CCUS: 4427254

A Critical Review of Machine Learning Applications in Monitoring and Predicting CO_2 Storage Behavior

Alireza Bigdeli*[1], Cenk Temizel[2], 1. State University of Campinas, 2. Saudi Aramco/Terra Altai.

This paper was prepared for presentation at the Carbon Capture, Utilization, and Storage conference held in The Woodlands, TX, 30 March – 01 April.

Abstract

Machine learning (ML) is becoming a vital tool for monitoring, prediction and optimization of geological CO_2 storage, although there has been considerable progress, several open questions remain regarding how robustly models trained in the lab generalize to the field, transfer to new sites, and respect known physics, all of which are important for practical deployment. Here, we review ML applications in three major tasks for geological CO_2 storage: (1) tracking CO_2 plume migration, (2) leakage detection and (3) reservoir pressure forecasting, and we cover various ML models, including CNNs, recurrent/transformer models, neural operators, and surrogate models. We show the potential of ML models in achieving high accuracy, reducing the turnaround time and improving the scalability compared to conventional numerical simulators, especially when high-quality and large datasets are available, which can result in rapid CO_2 plume imaging and earlier anomaly detection. However, there are a number of challenges, including the limited generalizability of models to different reservoirs, sensitivity of models to sparse/noisy measurements, and the lack of robust uncertainty quantification. Hybrid models that combine physics with data-driven learning, and transfer learning are also promising future directions, and in conclusion, ML-based monitoring has proven to be a powerful tool for complementing the physics-based modeling of geological CO_2 storage. Next steps include the development of standardized benchmark datasets, the better incorporation of physical constraints into ML models, and digital twins for real-time reservoir management and risk mitigation.

Introduction

There is now conclusive evidence that the increasing levels of carbon dioxide in the Earth's atmosphere are directly linked to rising temperatures and climate instability globally (Friedlingstein et al., 2022). Without immediate action, CO_2 concentrations in the Earth's atmosphere will reach over 450 parts per million by 2050. The geological carbon dioxide (CO_2) sequestration involves the long term or permanent storage of CO_2 in permeable underground rock units such as depleted oil and gas fields, saline aquifers and coal deposits (Cao et al., 2020). Conventional methods for monitoring rely upon the use of numerical simulations and the interpretation of geophysical data. However, these methods have several limitations when applied to large-scale reservoirs. The time required to simulate large systems can be excessive; the amount of data generated by monitoring instruments is often too large to analyze manually and it may take months before the results of monitoring are available (Tariq et al., 2021). Combining machine learning techniques with those used in geophysical monitoring, well logs and reservoir simulation provides the opportunity to develop real-time decision-support tools and reduce the risks associated with storing CO_2 in large underground reservoirs.

Machine Learning Fundamentals for CO_2 Storage Applications

Supervised Learning Methods in Reservoir Characterization: Reservoir characterization has been heavily reliant on supervised machine learning techniques for CO_2 storage. There are two supervised learning techniques widely applied to CO_2 storage: Random Forest (RF) and Extreme Gradient Boosting (XGBoost) techniques. The two supervised learning techniques can predict CO_2 saturation from pre-stack seismic attributes. For instance, a pioneering study at the Cranfield site demonstrated that RF and XGBoost regression models, which were trained on synthetic data, could determine CO_2 saturation levels with relatively low mean absolute error (MAE = 0.029 for RF model and MAE = 0.028 for XGBoost model) (Owusu & Zhang, 2025). This was because both RF and XGBoost are ensemble methods capable of identifying non-linear relationships between seismic attribute and fluid property in the subsurface as well as providing interpretability of feature importance ranking.

