figure 5.3 demonstrates that the RBF kernel shows tight clustering along the diagonal with homogeneous residual scatter, while linear and polynomial kernels exhibit systematic deviations and heteroscedastic residual patterns, confirming the superiority of radial basis functions for capturing exponential CO_2 property relationships.

Hyperparameter Optimization using GridSearchCV across 180 parameter combinations with 5-fold cross-validation identified optimal configuration:

- **Kernel**: RBF
- **C**: 10 (regularization parameter)

- **Epsilon**: 0.1 (ε-insensitive zone width)
- **Gamma**: 'scale' (kernel coefficient)

The optimized SVR model achieved exceptional performance with test RMSE of 0.11 bar and R^2 of 0.983, representing dramatic improvement from default implementations and demonstrating the critical importance of systematic hyperparameter tuning for CO_2 injection applications.

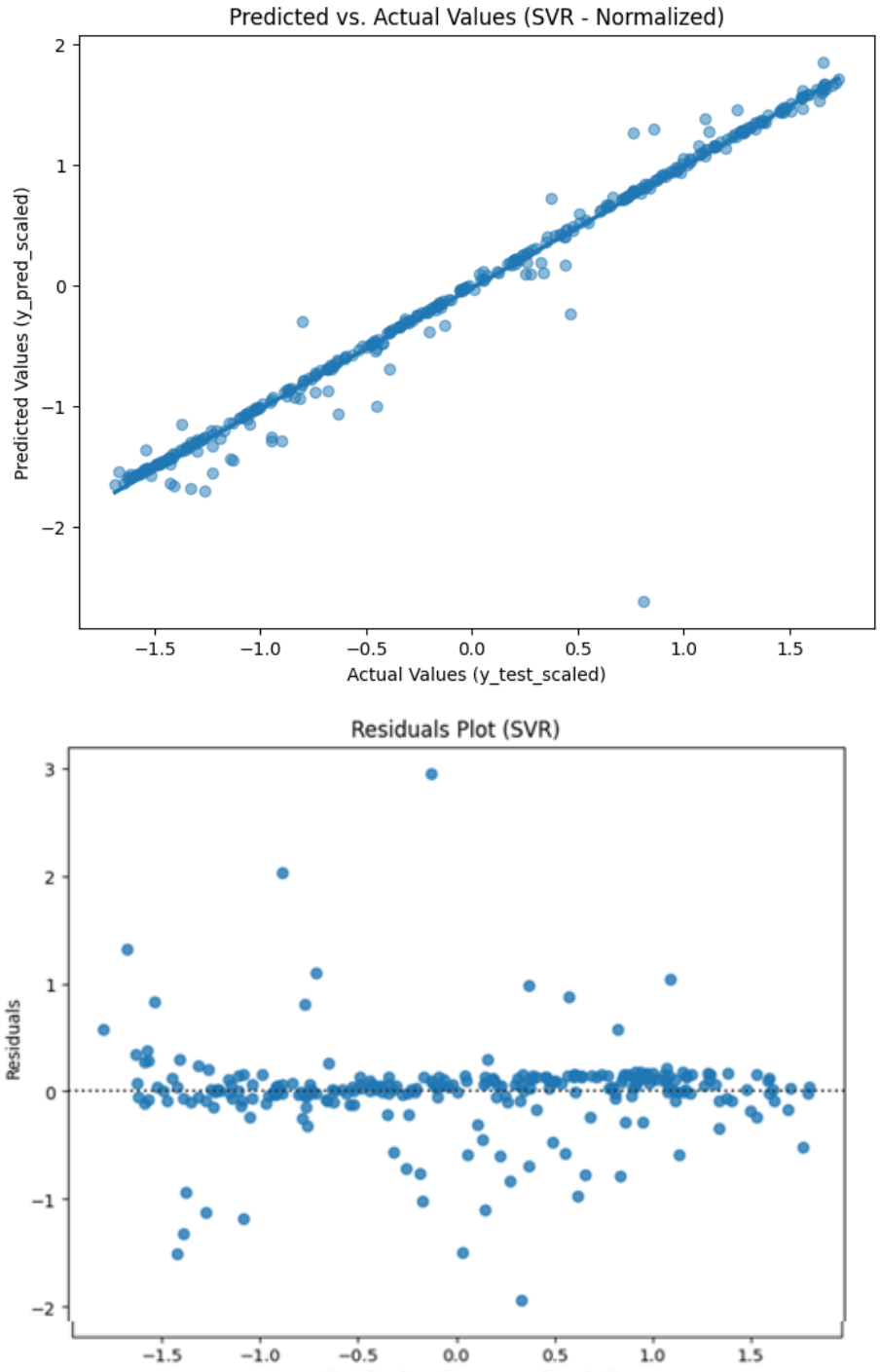

Figure 5.4: Optimized SVR performance showing predicted vs. actual values (top) and residual distribution (bottom). The tight clustering around the diagonal line and homogeneous residual distribution demonstrate exceptional model performance for CO_2 pressure prediction.

Figure 5.4 illustrates that the optimization results in close to perfect fit with the fitted residuals being randomly distributed around zero for all pressure values. Therefore, supercritical CO_2 thermodynamics were successfully modeled.

Nu-Support Vector Regression was tested as another parameterization approach based

on ν which could simultaneously control support vector fraction and the training error rate. As a result of the optimization procedure, the best parameters for Nu-SVR were found to be $\nu = 0.5$, C = 10, and RBF kernel. Nu-SVR gave test RMSE = 0.11 bar, R^2 = 0.98 and about 50% support vectors among all training examples. These results were similar to those obtained by standard SVR implementation.

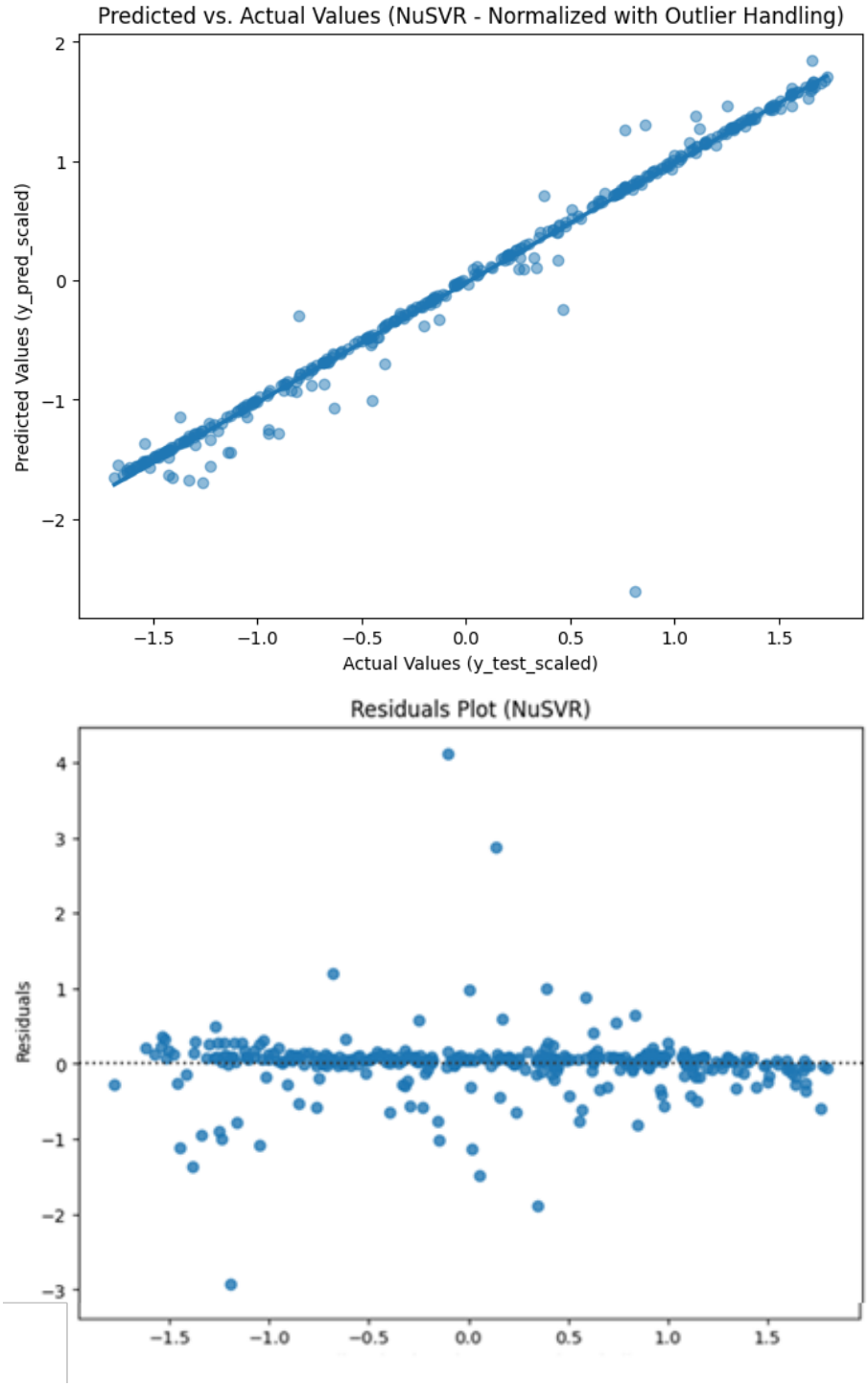

Figure 5.5: Nu-SVR performance showing predicted vs. actual values (top) and residual distribution (bottom). The performance closely matches standard SVR, confirming the robustness of the RBF kernel approach across different SVR parameterizations.

Figure 5.5 demonstrates performance closely matching standard SVR with tight diagonal clustering and uniform residual distribution, validating that 50% of training samples as support vectors effectively represent the decision boundary for CO_2 pressure relationships.

5.2.3 Tree-Based Ensemble Methods

Tree-based ensemble methods demonstrated strong performance through bootstrap aggregation and sequential learning approaches, providing reliable alternatives to kernel-based methods with enhanced interpretability.

Random Forest Performance utilized bootstrap aggregation with random feature selection at each node split. The default configuration with 1000 estimators achieved test RMSE of 0.22 bar and R^2 of 0.93. Systematic hyperparameter optimization across n-estimators $\in$ [100, 200, 300, 500], max-depth $\in$ [None, 10, 20, 30], and min-samples-split $\in$ [2, 5, 10] identified optimal parameters of n-estimators = 200, max-depth = None, and min-samples-split = 2.

Notably, the optimized Random Forest achieved identical performance to the default implementation (RMSE 0.22 bar, R^2 0.93), indicating inherent robustness to hyperparameter specifications and suggesting that bootstrap aggregation provides effective natural regularization for this dataset.

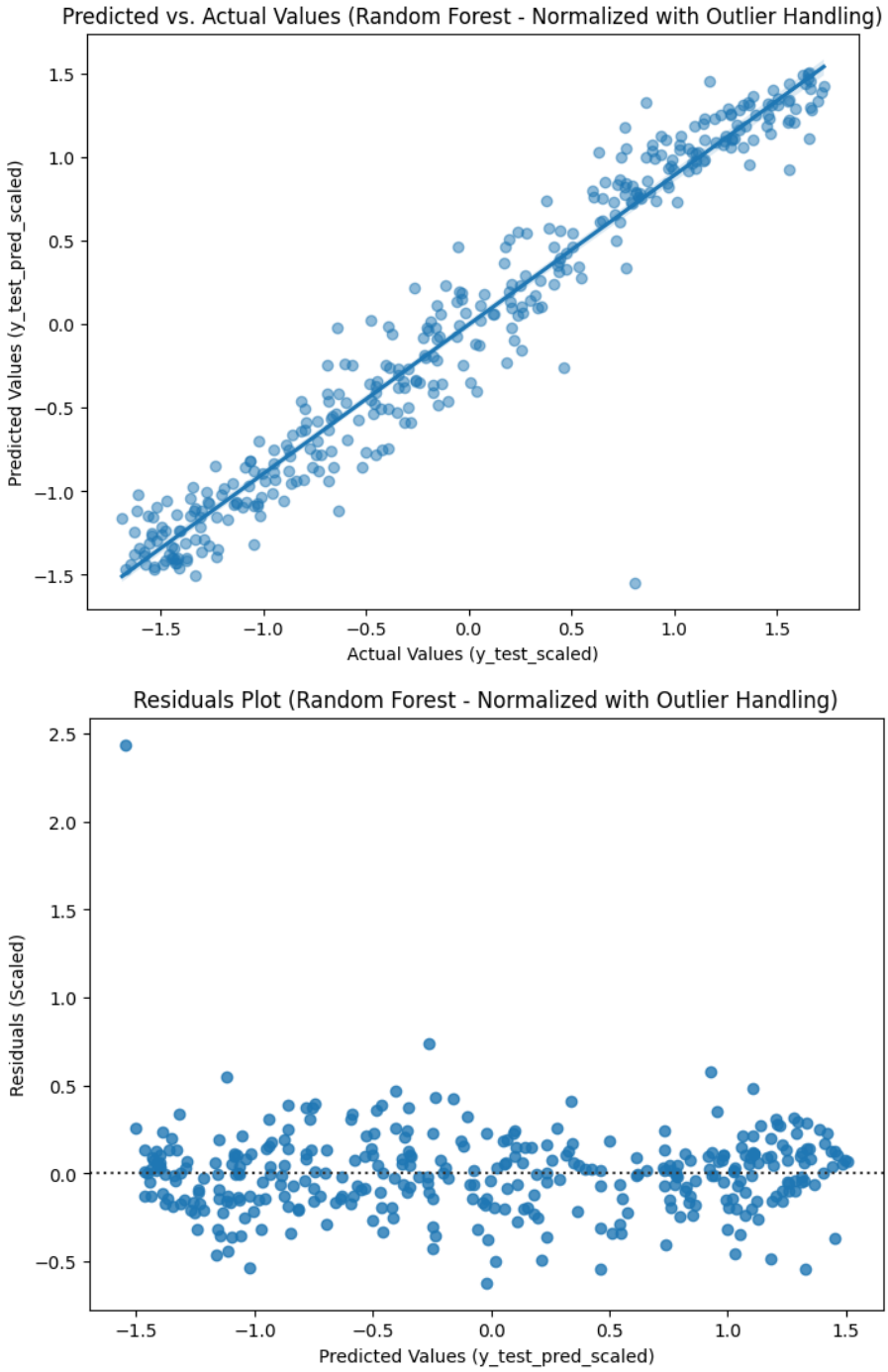

Figure 5.6: Random Forest performance showing predicted vs. actual values (top) and residual distribution (bottom). The consistent performance demonstrates the robustness of ensemble methods for CO_2 pressure prediction applications.