Physics-Based Proxy Modeling: SVM and their kernel variants have proven to be effective for both lithofacies classification and property prediction according to research findings. One example study was conducted in the Baikouquan Formation of the Junggar Basin to classify mudstone, sandstone and conglomerate facies based on well log data. It was determined that XGBoost had an overall classification accuracy of 0.882 and an area under the receiver operating characteristic (ROC) curve (AUC) of 0.947 (He et al.,2023). Additionally, multi-kernel learning approaches in combination with SVM have the capability to process high-dimensional well log data in an efficient manner and reduce computational time when compared to traditional methods. Physics-based machine learning proxy models represent a significant breakthrough which enables domain experts to link their knowledge with data-based methods. The Multilayer Perceptron (MLP) and Random Forest and Support Vector Regression models achieved successful CO_2 trapping mechanism prediction through their high predictive accuracy which reached R^2 values above 0.9989 for residual and mineralized and dissolution trapping mechanisms (Khanal & Shahriar, 2022). Proxy models utilize simulator output over a period of 275 years in order to determine how long CO_2 is trapped using fast computations that are used as an alternative to the computationally expensive forward simulation processes.

Seismic Monitoring and Time-Lapse Inversion for CO_2 Storage

Time-Lapse Seismic as Foundation for Monitoring: Time-Lapse (4-Dimensional) Seismic is the primary geophysics method used to track the evolution of CO_2 Plumes in storage reservoirs. A series of repeated seismic surveys are conducted and 3-D Data Volumes are collected over different periods of time; changes in elastic properties due to variations in CO_2 Saturation, changes in Pressure and Temperature Effects, are captured within each 3-D Volume. Differences between baseline and monitor survey data reveal subtle changes that cannot be detected from analyzing individual 3-D volumes. Machine learning has greatly affected the interpretation of complex multi-dimensional datasets (Ogu et al., 2023).

Deep Learning for Seismic Inversion: Deep Learning Methods called Cycle-Consistent Generative Adversarial Networks (CycleGAN) have been successfully applied to convert time-lapse seismic data to CO_2 Saturation Distributions (Zhong et al., 2020). Instead of performing computationally intensive forward problems multiple times, the CycleGAN learns the bidirectional mapping from seismic attribute changes to the associated CO_2 Saturation Changes from training data. The Physics-based Deep Learning methodology adheres to all the physics-based constraints and accelerates the inversion process by converting an hours-long computation into almost instantaneous predictions.

Real-Time Deep Learning Inversion for Operational Monitoring: The real-time application of deep learning inversion methodologies using U-Net Architectures allow for real-time estimation of CO_2 Saturation Distributions and their associated uncertainties immediately after acquiring new seismic data. Using proof-of-concept studies based on the Kimberlina Synthetic Model demonstrates that neural networks may accurately predict 2-D CO_2 Saturation Distributions, even when subjected to Gaussian Noise, and that Uncertainty Quantification can be obtained using Monte Carlo Dropout and Bootstrap Aggregating methods (Um et al., 2022).

Neural Network Architectures for CO_2 Saturation and Property Estimation

Convolutional Neural Networks for Spatial Pattern Recognition: The predictive modeling of CO_2 saturation properties can use deep learning methods to analyze large multi-dimensional geophysical data sets with the help of convolutional neural networks (CNN's). These tools are capable of directly identifying geospatial patterns in this type of data set (Wang & Alkhalifah, 2021), as has been demonstrated through the analysis of passive seismic array measurements to determine location of microseismic events. This indicates that CNN's have the capability to directly map the raw waveform data from seismic surveys to the source locations, eliminating the need to first pick the events (Wang & Alkhalifah, 2021).

U-Net Architecture and Skip Connections: The U-Net architecture was originally designed for performing medical image segmentation tasks, but it is now widely used in the area of geophysical data interpretation. This architecture includes an encoder-decoder structure with skip connections, which allows spatial information to be preserved while progressive feature extraction is performed.

Physics-Informed Neural Networks (PINNs): Physics-informed neural networks include partial differential equations describing multiphase flow, reactive transport, and geomechanical processes into the loss function when the network is being trained (Li et al., 2023). Adaptive weights are included in self-adaptive PINNs that allow the network to learn where the most difficult areas of the solution space lie, i.e. shock fronts in CO_2 saturation fronts.

Data Assimilation Methods for CO_2 Storage

Ensemble-Based Methods: Ensemble-based approaches to assimilating time-lapse data into reservoir models, primarily through the use of the Ensemble Kalman Filter (EnKF) and Ensemble Smoother (ES), and ES using Multiple Data Assimilation (ES-MDA) have shown great success in the integration of time-lapse data into reservoir models (Tavakoli et al., 2013). The ensemble-based approaches described above consider the reservoir characterization problem as a sequence of inference problems. An ensemble of potential realizations, which are then updated as new observations are assimilated, and posterior uncertainty represents the remaining unknowns within the reservoir, represents prior uncertainty. **Geostatistical Inversion for Spatial Distribution Estimation**: The combination of ES optimization with spatial constraints has been used successfully in geostatistical inversion to estimate the spatial distribution of CO_2 saturations from time-lapse geophysical data (Grana et al., 2021). The method preserves geological reality by modifying prior geostatistically generated realizations, rather than smoothing solutions; and it uses hydrologic constraints such as relationships between phase behavior and multiphase flow physics.

Deep Learning-Accelerated Data Assimilation: Deep learning-based data assimilation workflows, which employ deep learning methods, reduce the costs required for history matching operations. Neural networks that learn to map monitoring data directly to model parameters or CO_2 saturations have been utilized to rapidly calibrate reservoir models at the Illinois Basin-Decatur Project (Chan et al.,2024). For example, a deep learning workflow was developed to train a neural network that could take downhole pressure and temperature measurements and produce a predictive model for diffusive time-of-flight maps (i.e. the flow field), reducing the computational time required to perform history matching from hours of computationally expensive multiphase flow simulations to milliseconds of neural network inference.

Physics-Guided and Physics-Informed Machine Learning Approaches

Integration of Domain Knowledge: The most effective uses of machine learning in CO_2 Storage are those which utilize a combination of the scientific domain knowledge and data-driven learning. For example, Physics-Guided Modular Networks represent a combination of scientific domain knowledge and data-driven learning. While the network architecture is based upon scientific principles (Li et al., 2023), it is still based upon data. This allows the network to simultaneously optimize for four important aspects of model performance; namely, fitting the training data, generalizing to new data, removing noise from the input data, and satisfying the physics governing the problem being modeled.

Rock Physics Relationships in ML Models: Rock physics provides some of the necessary relations to link subsurface fluid properties (e.g., saturation, pressure, temperature) to seismic attribute values (i.e., velocity, impedance, attenuation). When the established rock physics relations (i.e., Gassmann's equation or Raymer's porosity-velocity relationship) are embedded into neural network architectures, the resulting models will be able to generalize better to new geological conditions.

Multi-Physics Integration: CO_2 storage behavior is influenced by several coupled process domains, including: multiphase flow, geochemistry, geomechanics and thermodynamics. As a result, when developing machine learning models that include multi-domain physics, they should be constrained using relevant information from each domain. As a result, the models can capture the dominant coupling effects among the different physics domains. (Khanal et al., 2024)

Advanced ML Methodologies and Emerging Approaches

Generative Adversarial Networks: In addition to generating high-quality synthetic seismic data, synthetic well-logs and physically-consistent synthetic pressure/temperature fields, Generative Adversarial Networks (GANs), can also be used to generate large numbers of synthetic reservoir responses based on actual field data. These can then be added to the training data in a machine learning model to increase its size and scope (Ogu et al., 2024).

Graph Neural Networks for Complex Spatial Relationships: Graph Neural Networks are able to represent complex spatial relationships, such as those found in a subsurface reservoir as a graph where each node in the graph represents a grid cell or well location, and each edge in the graph represents the physical connection between the nodes. Because the graph representation inherently captures the spatial connectivity of the reservoir, it is very efficient at propagating information across the reservoir network. There are potential applications of this type of model in identifying faults in a reservoir, classifying facies within a reservoir, and predicting plumes on geological networks.

Conclusions

Machine learning has emerged as a transformative technology for CO_2 storage monitoring and prediction, offering computational advantages, pattern recognition capabilities, and integration of diverse data types that exceed classical approaches. This review has examined advances across multiple ML domains: supervised learning algorithms for property estimation, physics-informed methods embedding domain knowledge, and data assimilation frameworks for uncertainty quantification. Research priorities must address these through systematic validation studies, transfer learning investigations, robust uncertainty methods, and collaborative development of regulatory frameworks. The integration of ML with physics-informed approaches, physics-constrained data assimilation, and hybrid architectures leveraging complementary strengths of classical and data-driven methods represents the most promising pathway forward. Rather than viewing ML as a complete replacement for physics-based simulation, the most robust approach should integrate data-driven inference with domain knowledge constraints, physical laws, and probabilistic uncertainty representation. Long-term success in deploying ML for CO2 storage monitoring depends on sustained research investment, capacity building across disciplines, open collaboration in pre-competitive research, and development of consensus frameworks for validation and governance.