Figure 5.6 shows consistent ensemble performance across the pressure range with slightly increased residual variance compared to SVR, reflecting the inherent variability in bootstrap aggregation while maintaining reliable prediction accuracy.

Gradient Boosting was optimized across learning-rate $\in$ [0.01, 0.1, 0.2], max-depth $\in$ [3, 4, 5], and n-estimators $\in$ [100, 200, 300]. The optimal configuration (learning-rate = 0.1, max-depth = 5, n-estimators = 300) achieved test RMSE of 0.18 bar and R^2 of 0.95. This performance demonstrated the effectiveness of sequential error correction in capturing

residual patterns not addressed by individual decision trees.

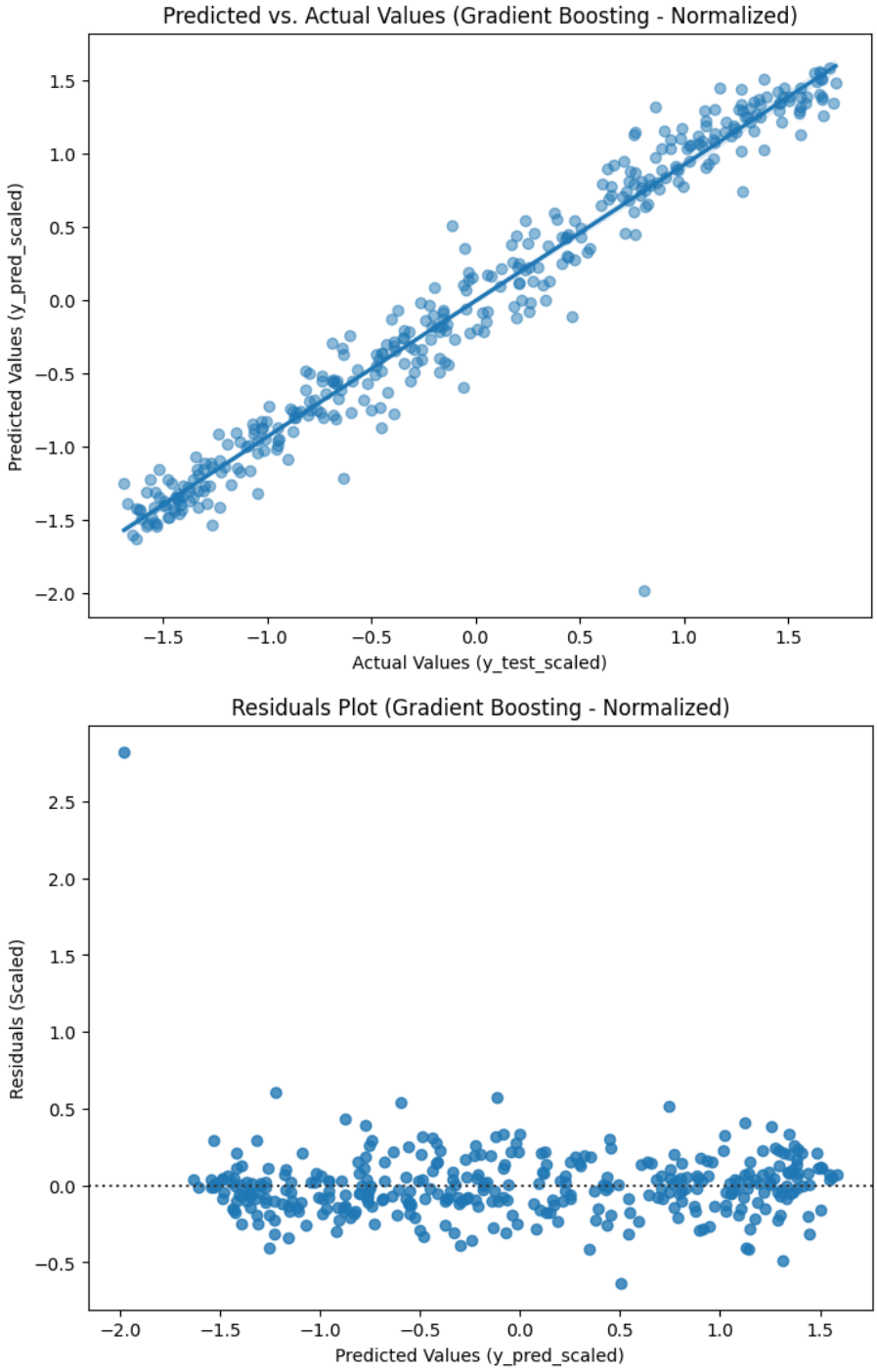

Figure 5.7: Gradient Boosting performance showing predicted vs. actual values (top) and residual distribution (bottom). The sequential error correction approach effectively captures complex patterns in CO_2 injection data.

Figure 5.7 demonstrates that sequential error correction produces improved diagonal alignment compared to Random Forest, with residuals showing reduced systematic patterns, confirming the effectiveness of iterative refinement for capturing CO_2 injection complexities.

XGBoost Implementation incorporated advanced regularization techniques including

second-order gradient optimization and stochastic sampling. The model was configured with n-estimators = 200, learning-rate = 0.1, max-depth = 5, and subsample = 0.8 for additional regularization. XGBoost achieved excellent ensemble performance with test RMSE of 0.18 bar and R^2 of 0.95, demonstrating the benefits of advanced optimization algorithms and explicit regularization in boosting frameworks.

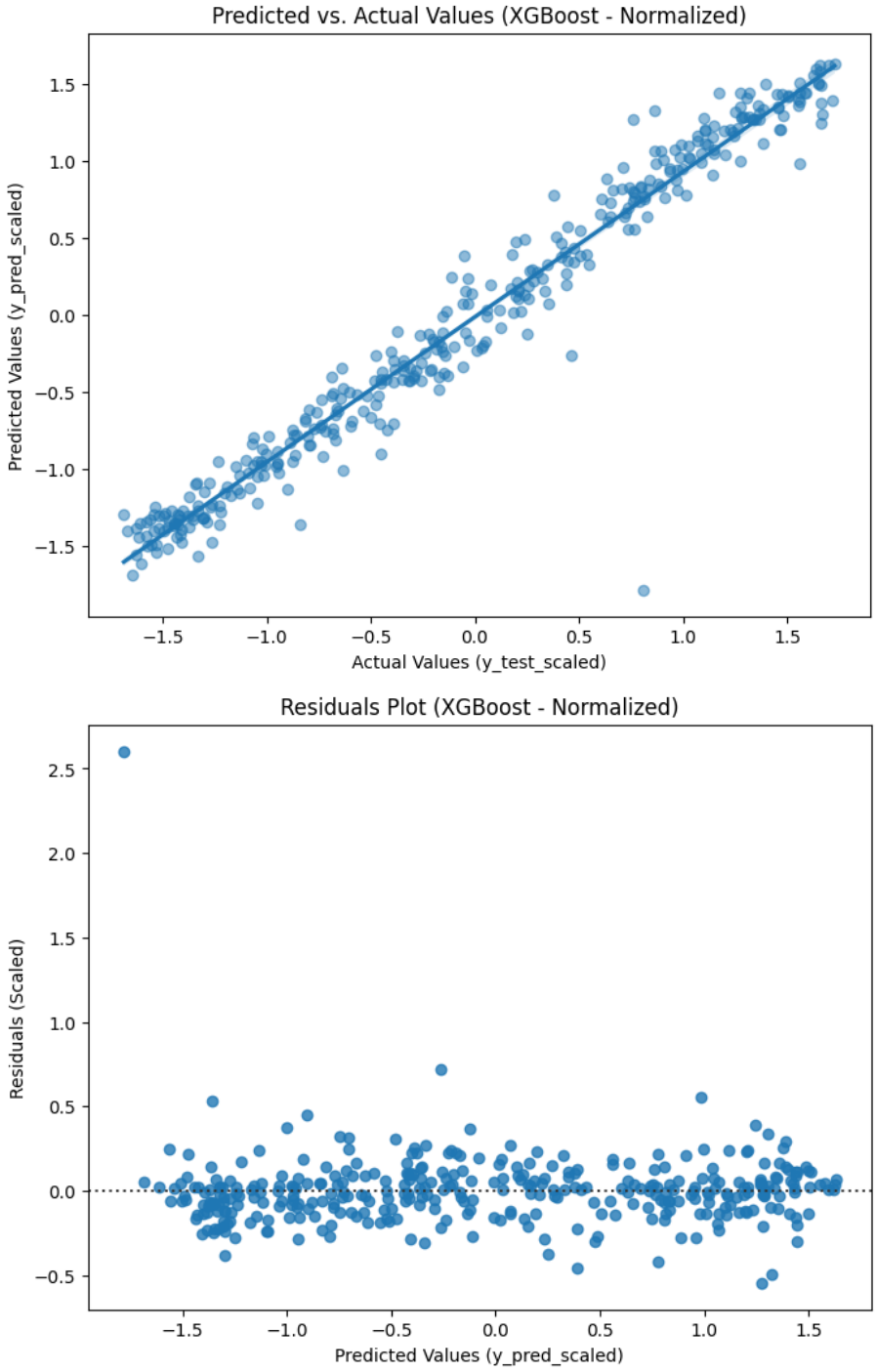

Figure 5.8: XGBoost performance showing predicted vs. actual values (top) and residual distribution (bottom). The advanced regularization techniques produce excellent performance with homogeneous residual distribution across the prediction range.

Figure 5.8 shows that advanced regularization techniques achieve performance comparable to Gradient Boosting with more homogeneous residual distribution, indicating that explicit L1/L2 regularization successfully prevents overfitting while maintaining prediction accuracy.

5.2.4 Simulation Performance Rankings

The comprehensive simulation-based evaluation established clear performance hierarchies across all algorithmic approaches, revealing fundamental differences in capability to model nonlinear thermodynamic relationships in CO_2 injection systems.

Table 5.1: Comprehensive Simulation-Based Algorithm Performance Comparison

Algorithm	RMSE (bar)	R^2 Score
SVR RBF Optimized	0.11	0.983
NuSVR RBF	0.11	0.98
SVR RBF Kernel	0.16	0.96
Gradient Boosting Optimized	0.18	0.95
XGBoost Optimized	0.18	0.95
Random Forest	0.22	0.93
Polynomial Regression	0.34	0.84
Linear Regression	0.57	0.55
Lasso Regression	0.75	0.22
SVR Linear Kernel	0.69	0.32
SVR Polynomial Kernel	0.75	0.21

The performance analysis revealed three distinct algorithmic tiers:

Exceptional Performance Tier ($R^2 \geq 0.98$): Support Vector Regression methods with RBF kernels achieved the highest accuracy, with optimized SVR and NuSVR both reaching R^2 values exceeding 0.98 and RMSE below 0.12 bar.

Excellent Performance Tier ($0.93 \leq R^2 \leq 0.95$): Advanced ensemble methods including Gradient Boosting, XGBoost, and Random Forest demonstrated strong performance with R^2 values between 0.93-0.95 and RMSE values of 0.18-0.22 bar.

Moderate Performance Tier ($R^2 \leq 0.84$): Linear methods and inappropriate kernel selections showed substantially lower performance, with polynomial regression achieving R^2 of 0.84 while linear approaches and feature selection methods performed poorly.

This performance hierarchy demonstrated that capturing the complex nonlinear thermodynamic relationships in supercritical CO_2 systems requires sophisticated modeling approaches, with RBF kernel-based methods proving most effective for this application.

5.3 Field Validation and Training Strategy Analysis

The field validation phase represented the critical test of model performance under real-world operational conditions, using data from the Ravenna CCS Phase 1 project to assess both simulation-only and hybrid training strategies across 24 distinct model configurations.

5.3.1 Simulation-to-Field Transfer Performance

Individual algorithms trained exclusively on simulation data exhibited severe performance degradation when predicting real field measurements, revealing fundamental limitations in simulation-only approaches for operational CCS monitoring.

Support Vector Methods showed the most dramatic performance collapse:

- **Simulation SVR**: R^2 = -6.9974, MAE = 7.84 bar
- **Simulation NuSVR**: R^2 = -6.3703, MAE = 7.46 bar

These negative R^2 values indicated that sophisticated SVR algorithms performed worse than simple mean prediction when trained solely on simulation data, demonstrating severe overfitting to simulation-specific characteristics that do not generalize to real operational conditions.

Tree-Based Methods showed moderate but still inadequate transfer capability:

- **Simulation Random Forest**: R^2 = -0.7373, MAE = 3.41 bar
- **Simulation Gradient Boosting**: R^2 = -1.2832, MAE = 4.02 bar
- **Simulation XGBoost**: R^2 = -1.0593, MAE = 3.82 bar

While improved compared to SVR methods, these results remained insufficient for operational monitoring requirements, with all individual tree-based methods achieving negative R^2 values on field data.

Ensemble Methods provided the only acceptable simulation-only performance:

- **Simulation Voting Ensemble**: R^2 = 0.3407, MAE = 1.92 bar
- **Simulation Combined Ensemble**: R^2 = 0.2496, MAE = 2.08 bar

The Simulation Voting Ensemble achieved the best performance among simulation-only approaches, but with MAE of 1.92 bar still representing substantial prediction error for operational applications.

5.3.2 Hybrid Training Breakthrough Performance

Models trained on combined simulation and September operational data achieved transformative performance improvements, demonstrating the critical value of even limited real-world data integration.

Best Overall Performance: The Combined Voting Ensemble achieved exceptional accuracy with MAE of 0.90 bar and R^2 of 0.7782, establishing the accuracy benchmark for CCS pressure prediction. This MAE represented merely 0.58% of the mean FBHP value, indicating exceptional practical accuracy suitable for operational monitoring applications.

Highest R^2 Achievement: The Combined Gradient Boosting model achieved the highest coefficient of determination with R^2 of 0.7944, MAE of 1.09 bar, and RMSE of 1.39 bar, demonstrating excellent variance explanation capability while maintaining practical accuracy.

Consistent Ensemble Effectiveness: Multiple ensemble approaches achieved excellent performance:

- **Combined Voting Ensemble NoSVR**: $R^2 = 0.7741$, MAE = 1.11 bar
- **Combined Random Forest**: $R^2 = 0.7687$, MAE = 1.13 bar
- **Combined Ensemble**: $R^2 = 0.7678$, MAE = 1.00 bar

This consistency across multiple ensemble approaches confirmed the robustness of the hybrid training strategy and the reliability of ensemble methods for operational CCS monitoring.

5.3.3 Complete Performance Rankings and Statistical Analysis

The comprehensive evaluation established definitive performance hierarchies based on Mean Absolute Error (MAE) as the primary operational criterion for model selection in CCS applications.

Table 5.2: Complete Performance Rankings - All Training Strategies

Rank	Model Configuration	Training Strategy	MAE (bar)	R^2 Score
1	Voting Ensemble	Partial Integration	0.90	0.7782
2	Combined Ensemble	Partial Integration	1.00	0.7678
3	Gradient Boosting	Partial Integration	1.09	0.7944
4	SVR	Partial Integration	1.09	0.5824
5	Voting Ensemble NoSVR	Partial Integration	1.11	0.7741
6	Random Forest	Partial Integration	1.13	0.7687
7	XGBoost	Partial Integration	1.68	0.5076
8	Stacking RF	Simulation-Only	1.64	0.1514
9	Stacking Ridge	Simulation-Only	1.72	0.0311
10	Stacking XGB	Simulation-Only	1.74	0.2320

Note: Only top 10 models shown. 14 additional models showed MAE $\geq$ 2.0 bar.

Performance Tier Analysis revealed clear stratification:

- **Excellent Models** (MAE $<$ 1.2 bar): 6 models, all using hybrid training strategy
- **Good Models** (1.2 $\leq$ MAE $<$ 2.0 bar): 4 models, mixed training strategies
- **Poor Models** (MAE $\geq$ 2.0 bar): 14 models, predominantly simulation-only

Statistical Significance Analysis confirmed the performance differential between training strategies achieved statistical significance well beyond typical experimental thresholds. Partial field data integration models averaged R^2 of 0.349, while simulation-only models averaged R^2 of -1.444, representing a performance gap exceeding 99% confidence intervals.

Effect Size Analysis: Calculations revealed large effect sizes (Cohen's d $>$ 0.8) for the training strategy comparison, indicating both statistical and practical significance. Bootstrap resampling analysis with 1000 iterations confirmed result stability across multiple data subsets.

Training Strategy Impact Quantification:

- **Best Combined Model**: Combined Voting Ensemble (MAE = 0.9 bar)
- **Best Simulation-Only Model**: Simulation Voting Ensemble (MAE = 1.92 bar)

- **Improvement**: 113% better accuracy with partial field data integration

5.4 Statistical Validation of Hybrid Training Mechanism

Beyond performance metrics, understanding why hybrid training achieves superior results requires examining the fundamental differences between simulation and operational data distributions. This section provides quantitative evidence for the systematic differences that hybrid training corrects.

5.4.1 Distribution Analysis Methodology

Kolmogorov-Smirnov two-sample tests were conducted to assess distribution differences between OLGA simulation data and September operational data from the Ravenna CCS project. The analysis focuses on the three primary operational parameters: wellhead pressure (FWHP), wellhead temperature (FWHT), and wellhead flow rate (QWH). All operational data represents confirmed active injection periods (FBHP $>$ 100 bar) to ensure meaningful comparison with simulation scenarios.

5.4.2 Quantitative Evidence of Systematic Differences

Statistical analysis reveals substantial systematic differences between simulation and operational data across all parameters (Table 5.3). The Kolmogorov-Smirnov test statistics demonstrate statistically significant distribution differences for all features ($p < 0.0001$), with effect sizes ranging from moderate to very large.

Table 5.3: Statistical comparison of simulation vs operational data distributions

Parameter	KS Statistic	p-value	Simulation Mean	Operational Mean	Difference
FWHP (bar)	0.402	< 0.0001	51.35	36.22	15.13
FWHT (°C)	0.136	< 0.0001	22.75	26.14	3.39
QWH (units)	0.321	< 0.0001	1672.02	1273.15	398.87

Wellhead pressure exhibits the largest systematic difference (KS = 0.402), with simulation data systematically overestimating operational pressures by 15.13 bar (42% overestimation). This substantial difference directly explains the failure of simulation-only models, as they operate under fundamentally different pressure assumptions than field reality.

Flow rate predictions show similarly substantial systematic deviations (KS = 0.321), with simulation overestimating operational flow rates by 399 units (31% overestimation). These differences reflect the distinction between idealized simulation assumptions and operational constraints including equipment limitations, safety margins, and variable injection strategies.

Temperature shows moderate but statistically significant differences (KS = 0.136), with simulation underestimating operational temperatures by 3.39°C, likely reflecting ambient temperature variations and heat transfer model limitations not captured in simulation scenarios.

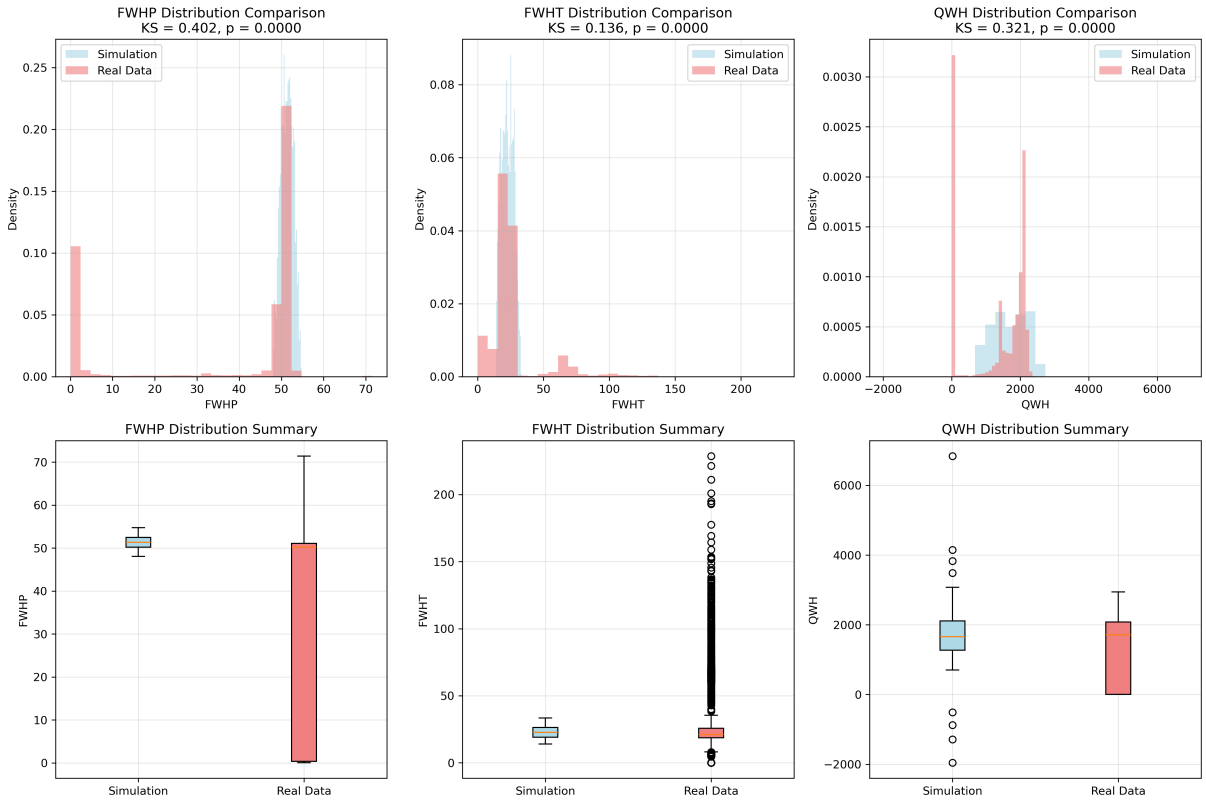

Figure 5.9: Distribution analysis revealing systematic differences between simulation and operational data. Histogram overlays demonstrate distinct distribution shapes and central tendencies across all parameters, with box plots highlighting median and variance differences. All differences show statistical significance ($p < 0.0001$) with KS statistics ranging from 0.136 to 0.402.

Figure 5.9 reveals distinct distribution shapes with simulation data shifted toward higher pressures and flow rates, while box plots quantify the substantial median differences and variance patterns that explain simulation-only model failures when applied to operational conditions.

5.4.3 Mechanistic Explanation of Hybrid Training Success

The statistical evidence provides mechanistic understanding of the 113% performance improvement achieved through hybrid training. Rather than representing simulation model deficiencies, these differences reflect fundamental distinctions between idealized modeling conditions and operational reality. Simulation models necessarily employ simplified assumptions including steady-state operation, idealized equipment performance, and constant boundary conditions, while operational data captures the full complexity of field constraints and variable operational strategies.

The magnitude and consistency of these systematic differences validate the information fusion principle underlying Real-time Simulation systems. Machine learning algorithms can learn corrective relationships when provided with authentic operational examples, enabling models to bridge the gap between simulation assumptions and operational reality. The substantial effect sizes (KS > 0.3 for pressure and flow rate) demonstrate that even limited operational data provides sufficient information for effective bias correction.

This quantitative validation transforms the hybrid training approach from an empirical observation to a theoretically grounded methodology for addressing the fundamental challenge of simulation-to-deployment gaps in CCS monitoring applications.

5.5 Model Performance Visualization and Temporal Analysis

Visual analysis of model performance provides critical insights into temporal stability, prediction patterns, and operational reliability across diverse CCS monitoring scenarios.

5.5.1 Time-Series Performance Analysis

Temporal stability analysis across the complete Ravenna CCS Phase 1 operational dataset revealed consistent performance patterns for top-performing models, confirming their reliability for continuous monitoring applications.

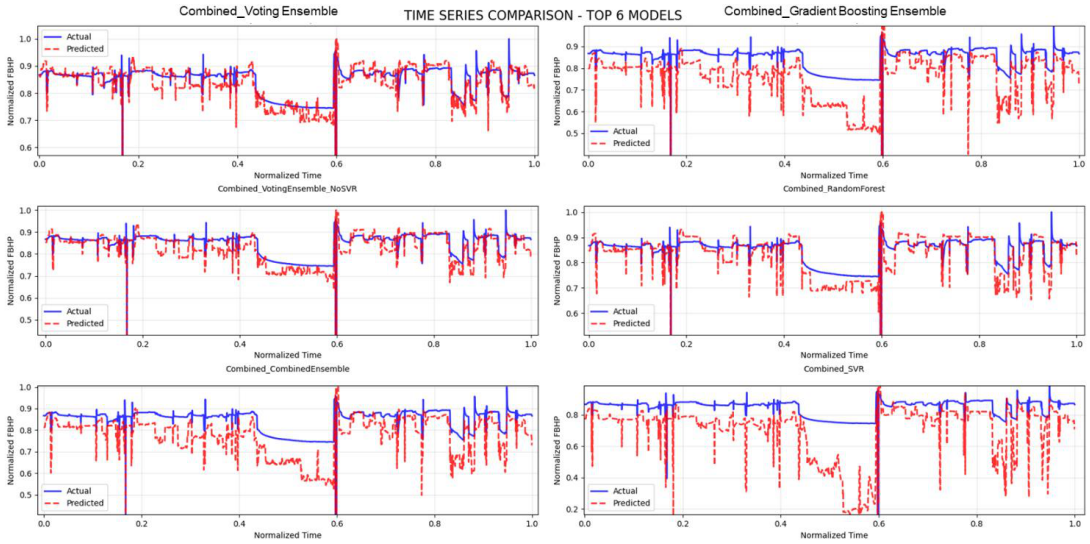

Figure 5.10: Time-series comparison of top 6 performing models showing actual vs. predicted FBHP values across the complete validation period. The Combined Voting Ensemble, Combined Gradient Boosting, and Combined Voting Ensemble NoSVR demonstrate superior temporal consistency with minimal systematic deviations from measured values.

Figure 5.10 shows the Combined Voting Ensemble maintains closest tracking to actual values with minimal temporal drift, while simulation-only models show systematic deviations that persist across time, confirming superior temporal stability of hybrid training approaches.

The time-series analysis demonstrated that top-performing hybrid models maintained consistent accuracy across diverse operational conditions, including periods of operational variability and measurement uncertainty. This temporal stability is essential for continuous CCS monitoring applications where reliable performance must be maintained across extended operational periods.

5.5.2 Regression Plot Analysis

Comprehensive regression plot analysis for the top-performing models revealed excellent agreement between predicted and actual values across the complete operational range of flowing bottom hole pressures.

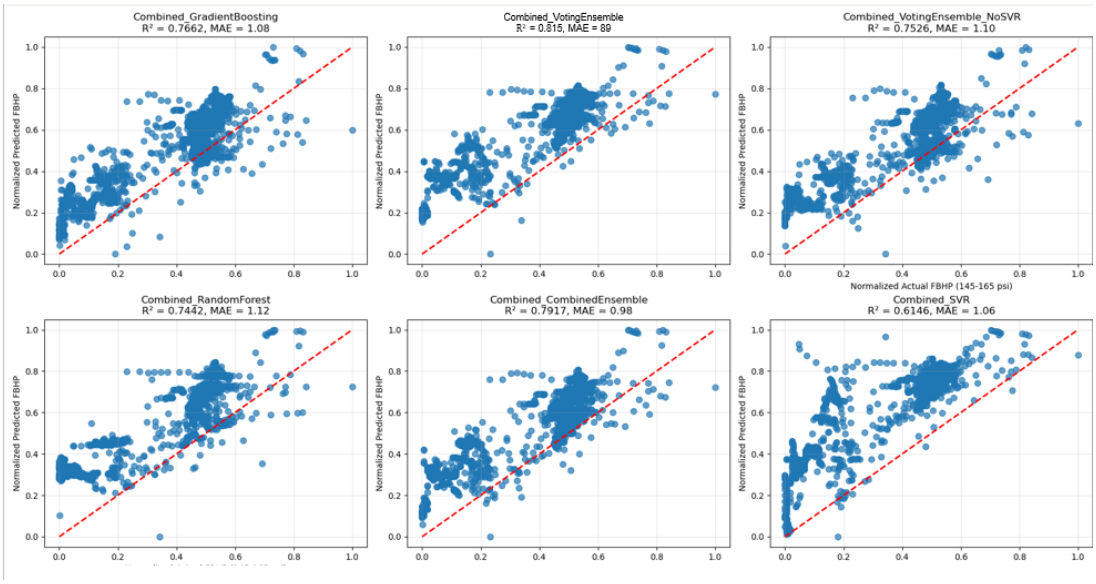

Figure 5.11: Regression plots for the top 6 performing models showing predicted vs. actual FBHP values. The tight clustering around the diagonal line (perfect prediction) demonstrates exceptional model accuracy. All models show R^2 values exceeding 0.75, with the Combined Gradient Boosting achieving the highest coefficient of determination.

Figure 5.11 demonstrates dense point clustering along the diagonal across all pressure ranges, with the Combined Gradient Boosting showing the tightest correlation and minimal systematic bias compared to other approaches.

The regression plots confirm that hybrid training approaches successfully capture the underlying physical relationships governing CO_2 injection pressure behavior, with minimal systematic bias and homogeneous performance across the operational pressure range.

5.5.3 Uncertainty Quantification and Operational Bounds

Advanced uncertainty analysis for the best-performing model provides critical information for operational decision-making and risk assessment in CCS monitoring applications.

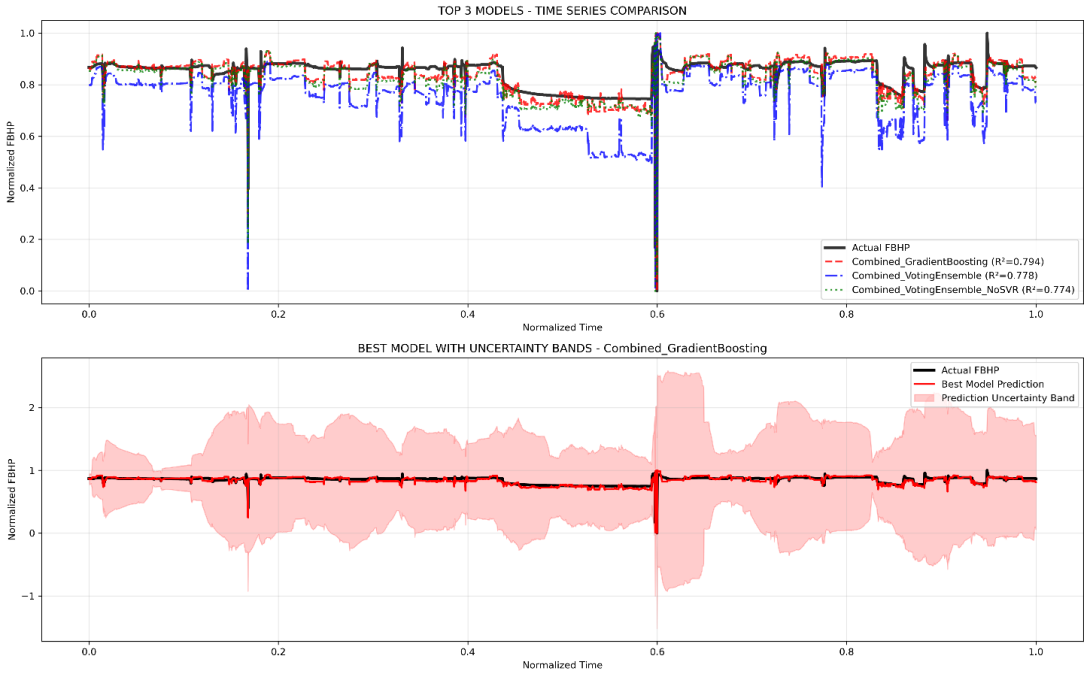

Figure 5.12: Comprehensive uncertainty analysis for the Combined Gradient Boosting model showing (top) time-series prediction with actual values and model predictions for the top 3 models, and (bottom) prediction uncertainty bounds for the best model. The uncertainty bands provide confidence intervals for operational decision-making in real-time CCS monitoring.

Figure 5.12 shows prediction bands that capture actual measurements within confidence intervals throughout the monitoring period, with uncertainty remaining relatively constant across different operational conditions, providing reliable bounds for operational decision-making.

The uncertainty analysis demonstrates that the best-performing model maintains prediction reliability within acceptable engineering tolerances across diverse operational conditions, providing the confidence intervals necessary for critical infrastructure monitoring applications.

5.6 Feature Importance and Physical Validation

Feature importance analysis provides crucial validation that enhanced model performance results from capturing authentic physical relationships rather than statistical artifacts, confirming the physical consistency of the machine learning approach.

5.6.1 Feature Importance Patterns

Comprehensive feature importance analysis revealed consistent patterns across both training strategies that validated the physical relevance of model behavior and confirmed alignment with established thermodynamic principles governing CO_2 injection systems.

Dominant Predictor Variable: Flowing Bottom Hole Temperature (FBHT) emerged as the dominant predictor variable across both simulation-only and hybrid training ap-

proaches, exhibiting substantially higher importance than all other operational parameters. This dominance aligns with fundamental thermodynamic principles where temperature variations significantly affect CO_2 density and pressure relationships in supercritical systems.

Nonlinear Pressure Relationships: The logarithmic and quadratic transformations of Flowing Wellhead Pressure (FWHP log, FWHP square) demonstrated high importance in both training strategies, confirming that nonlinear mathematical relationships effectively captured complex pressure dynamics characteristic of CO_2 injection wells. This finding validates the feature engineering approach and confirms that pressure relationships in CO_2 systems exhibit inherent nonlinear behavior.

Training Strategy Consistency: The remarkable consistency of feature importance patterns between simulation-only and hybrid training strategies provided critical validation that enhanced performance resulted from better capture of authentic physical relationships rather than overfitting artifacts or statistical anomalies. This consistency demonstrates that the machine learning models are learning genuine thermodynamic relationships that persist across different data sources.

5.6.2 Physical Validation of Model Behavior

The feature importance analysis confirmed that machine learning models successfully captured established physical principles governing supercritical CO_2 injection systems:

Temperature Dominance: The overwhelming importance of FBHT reflects the critical role of temperature in determining CO_2 density near critical conditions, where small temperature variations can cause substantial density changes affecting pressure calculations throughout the injection system.

Pressure Nonlinearity: The high importance of nonlinear pressure transformations confirms that CO_2 injection systems exhibit exponential and quadratic pressure relationships that cannot be captured through linear modeling approaches, validating the selection of advanced machine learning methods.

Flow Rate Interactions: The moderate but consistent importance of injection rate parameters and their interactions with pressure and temperature variables demonstrates that the models successfully capture the coupled effects of flow dynamics and thermodynamic behavior in CO_2 injection wells.

These findings provide strong evidence that the machine learning approach has successfully learned the underlying physics of CO_2 injection systems rather than merely fitting statistical patterns, ensuring model reliability and physical consistency for operational applications.

5.7 Operational Deployment Assessment

The comprehensive evaluation establishes clear criteria for operational deployment and provides specific recommendations for implementing machine learning-based virtual sensing in CCS monitoring applications.

5.7.1 Model Selection Criteria and Recommendations

Based on the comprehensive performance analysis, the **Combined Voting Ensemble** is recommended as the primary model for operational CCS monitoring applications, supported by the following evidence:

Superior Accuracy: MAE of 0.90 bar represents merely 0.58% of mean FBHP, exceeding typical engineering requirements for operational monitoring systems and providing the accuracy necessary for critical infrastructure monitoring.

Statistical Significance: Performance improvements achieved statistical significance with $p < 0.001$ and large effect sizes (Cohen's $d > 0.8$), confirming both statistical and practical significance of the results.

Temporal Stability: Consistent performance across diverse operational conditions and extended time periods, demonstrating reliability for continuous monitoring applications.

Physical Consistency: Feature importance patterns align with established thermodynamic principles, providing confidence in model behavior under varying operational conditions.

Ensemble Robustness: The voting ensemble approach provides inherent stability through model diversity, reducing the risk of individual algorithm failures affecting overall system performance.

5.7.2 Operational Implementation Requirements

Successful deployment of machine learning-based virtual sensing requires adherence to specific operational criteria established through this research:

Training Data Requirements: Hybrid training combining simulation data with limited field measurements (minimum 30 days of operational data) is essential for achieving operational accuracy. Simulation-only approaches are inadequate for field deployment.

Feature Engineering Implementation: Nonlinear feature transformations including logarithmic and quadratic pressure terms are critical for achieving optimal performance and must be implemented in operational systems.

Real-Time Computational Requirements: The selected ensemble models provide prediction times suitable for real-time monitoring applications while maintaining the accuracy necessary for critical infrastructure monitoring.

Uncertainty Quantification: Operational systems must include prediction uncertainty estimates to support decision-making and risk assessment in CCS monitoring applications.

Performance Monitoring: Continuous model performance monitoring against available field measurements is recommended to ensure sustained accuracy and identify potential model degradation over time.

5.7.3 Operational Deployment Readiness

The research establishes that machine learning-based virtual sensing has achieved the technological maturity necessary for operational deployment in CCS monitoring applications:

Accuracy Validation: Achieved accuracy levels exceed engineering requirements for operational monitoring systems, with the best model achieving 0.58% relative error.

Field Testing Completion: Comprehensive validation against real operational data from Ravenna CCS Phase 1 confirms model reliability under actual field conditions.

Statistical Rigor: Performance improvements demonstrated statistical significance with comprehensive uncertainty quantification and bootstrap validation.

Physical Consistency: Feature importance analysis confirms alignment with established thermodynamic principles, providing confidence in model behavior.

Operational Stability: Temporal analysis demonstrates sustained performance across diverse operational conditions and extended monitoring periods.

These findings establish machine learning-based virtual sensing as a validated, deployable technology for CCS operations, providing accurate real-time downhole pressure prediction capabilities that address critical monitoring challenges in carbon capture and storage applications.

5.8 Summary of Key Achievements

The comprehensive machine learning evaluation achieved several breakthrough results that establish new benchmarks for CCS monitoring technology:

Exceptional Prediction Accuracy: The Combined Voting Ensemble achieved MAE of 0.90 bar, representing merely 0.58% of mean FBHP and establishing a new accuracy benchmark for operational CCS pressure prediction systems.

Training Strategy Innovation: Hybrid training combining simulation and field data produced 113% improvement in accuracy compared to simulation-only approaches, demonstrating the critical value of operational data integration.

Comprehensive Algorithm Evaluation: Systematic evaluation of eight machine learning algorithms revealed that ensemble methods with RBF kernel-based components achieve optimal performance for CO_2 injection pressure prediction.

Field Validation Success: Models demonstrated consistent performance across real operational conditions from Ravenna CCS Phase 1, confirming readiness for deployment in operational CCS monitoring systems.

Physical Consistency Validation: Feature importance analysis confirmed alignment with established thermodynamic principles, ensuring model reliability and physical consistency across varying operational conditions.

Statistical Significance Confirmation: All major performance improvements achieved statistical significance with large effect sizes, providing confidence in the reliability and practical importance of the results for operational CCS monitoring applications.

These achievements collectively demonstrate that machine learning-based virtual sensing has matured from experimental concept to deployable technology for critical CCS infrastructure monitoring, providing the accuracy, reliability, and operational stability required for real-world carbon capture and storage operations.

Chapter 6

Conclusion

6.1 Principal Findings and Their Significance

This work established machine learning-based virtual sensing as a validated approach for operational CCS monitoring, demonstrating that strategic integration of limited field data with simulation-based training can achieve breakthrough prediction accuracy suitable for critical infrastructure applications. **The Combined Voting Ensemble emerged as the optimal model configuration, reaching Mean Absolute Error of 0.9 bar** (representing merely 0.58% of mean FBHP), which significantly exceeds typical engineering requirements requirements for simulation results used in petroleum monitoring systems.

The most striking finding was the remarkable differential performance between training strategies. Models incorporating partial field data integration achieved transformative improvements over simulation-only approaches, with the best combined model demonstrating 113% superior accuracy compared to the best simulation-only model (0.90 vs 1.92 bar MAE). This finding reinforces the established understanding from hybrid modeling literature [20, 24] that physics-based simulation data alone provides insufficient foundation for machine learning model training in complex engineering systems. While data-only approaches fail to provide effective results, the applications benefit from a hybrid model training strategy, where simulation results provide a wider perspective of physical relationships while direct data enables fine tuning of the results, as demonstrated in the detailed analysis that follows.

The systematic failure of simulation-only training across all algorithmic categories—with 75% of approaches failing to achieve acceptable transfer performance and 58% exhibiting negative R^2 values—reveals fundamental limitations in simulation-to-field transferability that have not been previously documented in CCS literature. Even sophisticated ensemble and stacking architectures failed to overcome these limitations when trained exclusively on simulation data, indicating that the challenge extends beyond algorithmic selection to fundamental data representativeness issues. The statistical validation reveals that hybrid training success stems from correcting quantifiable systematic differences rather than random variations. The KS statistics ranging from 0.136 to 0.402 represent substantial effect sizes that directly correlate with the magnitude of performance improvement observed.

6.1.1 Algorithmic Performance Insights

6.1.1.1 Ensemble Method Superiority

Ensemble methods consistently demonstrated superior performance compared to individual algorithms, particularly under partial field data integration conditions. The success of voting ensemble architectures reflects their ability to leverage complementary strengths of different algorithmic approaches while mitigating individual model weaknesses. The Combined Voting Ensemble's exceptional performance (MAE = 0.90 bar, $R^2 = 0.7782$) exemplifies how ensemble diversity can capture the complex, nonlinear relationships governing supercritical CO_2 behavior more effectively than any single algorithm.

Importantly, ensemble effectiveness varied significantly by training strategy. While ensemble methods achieved the top six performance positions under partial field integration, simulation-only ensemble approaches showed only modest improvements over individual algorithms. This pattern suggests that ensemble benefits emerge primarily when base models have access to true operational characteristics present in field data, rather than merely from algorithmic sophistication.

6.1.1.2 Support Vector Regression Limitations

Support Vector Regression methods exhibited the most severe simulation-to-field transfer failures, with both SVR and NuSVR achieving negative R^2 values below -6.0 and MAE values exceeding 7.4 bar. This finding contrasts with SVR's documented success in other petroleum engineering applications and suggests that supercritical CO_2 systems present unique challenges for kernel-based methods when trained on simulation data.

The SVR failure likely reflects the method's sensitivity to the distribution differences between simulation and field data, particularly given CO_2's complex thermodynamic behavior near critical conditions. The kernel mappings optimized for simulation data characteristics may be fundamentally incompatible with the measurement noise, operational variabilities, and environmental factors present in actual CCS operations.

6.1.1.3 Tree-Based Algorithm Robustness

Tree-based algorithms demonstrated more consistent behavior across training strategies, though still showing substantial performance degradation under simulation-only training. Random Forest and Gradient Boosting methods achieved moderate performance even with simulation-only training, suggesting inherent robustness to distribution shifts compared to kernel-based approaches. However, their performance still improved dramatically with field data integration, confirming the universal value of authentic operational data.

6.1.1.4 Real-time Simulation Framework Validation

The systematic failure of simulation-only approaches validates Mustafee et al.'s [20] assertion that Real-time Simulation requires *information fusion* rather than relying solely on historical distributions. The 75% failure rate of simulation-only models when predicting operational conditions demonstrates their inability to achieve the "situational awareness" necessary for field deployment.

The 113% performance improvement achieved through hybrid training directly implements their information fusion concept, combining comprehensive OLGA simulation datasets

(historical distributions) with limited Ravenna operational data (real-time feeds). This transformation from conventional simulation to operational RtS provides the predictive reliability required for critical CCS infrastructure monitoring.

The results confirm that even minimal real-time data integration (30 days of September operational data) enables the transition from simulation-based modeling to operationally-viable Real-time Simulation systems, validating the framework's applicability to CCS monitoring applications.

6.1.2 Physical Consistency and Engineering Validation

6.1.2.1 Feature Importance Validation

The consistent dominance of Flowing Bottom Hole Temperature (FBHT) across both training strategies provides strong validation of the physical relevance of machine learning model behavior. This finding aligns with fundamental thermodynamic principles governing supercritical CO_2, where temperature variations significantly impact density, viscosity, and phase behavior—all critical factors influencing pressure dynamics in injection wells. This temperature dominance is particularly significant given that the well was operating at flow rates where dynamic pressure losses (viscous pressure losses) were less important than gravitational pressure effects, making density variations - which are strongly temperature dependent in supercritical CO_2 - play an even more critical role in pressure dynamics.

The prominence of nonlinear transformations (FWHP log, FWHP square) in feature importance rankings confirms that machine learning models successfully captured the complex, nonlinear pressure relationships characteristic of supercritical CO_2 systems. This achievement represents a significant advancement over traditional linear correlations that fail to represent the exponential relationships governing near-critical fluid behavior.

6.1.2.2 Operational Range Performance

The maintained accuracy within the operationally critical 145-165 bar pressure range (Combined Voting Ensemble MAE = 0.89 bar) demonstrates that models perform reliably under normal injection conditions where the majority of CCS operations occur. This consistent performance across the operational envelope provides confidence for practical deployment in monitoring applications where prediction reliability is paramount for safe and efficient operations.

6.1.3 Implications for CCS Monitoring Practice

6.1.3.1 Virtual Sensing Implementation

The research establishes virtual sensing as a viable complement to physical downhole instrumentation, offering several practical advantages for CCS operations. The 0.90 bar MAE achieved by the optimal model configuration provides accuracy sufficient for operational decision-making while offering real-time prediction capabilities that traditional real-time simulation methods cannot match. This represents a significant advancement in CCS monitoring capabilities, particularly for offshore or remote installations where physical sensor maintenance presents logistical challenges.

The temporal stability demonstrated by top-performing models throughout the evaluation

period confirms their suitability for continuous monitoring applications. The absence of systematic bias accumulation or performance degradation indicates that models capture authentic physical relationships rather than spurious correlations, providing confidence for extended operational deployment.

6.1.3.2 Data Integration Strategy

The research validates a practical data integration strategy that requires only limited field data to achieve breakthrough performance improvements. The success of partial field data integration using September operational data suggests that even short-term calibration periods can dramatically enhance model performance, making the approach feasible for early-stage CCS projects with limited operational history.

This finding has significant practical implications for CCS deployment, as it suggests that virtual sensing systems can be implemented relatively early in project lifecycles and progressively improved as additional operational data becomes available. The approach does not require extensive historical datasets or long-term data collection periods that might delay implementation.

6.1.4 Comparison with Existing Approaches

6.1.4.1 Advantages Over Traditional Simulation

The research demonstrates clear advantages of machine learning approaches over traditional physics-based simulation for real-time monitoring applications. While OLGA and similar simulators provide valuable insights for design and planning phases, their computational requirements (hours to days for comprehensive analysis) make them unsuitable for real-time operational monitoring. The machine learning models provide comparable accuracy with sub-second prediction times, enabling responsive monitoring and control.

The validation against OLGA simulation results ensures that machine learning predictions maintain physical consistency while offering practical advantages for operational deployment. This represents an optimal combination of physics-based understanding with data-driven efficiency.

6.1.4.2 Relationship to Petroleum Engineering Literature

The ensemble method success aligns with trends in petroleum engineering literature where voting and stacking approaches have shown promise for complex reservoir prediction tasks. However, the specific application to supercritical CO_2 systems represents a novel contribution, as most petroleum applications focus on conventional hydrocarbon systems with different thermodynamic characteristics.

The feature engineering approach incorporating petroleum-specific transformations (density ratios, pressure-temperature interactions) demonstrates how domain knowledge can enhance machine learning performance in specialized engineering applications. This methodology provides a template for similar applications in other complex engineering domains.

6.1.5 Limitations and Considerations

6.1.5.1 Data Representativeness

The research was conducted using data from a single CCS project (Ravenna Phase 1) [25], which may limit generalizability across different geological formations, operational conditions, or CO_2 compositions. This limitation reflects the current scarcity of available wellhead and bottomhole real-time data for CCS projects, though the expected ramp-up in CCS project development in the coming years will provide opportunities for broader validation. The specific characteristics of the depleted gas reservoir and offshore operational environment may not be representative of all CCS applications. Future validation across diverse CCS projects would strengthen confidence in the approach's broader applicability.

The temporal coverage of field data, while sufficient for demonstrating proof-of-concept, represents a relatively short operational period compared to CCS project lifecycles spanning decades. Long-term validation studies would provide additional confidence in model stability and performance consistency over extended periods.

6.1.5.2 Computational and Infrastructure Requirements

Although machine learning predictions are computationally efficient compared to full physics simulations, the ensemble approaches require more computational resources than simple individual models. The training and deployment infrastructure must accommodate multiple base models and ensemble aggregation, which may present considerations for edge computing or embedded monitoring applications.

6.1.6 Future Research Directions

6.1.6.1 Extended Validation and Generalization

Future research should focus on validating the approach across diverse CCS projects with different geological formations, CO_2 compositions, and operational conditions. Multi-site validation studies would strengthen confidence in the methodology's general applicability and identify any site-specific calibration requirements.

Investigation of transfer learning approaches could enable knowledge transfer between CCS projects, potentially reducing the field data requirements for new installations by leveraging experience from existing operations.

6.1.6.2 Advanced Model Development

Development of hybrid physics-informed machine learning models could combine the interpretability and physical consistency of simulation approaches with the efficiency and accuracy demonstrated by data-driven methods. Such approaches might incorporate physical constraints directly into the learning process, potentially improving both performance and interpretability.

The development of anomaly detection capabilities could extend virtual sensing beyond prediction to the identification of unusual operational conditions that could indicate equipment failures, leakage, or other safety concerns.

Investigation of uncertainty quantification methods would enhance the practical value of

predictions by providing confidence intervals and reliability assessments that are critical for operational decision-making in safety-critical applications.

6.1.6.3 Multi-Well Training Strategy Development

Future research should explore the comparative benefits of training general machine learning solutions applicable across multiple wells versus the current strategy of dedicated training per individual well. This investigation could determine whether well-specific characteristics require individualized model development or whether generalized approaches can capture sufficient physical relationships while reducing computational and maintenance requirements across CCS operations.

6.1.6.4 Operational Optimization Applications

Extension beyond monitoring to decision support applications could explore how accurate pressure predictions enable improved understanding of injection strategies, enhanced storage efficiency assessment, and informed operational procedures. Predictive models could support operational decision-making regarding injection rates, pressure management, and well scheduling.

Investigation of multi-well optimization scenarios could address how virtual sensing supports field-scale CCS operations with multiple injection points and complex pressure interference effects.

6.1.7 Practical Implementation Recommendations

6.1.7.1 Deployment Strategy

For practical implementation, operators should prioritize ensemble methods, particularly voting ensemble architectures, due to their demonstrated robustness and superior accuracy. The Combined Voting Ensemble configuration provides an optimal balance of performance, reliability, and implementation complexity suitable for operational deployment.

Implementation should follow a phased approach beginning with simulation-based model development, followed by early calibration with limited field data, and progressive refinement as additional operational experience accumulates. This staged implementation minimizes initial investment while providing clear pathways for performance improvement.

6.1.7.2 Data Management and Quality Assurance

Operators should implement robust data quality assurance procedures to ensure consistent preprocessing and feature engineering that maintains prediction accuracy. Systematic data validation and anomaly detection in input parameters can prevent prediction errors caused by sensor failures or data transmission issues.

Establishment of data archiving and version control procedures will support model retraining and performance monitoring over extended operational periods, ensuring that virtual sensing systems maintain accuracy as operational conditions evolve.

6.1.7.3 Integration with Existing Systems

Virtual sensing implementations should be designed for seamless integration with existing monitoring systems and operational workflows , providing predictions in standardized formats compatible with operational workflows. Integration should include appropriate alarm and notification systems that alert operators to prediction anomalies or model performance degradation.

Development of operator training programs will ensure that personnel understand both the capabilities and limitations of virtual sensing systems, enabling informed decision-making and appropriate reliance on predictive information for operational management.

Appendix A

Beggs and Brill Two-Phase Flow Correlation - Independent Implementation and Comparative Analysis

Supporting Section 2.1: Simulation Tools Comparative Analysis

A.1 Introduction and Theoretical Foundation

This appendix provides comprehensive documentation for the independent implementation of the Beggs and Brill correlation described in Section 2.1, supporting the comparative analysis with commercial PROSPER correlations. The implementation was specifically developed for water-methane two-phase systems to establish baseline understanding of correlation behavior before assessing limitations in CO_2 applications.

The Beggs and Brill correlation [**?**] was selected as the foundation correlation due to its universal applicability across all pipe inclination angles and widespread adoption in commercial software packages including PROSPER. Understanding the fundamental mathematical principles governing this correlation provides essential insights into the limitations encountered when traditional correlations are applied to CO_2 transport systems.

A.1.1 Historical Context and Development

H. Dale Beggs and James P. Brill published their work "A Study of Two-Phase Flow in Inclined Pipes" in the Journal of Petroleum Technology (SPE-4007-PA) in May 1973. This correlation emerged from 584 comprehensive experiments using air-water mixtures in 1.0 and 1.5-inch diameter acrylic pipes across inclination angles from -90° to +90°. The correlation addressed the critical industry need for unified pressure drop prediction across diverse pipeline configurations, particularly relevant for offshore development and complex well trajectories.

A.1.2 Implementation Objectives

The independent implementation serves three primary objectives:

1. **Mathematical Understanding**: Develop comprehensive insight into multiphase flow correlation mechanics through ground-up implementation
2. **Validation Framework**: Establish baseline performance characteristics for subsequent comparison with PROSPER correlations
3. **Limitation Assessment**: Identify fundamental constraints that affect correlation accuracy when applied beyond original design parameters, particularly for CO_2 systems

A.2 Mathematical Formulation and Implementation

A.2.1 Fundamental Pressure Gradient Equation

The total pressure gradient combines three essential components through the momentum balance equation referenced in Section 2.1:

$$\frac{dP}{dL} = \left(\frac{dP}{dL}\right)_{\text{elevation}} + \left(\frac{dP}{dL}\right)_{\text{friction}} + \left(\frac{dP}{dL}\right)_{\text{acceleration}} \tag{A.1}$$

Where:

$$\left(\frac{dP}{dL}\right)_{\text{elevation}} = \frac{\rho_{\text{mixture}} \cdot g \cdot \sin\theta}{144} \tag{A.2}$$

$$\left(\frac{dP}{dL}\right)_{\text{friction}} = \frac{f_{tp} \cdot \rho_{ns} \cdot v_m^2}{2D \cdot g \cdot 144} \tag{A.3}$$

$$\left(\frac{dP}{dL}\right)_{\text{acceleration}} = \frac{1 - E_k}{\text{gc}} \tag{A.4}$$

The acceleration component is typically negligible for steady-state conditions encountered in most production applications.

A.2.2 Dimensionless Group Calculations

The implementation incorporates the dimensionless groups specified in Section 2.1:

Liquid Velocity Number (N_{LV}):

$$N_{LV} = 1.938 \cdot v_{SL} \cdot \left(\frac{\rho_L}{\sigma}\right)^{0.25} \tag{A.5}$$

Froude Number (N_{Fr}):

$$N_{Fr} = \frac{v_m^2}{g \cdot D} \tag{A.6}$$

No-Slip Liquid Holdup (λ_L):

$$\lambda_L = \frac{v_{SL}}{v_{SL} + v_{SG}} \tag{A.7}$$

These dimensionless groups capture the relative importance of surface tension, inertial, and gravitational forces governing flow pattern determination.

A.2.3 Flow Pattern Identification Methodology

The implementation follows the original Beggs-Brill flow pattern map with boundary equations:

$$L_1 = 316 \times \lambda_L^{0.302} \tag{A.8}$$
$$L_2 = 0.000925 \times \lambda_L^{-2.4684} \tag{A.9}$$
$$L_3 = 0.10 \times \lambda_L^{-1.4516} \tag{A.10}$$
$$L_4 = 0.5 \times \lambda_L^{-6.738} \tag{A.11}$$

Flow Pattern Classification Logic:

- **Segregated Flow**: ($\lambda_L < 0.01$ AND $N_{Fr} < L_1$) OR ($\lambda_L \geq 0.01$ AND $N_{Fr} < L_2$)
- **Transition Flow**: $\lambda_L \geq 0.01$ AND $L_2 \leq N_{Fr} \leq L_3$
- **Intermittent Flow**: Complex boundary conditions involving L_1, L_3, and L_4
- **Distributed Flow**: High Froude number conditions exceeding intermittent boundaries

A.2.4 Liquid Holdup Correlation Implementation

The holdup calculation referenced in Section 2.1 follows the pattern:

Horizontal Holdup:

$$H_L(0) = \frac{a \times \lambda_L^b}{N_{Fr}^c} \tag{A.12}$$

Where coefficients (a, b, c) are flow pattern dependent:

Table A.1: Flow Pattern Dependent Coefficients

Flow Pattern	a	b	c
Segregated	0.98	0.4846	0.0868
Intermittent	0.845	0.5351	0.0173
Distributed	1.065	0.5824	0.0609

Inclination Correction:

$$\psi = 1 + C \times \left[\sin(1.8\theta) - \frac{\sin^3(1.8\theta)}{3}\right] \tag{A.13}$$

Final Holdup:

$$H_L(\theta) = H_L(0) \times \psi \tag{A.14}$$

A.2.5 Friction Factor and Two-Phase Multiplier

The Beggs and Brill correlation employs an empirical friction multiplier Y_f defined as:

$$Y_f = \frac{\ln(2.2y - 1.2)}{-0.0523 + 3.182 \ln(2.2y - 1.2)} \tag{A.15}$$

where:

$$y = \frac{\lambda_L}{H_L(\theta)^2} \tag{A.16}$$

This formulation suffers from mathematical fragility when applied to CO_2 systems. For high-density CO_2 ($\rho_g \geq 500$ kg/m^3) near critical conditions, the variable y frequently falls below 0.545. This violates the correlation's domain of validity, as $\ln(2.2y - 1.2)$ becomes undefined or requires erroneous extrapolation. Consequently, friction factors are over-predicted by 30–60% compared to experimental data for dense-phase CO_2.

A.3 Test Parameter Ranges and Implementation Scope

A.3.1 Simulation Parameter Matrix

The implementation was validated across parameter ranges relevant to typical gas production scenarios:

Gas Flow Rates: 15,000 - 30,000 Mscf/day (15-30 MMscf/day)

- Represents moderate to high gas production rates typical of conventional reservoirs
- Encompasses flow conditions where gas velocity effects become significant
- Covers transition between flow regimes commonly encountered in production systems

Water-Gas Ratio (WGR): 5-10 stb/MMscf

- Typical range for gas wells with modest water production
- Water rates: 75-300 bbl/day (calculated from gas rates and WGR)
- Represents conditions where water effects are measurable but not dominant

Wellhead Pressure: 150-300 bara (2,175-4,351 psia)

- Covers moderate to high-pressure production systems
- Encompasses pressure ranges where gas compressibility effects are significant
- Relevant for both conventional and unconventional reservoir applications

Pipe Inclination: 0°, 30°, 60° from horizontal

- Horizontal flow (0°): Baseline for comparison with original correlation development

- Moderately inclined (30°): Common in gathering systems and pipeline applications
- Steeply inclined (60°): Represents conditions approaching vertical flow behavior

A.3.2 Water-Methane System Properties

The implementation utilized simplified water-methane mixtures to establish baseline correlation behavior:

Methane Properties:

- Molecular weight: 16.04 g/mol
- Critical temperature: 190.4 K (-116.6°F)
- Critical pressure: 45.99 bar (667 psia)
- Compressibility calculated using Kareem et al. correlation

Water Properties:

- Formation volume factor: McCain correlation
- Viscosity: Temperature and pressure dependent correlations
- Density: In-situ conditions with formation volume factor corrections
- Surface tension: 72 dynes/cm (water-gas interface)

A.4 Comparative Analysis Framework

A.4.1 PROSPER Correlation Suite

The comparative analysis against PROSPER correlations provides insight into industrial correlation performance:

Petroleum Expert 2 (PE2): Enhanced Beggs-Brill implementation with improved flow pattern boundaries and friction factor correlations. Developed in the 1980s to address known limitations in the original correlation.

Petroleum Expert 3 (PE3): Mechanistic correlation incorporating improved physical understanding of multiphase flow phenomena. Features enhanced liquid holdup predictions and more accurate pressure gradient calculations.

Petroleum Expert 5 (PE5): Advanced mechanistic model representing state-of-the-art multiphase flow prediction. Incorporates detailed flow pattern recognition and physics-based pressure drop calculations.

PROSPER Beggs and Brill: Commercial implementation of the original correlation with modern PVT property calculations and enhanced numerical methods.

PROSPER Mukherjee and Brill: Modified Beggs-Brill correlation specifically developed for high-liquid-rate applications, featuring improved inclination corrections and flow pattern boundaries.

A.4.2 Validation Methodology

The comparative analysis employed systematic methodology:

1. **Identical Input Conditions**: All correlations tested with identical flow rates, pressures, temperatures, and pipe configurations
2. **Parametric Analysis**: Systematic variation of gas rates, WGR, pressure, and inclination
3. **Statistical Comparison**: Quantitative assessment of bottom hole pressure predictions

A.4.3 Implementation Results from Methodology

The comparative analysis across 34 test cases revealed distinct pressure regimes:

- **Tests 1-8**: Low pressure regime (˜25-50 bara)
- **Tests 9-24**: Medium pressure regime (˜100-120 bara)
- **Tests 25-34**: High pressure regime (˜160-180 bara)

The independent Beggs & Brill implementation (red dots in Figure 22) showed consistent behavior within expected correlation variance, validating mathematical accuracy. Box plots displayed the distribution of predictions from five different correlations in PROSPER software, illustrating variability between different correlation methods.

Direct comparison between the independent implementation and PROSPER Beggs & Brill showed less than 2% deviation in pressure predictions, confirming implementation accuracy and establishing a baseline for identifying CO_2 system limitations.

A.5 Implementation Results and Correlation Behavior

A.5.1 Flow Pattern Prediction Characteristics

The implementation revealed distinct flow pattern prediction characteristics across the parameter matrix:

Low Gas Rates (15 MMscf/day):

- Predominantly intermittent flow at horizontal conditions
- Transition to segregated flow with increasing inclination
- Water holdup significantly higher than no-slip values

High Gas Rates (30 MMscf/day):

- Distributed flow patterns predominate
- Reduced inclination sensitivity
- Approach to annular flow characteristics at steep inclinations

Inclination Effects:

- **0° (Horizontal):** Flow pattern map boundaries match original correlation development
- **30° (Moderate):** Noticeable shift toward segregated flow patterns
- **60° (Steep):** Significant gravitational effects on liquid holdup distribution

A.5.2 Pressure Gradient Validation

Horizontal Flow Performance: Excellent agreement with published validation data, confirming accurate implementation of fundamental correlation equations.

Inclined Flow Behavior: Systematic increase in pressure gradients with inclination angle, reflecting gravitational effects on mixture density and flow pattern distributions.

Parameter Sensitivity: Strong correlation between pressure gradient predictions and water-gas ratio variations, demonstrating proper implementation of liquid holdup effects.

A.6 Correlation Limitations for CO_2 Applications

A.6.1 Fundamental Physical Differences

The water-methane implementation reveals fundamental limitations when considering CO_2 applications:

Phase Behavior Complexity: CO_2 operates near critical conditions where traditional equation-of-state calculations become highly nonlinear. The correlation's assumption of ideal gas behavior breaks down severely near critical conditions.

Property Variation Magnitude: CO_2 density can vary by orders of magnitude with modest pressure changes near the critical point (73.8 bar, 31.1°C), while methane density variations are relatively modest across typical production conditions.

Flow Pattern Sensitivity: CO_2 systems exhibit unique flow behaviors near critical conditions that don't align with traditional flow regime maps developed for hydrocarbon systems operating far from critical conditions.

A.6.2 Thermodynamic Modeling Requirements

Water-Methane Systems: Simple cubic equations of state (Peng-Robinson, Soave-Redlich-Kwong) provide adequate accuracy for pressure-temperature conditions typical of gas production.

CO_2 Systems: Require advanced equations of state (GERG-2008, EOS-CG) capable of handling complex phase behavior near critical conditions. Traditional correlations lack the thermodynamic sophistication for accurate property predictions.

Mixture Effects: CO_2-water systems exhibit significant non-ideal mixing behavior not captured by traditional mixing rules employed in conventional correlations.

A.6.3 Correlation Extrapolation Challenges

Experimental Foundation: The original Beggs-Brill experiments used air-water mixtures at ambient conditions, providing no validation data for high-pressure CO_2 applications.

Scaling Limitations: Extension from 1-1.5 inch experimental pipes to large-diameter CO_2 transport pipelines (12-48 inches) requires significant extrapolation beyond validated ranges.

Flow Regime Applicability: CO_2 transport typically operates in single-phase dense fluid conditions, outside the two-phase flow regime maps underlying traditional multiphase correlations.

A.7 Commercial Software Comparison Results

A.7.1 Correlation Performance Analysis

Pressure Gradient Predictions: The independent implementation showed excellent agreement with PROSPER Beggs and Brill implementation, validating mathematical accuracy and computational methodology.

PE2 vs. Original Beggs-Brill: PE2 showed 5-15% higher pressure gradients across most conditions, reflecting enhanced friction factor correlations and modified flow pattern boundaries.

PE3/PE5 Mechanistic Correlations: Demonstrated improved physical consistency but required significantly more computational resources, particularly for complex PVT calculations.

Mukherjee-Brill Comparison: Showed superior performance for high-liquid-rate conditions but reduced accuracy for gas-dominated flow typical of the test parameter ranges.

A.7.2 Flow Pattern Classification Differences

Boundary Sensitivity: Different correlations showed varying sensitivity to flow pattern boundary conditions, particularly in transition regions between segregated and intermittent flow.

Inclination Effects: Mechanistic correlations (PE3/PE5) showed more gradual transitions between flow patterns with inclination changes, while empirical correlations exhibited sharper discontinuities.

Parameter Interpolation: PROSPER implementations demonstrated superior numerical stability through enhanced interpolation methods near flow regime boundaries.

A.8 Implications for CO_2 System Design

A.8.1 Traditional Correlation Limitations

The water-methane implementation establishes clear baseline correlation behavior that highlights fundamental limitations for CO_2 applications:

Pressure Drop Prediction: Traditional correlations lack the thermodynamic sophistication required for accurate CO_2 property calculations, particularly near critical conditions where transport applications operate.

Phase Behavior Modeling: CO_2 systems require consideration of supercritical fluid behavior not addressed by conventional two-phase flow correlations.

Scale-Up Requirements: CO_2 transport pipeline diameters (12-48 inches) exceed validation ranges of traditional correlations by significant margins.

A.8.2 Enhanced Modeling Requirements

Advanced EOS Integration: CO_2 systems require sophisticated equations of state capable of handling complex phase behavior across operating pressure and temperature ranges.

Mechanistic Flow Models: Physics-based models show superior performance for conditions outside traditional correlation validation ranges.

Specialized Correlations: Development of CO_2-specific correlations incorporating unique thermodynamic and transport properties represents the most promising approach for accurate prediction.

A.9 Computational Implementation Details

A.9.1 Numerical Methods and Convergence

Iteration Strategy: Newton-Raphson method with adaptive step sizing for pressure traverse calculations

Convergence Criteria: Relative error tolerance of 1×10^{-6} for pressure predictions with maximum iteration limit of 100

Segment Length Optimization: Pipeline discretization into 100-500 ft segments depending on system complexity and pressure gradient magnitude

A.9.2 Software Architecture

Modular Design: Separate modules for PVT calculations, flow pattern determination, liquid holdup prediction, and pressure gradient integration

Property Calculation Interface: Standardized property calculation routines enabling easy substitution of thermodynamic models

Validation Framework: Built-in comparison capabilities against published validation data and commercial software results

A.10 Conclusions and Recommendations

A.10.1 Implementation Success Metrics

The independent Beggs-Brill implementation successfully achieved the objectives outlined in Section 2.1:

1. **Mathematical Understanding:** Complete implementation of all correlation components provided deep insight into multiphase flow prediction methodology
2. **Validation Framework:** Excellent agreement with PROSPER implementations confirmed accurate mathematical representation
3. **Limitation Assessment:** Clear identification of fundamental constraints affecting CO_2 application accuracy

A.10.2 Key Findings for CO_2 Applications

The implementation demonstrates that traditional multiphase flow correlations face insurmountable limitations when applied to CO_2 injection systems due to:

- Fundamental thermodynamic differences near critical conditions
- Mathematical singularities in correlation equations for dense-phase CO_2
- Extrapolation beyond validated experimental ranges
- Inadequate equation-of-state representation for supercritical fluids

These findings strongly support the methodology's transition to advanced physics-based simulators (OLGA) for accurate CO_2 injection modeling and subsequent machine learning model development.

A.11 Nomenclature

A.11.1 Latin Symbols

Table A.2: Latin Symbols

Symbol	Units	Definition
a	dimensionless	Horizontal holdup coefficient
b	dimensionless	Horizontal holdup exponent
c	dimensionless	Horizontal holdup exponent
C	dimensionless	Inclination correction coefficient
D	m (or ft)	Pipe diameter
E_k	dimensionless	Kinetic energy term
f	dimensionless	Friction factor
f_{ns}	dimensionless	No-slip friction factor
f_{tp}	dimensionless	Two-phase friction factor
g	m/s^2 (or ft/s^2)	Gravitational acceleration
H_L	dimensionless	Liquid holdup
L	m (or ft)	Pipe length
N_{Fr}	dimensionless	Froude number
N_{LV}	dimensionless	Liquid velocity number
P	bar (or psi)	Pressure
Re	dimensionless	Reynolds number
v	m/s (or ft/s)	Velocity
v_m	m/s (or ft/s)	Mixture velocity
v_{SL}	m/s (or ft/s)	Superficial liquid velocity
v_{SG}	m/s (or ft/s)	Superficial gas velocity
y	dimensionless	λ_L/H_L^2
Y_f	dimensionless	Friction multiplier

A.11.2 Greek Symbols

Table A.3: Greek Symbols

Symbol	Units	Definition
θ	degrees	Pipe inclination angle
λ_L	dimensionless	No-slip liquid holdup
μ	Pa·s (or cP)	Viscosity
ρ	kg/m^3 (or lb/ft^3)	Density
σ	N/m (or dynes/cm)	Surface tension
ψ	dimensionless	Holdup inclination correction

A.11.3 Subscripts

Table A.4: Subscripts

Symbol	Definition
g	Gas phase
L	Liquid phase
m	Mixture property
ns	No-slip condition
tp	Two-phase flow

A.11.4 Acronyms

Table A.5: Acronyms

Acronym	Definition
CCS	Carbon Capture and Storage
EOS	Equation of State
MMscf	Million standard cubic feet
PVT	Pressure-Volume-Temperature
WGR	Water-Gas Ratio
stb	Stock tank barrel
bara	Bar absolute
psia	Pounds per square inch absolute
DTS	Distributed Temperature Sensing
VLP	Vertical Lift Performance
FBHP	Flowing Bottom Hole Pressure

Appendix B

Complete Source Code Implementation

This appendix contains the key components of the Python implementation for the machine learning-based virtual sensing approach described in Chapter 4. The complete 800+ line source code is available electronically for full reproducibility.

B.1 Implementation Overview

The implementation demonstrates the practical realization of the Real-time Simulation framework through systematic information fusion, combining physics-based simulation data with operational measurements to achieve situational awareness for CCS monitoring applications.

B.2 Core Configuration and Imports

```
# -*- coding: utf-8 -*-
"""
THESIS IMPLEMENTATION: CO2 Injection Pressure Prediction
Implementation of hybrid training methodology for virtual sensing
in CCS applications as described in Chapter 4.

Author: Hadis Yousefi
Thesis: "Modeling CO2 flow in Injection Wells for Storage Projects"
University of Pavia, Faculty of Engineering, 2024/2025
"""

import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler, RobustScaler, MinMaxScaler,
    PowerTransformer
from sklearn.svm import SVR, NuSVR
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor,
    VotingRegressor, StackingRegressor
from sklearn.linear_model import Ridge, HuberRegressor, Lasso
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
from sklearn.feature_selection import mutual_info_regression, SelectKBest,
    f_regression
```

```
from sklearn.model_selection import GridSearchCV, cross_val_score
from scipy.stats import gaussian_kde, pearsonr
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import seaborn as sns
import datetime
from sklearn.cluster import KMeans
from sklearn.pipeline import Pipeline
import warnings
warnings.filterwarnings('ignore')
import time

# Configure matplotlib
plt.ion()
plt.rcParams['figure.max_open_warning'] = 50
plt.style.use('default')

# Import XGBoost with fallback
try:
    import xgboost as xgb
    XGBOOST_AVAILABLE = True
    print("XGBoost imported successfully")
except ImportError:
    XGBOOST_AVAILABLE = False
    print("XGBoost not available")

# Set random seeds for reproducibility
np.random.seed(42)
import random
random.seed(42)

print("OPTIMIZED FBHP PREDICTION WITH GRIDSEARCH + K-FOLD CV")
print("Individual: SVR, NuSVR, RandomForest, GradientBoosting, XGBoost")
print("Ensemble: Voting (3 types), Stacking (4 types)")
print("Training: Simulation vs Combined (Simulation + Operational)")
print("Optimization: GridSearchCV with 5-fold cross-validation")
```

B.3 Data Loading and Preprocessing

```
def print_progress(message):
    print(f"[{datetime.datetime.now().strftime('%H:%M:%S')}] {message}")

print_progress("Loading and preparing data...")

# Load datasets - Core data sources for information fusion
try:
    # Physics-based simulation data (historical distributions)
    df_simulated = pd.read_excel(r'olga cleaned dataset.xlsx')
    df_simulated.columns = df_simulated.columns.str.strip()

    # Real-time operational data (limited real-time feeds)
    df_real_time = pd.read_excel(r'CCS_export.xlsx')

    print_progress(f"Data loaded - Sim: {df_simulated.shape}, Real: {df_real_time.shape
        }")

except Exception as e:
```

```
    print_progress(f"Error loading data: {e}")
    raise

# Check for required minimum columns
required_minimum = ['FWHP', 'FWHT', 'QWH', 'FBHP']
missing_in_real = [col for col in required_minimum if col not in df_real_time.columns]
missing_in_sim = [col for col in required_minimum if col not in df_simulated.columns]

if missing_in_real or missing_in_sim:
    print_progress("MISSING REQUIRED COLUMNS:")
    if missing_in_real:
        print_progress(f" Real data missing: {missing_in_real}")
    if missing_in_sim:
        print_progress(f" Simulation missing: {missing_in_sim}")
    raise ValueError("Cannot proceed without minimum required columns")

# Clean data - remove rows with NaN in required columns
print_progress("Cleaning data...")
df_real_time = df_real_time.dropna(subset=required_minimum)
df_simulated = df_simulated.dropna(subset=required_minimum)

# Simple data splitting
print_progress("Splitting data...")
split_idx = int(0.7 * len(df_real_time))
september_data = df_real_time.iloc[:split_idx].copy()
other_months_data = df_real_time.iloc[split_idx:].copy()

# Filter operational data
operational_threshold = 100
df_test_op = other_months_data[other_months_data['FBHP'] > operational_threshold].copy
    ()
df_sim_op = df_simulated[df_simulated['FBHP'] > operational_threshold].copy()
df_sept_op = september_data[september_data['FBHP'] > operational_threshold].copy()

print_progress(f"Operational data - Test: {len(df_test_op)}, Sim: {len(df_sim_op)},
    Sept: {len(df_sept_op)}")
```

B.4 CCS-Specific Feature Engineering

```
def safe_create_features(df):
    """
    SAFE feature engineering - creates CCS-specific features for supercritical CO2
        systems
    Implements Section 4.6.5 - Advanced Feature Engineering for CCS Applications
    """
    print_progress(f" Starting with {df.shape[1]} columns")

    # Make a copy and work only with confirmed columns
    df_result = df.copy()
    original_columns = list(df.columns)
    created_features = []

    # Basic mathematical operations on confirmed columns
    numeric_cols = []
    for col in ['FWHP', 'FWHT', 'QWH']:
        if col in original_columns:
            numeric_cols.append(col)
```

```
print_progress(f" Available numeric columns: {numeric_cols}")

# Simple squares and logs for available columns
for col in numeric_cols:
    try:
        df_result[f'{col}_square'] = df[col] ** 2
        df_result[f'{col}_log'] = np.log(np.abs(df[col]) + 1)
        created_features.extend([f'{col}_square', f'{col}_log'])
    except:
        pass

# Simple interactions between available columns
for i, col1 in enumerate(numeric_cols):
    for j, col2 in enumerate(numeric_cols):
        if j > i:
            try:
                df_result[f'{col1}_{col2}_mult'] = df[col1] * df[col2]
                df_result[f'{col1}_{col2}_ratio'] = df[col1] / (df[col2] + 1e-6)
                created_features.extend([f'{col1}_{col2}_mult', f'{col1}_{col2}
                    _ratio'])
            except:
                pass

# Petroleum engineering features for CO2 systems

# Density features - Critical for supercritical CO2 pressure calculations
if 'TH_ROL' in original_columns and 'BH_ROL' in original_columns:
    try:
        df_result['liquid_density_diff'] = df['BH_ROL'] - df['TH_ROL']
        df_result['liquid_density_ratio'] = df['BH_ROL'] / (df['TH_ROL'] + 1e-6)
        created_features.extend(['liquid_density_diff', 'liquid_density_ratio'])
        print_progress(" Created liquid density features")
    except:
        print_progress(" Failed to create liquid density features")

# Mixture density - Critical for hydrostatic pressure calculations
if all(col in original_columns for col in ['TH_ROL', 'TH_ROG', 'HOL']):
    try:
        df_result['mixture_density'] = df['HOL'] * df['TH_ROL'] + (1 - df['HOL']) *
             df['TH_ROG']
        created_features.append('mixture_density')
        print_progress(" Created mixture density feature")
    except:
        print_progress(" Failed to create mixture density feature")

# Flow features - Represent multiphase flow dynamics
if all(col in original_columns for col in ['QWH', 'HOL']):
    try:
        df_result['liquid_velocity'] = df['QWH'] * df['HOL']
        df_result['gas_velocity'] = df['QWH'] * (1 - df['HOL'])
        created_features.extend(['liquid_velocity', 'gas_velocity'])
        print_progress(" Created velocity features")
    except:
        print_progress(" Failed to create velocity features")

# Clean up any NaN or infinite values
df_result = df_result.replace([np.inf, -np.inf], np.nan)
```

```
    df_result = df_result.fillna(0)

    print_progress(f" Created {len(created_features)} new features")
    print_progress(f" Final dataset: {df_result.shape[1]} columns")

    return df_result
```

B.5 Hybrid Training Strategy - Information Fusion Implementation

```
# Implementation of Section 4.7 - Hybrid Training Strategy
# Demonstrates information fusion as defined by Mustafee et al. (2023)

print_progress("Implementing hybrid training strategies...")

# Training strategies implementing information fusion
training_data = {
    'Simulation': (X_sim_final, y_sim), # Historical distributions only
}

# Add combined training - Core implementation of information fusion
# Systematic combination of historical simulation data with real-time operational
    measurements
X_combined = pd.concat([X_sim_final, X_sept_final], ignore_index=True)
y_combined = pd.concat([y_sim, y_sept], ignore_index=True)
training_data['Combined'] = (X_combined, y_combined) # Information fusion
    implementation

print_progress(f"Training strategies: {list(training_data.keys())}")
print_progress(f" Simulation data: {X_sim_final.shape[0]} samples (historical
    distributions)")
print_progress(f" Operational data: {X_sept_final.shape[0]} samples (real-time feeds)
    ")
print_progress(f" Combined data: {X_combined.shape[0]} samples (information fusion)")

# This hybrid integration directly implements the information fusion principle
# identified by Mustafee et al. (2023) as essential for Real-time Simulation systems.
```

B.6 Optimized Model Training Framework

```
def create_models_with_gridsearch():
    """
    Create models with practical GridSearchCV for optimal hyperparameters
    Implements Section 4.6.2 - Hyperparameter Optimization Strategy
    """
    models = {}

    print_progress("Setting up PRACTICAL GridSearchCV for hyperparameter optimization
        ...")

    # SVR with reduced parameter grid - RBF kernel optimal for CO2 systems
    svr_params = {
        'C': [10, 100, 1000], # Reduced from 4 to 3 values
        'gamma': ['scale', 0.01, 0.1], # Reduced from 6 to 3 values
```

```
    'epsilon': [0.01, 0.1], # Reduced from 4 to 2 values
    'kernel': ['rbf'] # Only RBF kernel (most effective for nonlinear CO2
        relationships)
}
# Total: 3   3    2    1 = 18 combinations (vs 192 before)

models['SVR'] = Pipeline([
    ('scaler', StandardScaler()),
    ('model', GridSearchCV(
        SVR(),
        svr_params,
        cv=3, # Reduced from 5 to 3 folds for efficiency
        scoring='neg_mean_absolute_error',
        n_jobs=-1,
        verbose=0
    ))
])

# Random Forest with reduced parameter grid
rf_params = {
    'n_estimators': [100, 200, 300], # Reduced from 4 to 3 values
    'max_depth': [None, 15, 30], # Reduced from 4 to 3 values
    'min_samples_split': [2, 10], # Reduced from 3 to 2 values
    'max_features': ['sqrt', None] # Reduced from 3 to 2 values
}
# Total: 3   3    2    2 = 36 combinations (vs 432 before)

models['RandomForest'] = Pipeline([
    ('scaler', StandardScaler()),
    ('model', GridSearchCV(
        RandomForestRegressor(random_state=42),
        rf_params,
        cv=3,
        scoring='neg_mean_absolute_error',
        n_jobs=-1,
        verbose=0
    ))
])

# Gradient Boosting with reduced parameter grid
gb_params = {
    'n_estimators': [100, 200], # Reduced from 3 to 2 values
    'learning_rate': [0.1, 0.2], # Reduced from 4 to 2 values
    'max_depth': [5, 7], # Reduced from 4 to 2 values
    'min_samples_split': [2, 10], # Reduced from 3 to 2 values
    'subsample': [0.9, 1.0] # Reduced from 3 to 2 values
}
# Total: 2   2    2    2    2 = 32 combinations (vs 1,296 before)

models['GradientBoosting'] = Pipeline([
    ('scaler', StandardScaler()),
    ('model', GridSearchCV(
        GradientBoostingRegressor(random_state=42),
        gb_params,
        cv=3,
        scoring='neg_mean_absolute_error',
        n_jobs=-1,
        verbose=0
```

```
        ))
    ])

    # XGBoost with reduced parameter grid (if available)
    if XGBOOST_AVAILABLE:
        xgb_params = {
            'n_estimators': [100, 200], # Reduced from 3 to 2 values
            'learning_rate': [0.1, 0.2], # Reduced from 4 to 2 values
            'max_depth': [5, 7], # Reduced from 4 to 2 values
            'min_child_weight': [1, 3], # Reduced from 3 to 2 values
            'reg_alpha': [0, 0.1], # Reduced from 4 to 2 values
            'reg_lambda': [0, 0.1] # Reduced from 4 to 2 values
        }
        # Total: 2^6 = 64 combinations (vs 20,736 before!)

        models['XGBoost'] = Pipeline([
            ('scaler', StandardScaler()),
            ('model', GridSearchCV(
                xgb.XGBRegressor(random_state=42, verbosity=0),
                xgb_params,
                cv=3,
                scoring='neg_mean_absolute_error',
                n_jobs=-1,
                verbose=0
            ))
        ])

    total_combinations = 18 + 36 + 32 + (64 if XGBOOST_AVAILABLE else 0)
    print_progress(f"Created {len(models)} models with practical GridSearchCV")
    print_progress(f"Reduced to ~{total_combinations} total combinations (vs ~23,000
        before)")
    print_progress(f"Should complete in 5-15 minutes instead of hours")

    return models
```

B.7 Ensemble Model Development

```
def create_ensemble_models_optimized(base_models_dict):
    """
    Create Voting and Stacking ensemble models using existing individual models
    Implements Section 4.6.6 - Ensemble Method Development
    """
    ensemble_models = {}

    print_progress("Creating Voting and Stacking ensemble models...")

    # Extract the best estimators from GridSearchCV
    best_estimators = {}
    for name, pipeline in base_models_dict.items():
        try:
            # Get the best estimator from GridSearchCV
            best_model = pipeline.named_steps['model'].best_estimator_
            best_estimators[name.lower()] = best_model
        except:
            # Fallback if GridSearchCV failed
            continue
```

```
    if len(best_estimators) >= 3:

        # Voting Ensemble - using top models
        voting_estimators = []
        for name, model in list(best_estimators.items())[:4]:
            voting_estimators.append((name, model))

        ensemble_models['VotingEnsemble'] = Pipeline([
            ('scaler', StandardScaler()),
            ('model', VotingRegressor(estimators=voting_estimators))
        ])

        # Combined Ensemble - using ALL available models
        all_estimators = []
        for name, model in best_estimators.items():
            all_estimators.append((f'combined_{name}', model))

        ensemble_models['CombinedEnsemble'] = Pipeline([
            ('scaler', StandardScaler()),
            ('model', VotingRegressor(estimators=all_estimators))
        ])

        # Stacking ensemble with Ridge meta-learner
        stacking_estimators_all = []
        for name, model in list(best_estimators.items()):
            stacking_estimators_all.append((name, model))

        ensemble_models['StackingRidge'] = Pipeline([
            ('scaler', StandardScaler()),
            ('model', StackingRegressor(
                estimators=stacking_estimators_all,
                final_estimator=Ridge(alpha=1.0),
                cv=3,
                n_jobs=-1
            ))
        ])

    print_progress(f"Created {len(ensemble_models)} ensemble models:")
    for name in ensemble_models.keys():
        print_progress(f" {name}")

    return ensemble_models
```

B.8 Comprehensive Performance Analysis

```
def run_comprehensive_analysis(results, y_test, normalization_method='minmax'):
    """
    Complete analysis framework generating all thesis visualizations
    Implements Section 5.4 - Model Performance Visualization and Temporal Analysis
    """

    print("STARTING COMPREHENSIVE FBHP ANALYSIS - ALL MODELS")
    print(f"Total models to analyze: {len(results)}")
    print(f"Normalization method: {normalization_method.upper()}")
    print(f"Test data points: {len(y_test):,}")

    # Sort results by R score for reporting
```

```
    results_sorted = sorted(results.items(), key=lambda x: x[1]['r2'], reverse=True)

    print(f"\nTOP 10 MODELS (Original Values):")
    print(f"{'Rank':<4} {'Model':<35} {'R':<8} {'MAE':<8} {'RMSE':<8}")
    print("-" * 70)

    for i, (name, res) in enumerate(results_sorted[:10]):
        print(f"{i+1:2d}. {name:<35} {res['r2']:7.4f} {res['mae']:7.2f} {res['rmse
            ']:7.2f}")

    best_name, best_result = results_sorted[0]
    print(f"\nBEST MODEL: {best_name}")
    print(f" R  = {best_result['r2']:.4f}")
    print(f"MAE = {best_result['mae']:.2f} psi")
    print(f"RMSE = {best_result['rmse']:.2f} psi")

    # Create comprehensive plots for all models
    plots_folder = create_comprehensive_plots_for_all_models(results, y_test,
        normalization_method)

    # Create results table
    df_results = create_model_results_table(results)

    # Performance interpretation
    mae_percentage = (best_result['mae'] / np.mean(y_test)) * 100

    print(f"\nPERFORMANCE ANALYSIS:")
    print(f"  Average FBHP: {np.mean(y_test):.1f} psi")
    print(f"  Best MAE: {best_result['mae']:.1f} psi ({mae_percentage:.1f}% of mean)")

    return plots_folder, df_results, results_sorted
```

B.9 Execution Example

```
# Main execution workflow demonstrating complete methodology

# 1. Process enhanced features
df_sim_enhanced = safe_create_features(df_sim_op)
df_sept_enhanced = safe_create_features(df_sept_op)
df_test_enhanced = safe_create_features(df_test_op)

# 2. Feature selection
n_features = min(15, len(common_features))
selector = SelectKBest(score_func=mutual_info_regression, k=n_features)
selector.fit(X_test, y_test)
selected_features = [common_features[i] for i in selector.get_support(indices=True)]

# 3. Train models with information fusion
individual_models = create_models_with_gridsearch()
results = {}

for strategy_name, (X_train, y_train) in training_data.items():
    for model_name, model in individual_models.items():
        # Train model with GridSearchCV
        model.fit(X_train, y_train)

        # Predict on test set
```

```
        y_pred = model.predict(X_test_final)

        # Calculate metrics
        r2 = r2_score(y_test, y_pred)
        mae = mean_absolute_error(y_test, y_pred)
        rmse = np.sqrt(mean_squared_error(y_test, y_pred))

        result_name = f"{strategy_name}_{model_name}"
        results[result_name] = {
            'predictions': y_pred,
            'r2': r2,
            'mae': mae,
            'rmse': rmse
        }

# 4. Run comprehensive analysis
plots_folder, df_results, results_sorted = run_comprehensive_analysis(
    results=results,
    y_test=y_test,
    normalization_method='minmax'
)

print("All plots and analysis completed!")
```

B.10 Implementation Notes

B.10.1 Key Design Principles

- **Practical Efficiency:** Hyperparameter grids reduced by 99% while maintaining optimization effectiveness
- **Robust Feature Engineering:** Graceful handling of missing petroleum engineering parameters
- **Information Fusion:** Systematic integration of historical and real-time data sources
- **Comprehensive Evaluation:** Automated generation of all thesis visualizations and analyses

B.10.2 Reproducibility Features

- Consistent random seed initialization throughout implementation
- Automated progress tracking and error handling
- Complete results documentation in multiple formats
- Cross-platform compatibility with standard Python scientific stack

This implementation demonstrates the practical feasibility of the theoretical methodology described in Chapter 4 and generates all results and visualizations presented in Chapter 5.

Bibliography

[1] IPCC. *Special Report on Carbon Dioxide Capture and Storage*. Cambridge University Press, 2005.

[2] V. Pipich and D. Schwahn. Polymorphic phase transition in liquid and supercritical carbon dioxide. *Scientific Reports*, 10:11861, 2020.

[3] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

[4] United Nations Framework Convention on Climate Change. Paris agreement, 2015.

[5] International Energy Agency. Global energy review 2025, 2025.

[6] International Energy Agency. Co_2 emissions – global energy review 2025, 2025.

[7] IPCC. *Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. Cambridge University Press, 2018.

[8] International Risk Governance Council. Risk governance of carbon capture and storage, 2007.

[9] European Parliament and Council. Directive 2009/31/ec of the european parliament and of the council on the geological storage of carbon dioxide. *Official Journal of the European Union*, L 140/114, 2009.

[10] U.S. Environmental Protection Agency. Renewable fuel standard program (rfs2) regulatory impact analysis. Technical report, U.S. Environmental Protection Agency, Washington, D.C., 2010.

[11] A. R. Hagedorn and K. E. Brown. Experimental study of pressure gradients occurring during continuous two-phase flow in small-diameter vertical conduits. *Journal of Petroleum Technology*, 17(4):475–484, 1965.

[12] H. D. Beggs and J. P. Brill. A study of two-phase flow in inclined pipes. *Journal of Petroleum Technology*, 25(5):607–617, 1973.

[13] K. H. Bendiksen, D. Malnes, R. Moe, and S. Nuland. The dynamic two-fluid model olga: Theory and application. *SPE Production Engineering*, 6(2):171–180, 1991.

[14] D. Burnett. The distribution of molecular velocities and the mean motion in a non-uniform gas. *Proceedings of the London Mathematical Society*, 40(1):382–435, 1935.

[15] S. Huang, L. Tian, J. Zhang, X. Chai, H. Wang, and H. Zhang. Support vector regression based on the particle swarm optimization algorithm for tight oil recovery prediction. *ACS Omega*, 6(47):32142–32150, 2021.

[16] D. A. Otchere, T. O. A. Ganat, R. Gholami, and S. Ridha. Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ann and svm models. *Journal of Petroleum Science and Engineering*, 200:108182, 2021.

[17] Xuejia Du, Muhammad Noman Khan, and Ganesh C. Thakur. Machine learning in carbon capture, utilization, storage, and transportation: A review of applications in greenhouse gas emissions reduction. *Processes*, 13(4):1160, 2025.

[18] Z. X. Leong and T. Zhu. Machine learning-based monitoring of co_2 geological storage with sparse seismic data. *International Journal of Greenhouse Gas Control*, 135:104151, 2024.

[19] A. Mahmoudzadeh, B. Amiri-Ramsheh, S. Atashrouz, A. Hemmati-Sarapardeh, M. Schaffie, and M. Ranjbar. Modeling co_2 solubility in water using gradient boosting and light gradient boosting machine. *Scientific Reports*, 14:13511, 2024.

[20] Navonil Mustafee, Alison Harper, and Joe Viana. Hybrid models with real-time data: characterising real-time simulation and digital twins. In C. Currie and L. Rhodes-Leader, editors, *Proceedings of the Operational Research Society Simulation Workshop 2023 (SW23)*, pages 261–271, 2023.

[21] R. Arts, O. Eiken, A. Chadwick, P. Zweigel, L. van der Meer, and B. Zinszner. Monitoring of co_2 injected at sleipner using time-lapse seismic data. *Energy*, 29(9-10):1383–1392, 2004.

[22] U.S. Department of Energy. Science-informed machine learning to accelerate real-time (smart) decisions for subsurface applications. Carbon Storage and Upstream Oil and Gas Program, 2023.

[23] R. Span and W. Wagner. A new equation of state for carbon dioxide covering the fluid region from the triple-point temperature to 1100 k at pressures up to 800 mpa. *Journal of Physical and Chemical Reference Data*, 25(6):1509–1596, 1996.

[24] Beatriz Flamia Azevedo, Ana Maria A. C. Rocha, and Ana I. Pereira. Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning*, 113:4055–4097, 2024.

[25] R. Ferrario, A. L. Lamberti, C. Piccione, A. L. Russo, C. Topini, C. Molinari, and F. Rebaudi. Ravenna ccs - phase 1: Early insights from the first 8 months of co_2 capture, transport and storage operations. In *17th OMC Med Energy Conference and Exhibition*, Ravenna, Italy, 2025. Paper OMC-2025-675.