References

Cao, C., Liu, H., Hou, Z., Mehmood, F., Liao, J., & Feng, W. (2020). A review of CO2 storage in view of safety and cost-effectiveness. Energies, 13(3), 600.

Chan, C., Sakai, T., Nagao, M., Datta-Gupta, A., & Yao, C. (2024). CO2 plume imaging with accelerated deep learning-based data assimilation using distributed pressure and temperature measurements at the Illinois Basin-Decatur carbon sequestration project. Proceedings of the CCUS Conferences.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., … & Canadell, J. G. (2022). Global carbon budget 2022. Earth System Science Data, 14(11), 4811-4900.

Grana, D., Liu, M., & Ayani, M. (2021). Prediction of CO saturation spatial distribution using geostatistical inversion of time-lapse geophysical data. IEEE Transactions on Geoscience and Remote Sensing, 59(5), 3690-3703.

He, B., Zhu, X., Cang, Z., Liu, Y., Lei, Y., Chen, Z., … & Zhang, L. (2023). Interpretation and prediction of the CO2 sequestration of steel slag by machine learning. Environmental Science & Technology, 57(36), 13319-13328.

Khanal, A., & Shahriar, M. F. (2022). Physics-based proxy modeling of CO2 sequestration in deep saline aquifers. Energies, 15(12), 4350.

Li, X., Li, B., Liu, F., Li, T., & Nie, X. (2023). Advances in the application of deep learning methods to digital rock technology. Advances in Geo-Energy Research, 8(1), 5-18.

Liu, E., Lu, X., & Chen, D. (2023). A systematic review of carbon capture, utilization and storage: Status, progress and challenges. Energies, 16(6), 2865.

Liu, X., Zhou, L., Chen, X., & Li, J. (2020). Lithofacies identification using support vector machine based on local deep multi-kernel learning. Petroleum Science, 17(6), 1609-1621.

Ogu, E., Egbumokei, P. I., Dienagha, I. N., & Digitemie, W. N. (2023). Seismic technologies in oilfield management: A conceptual approach to improving operational efficiency in dynamic reservoirs. International Journal of Multidisciplinary Research and Growth Evaluation, 5(1), 1181-1196.

Ogu, E., Egbumokei, P. I., Dienagha, I. N., & Digitemie, W. N. (2024). Elastic inversion as a key to understanding reservoir dynamics: A conceptual model for 4D seismic interpretation in oilfields. International Journal of Multidisciplinary Research and Growth Evaluation, 4(1), 724-736.

Owusu, P., & Zhang, R. (2025). Estimating CO2 saturation in subsurface geologic reservoirs using pre-stack seismic attributes and machine learning. Proceedings of the CCUS Conferences.

Tavakoli, R., Yoon, H., Delshad, M., Elsheikh, A. H., Wheeler, M. F., & Arnold, W. (2013). Comparison of ensemble filtering algorithms and null-space Monte Carlo for parameter estimation and uncertainty quantification using CO2 sequestration data. Water Resources Research, 49(11), 7837-7854.

Tariq, Z., Aljawad, M. S., Hasan, A., Murtaza, M., Mohammed, E. A., ElHusseiny, A., … & Abdulraheem, A. (2021). A systematic review of data science and machine learning applications to the oil and gas industry. Journal of Petroleum Exploration and Production Technology, 11, 3899-3928.

Um, E. S., Alumbaugh, D., Lin, Y., & Feng, S. (2022). Real-time deep-learning inversion of seismic full waveform data for CO2 saturation and uncertainty in geological carbon storage monitoring. Geophysical Journal International, 229(2), 1234-1251.

Wang, H., & Alkhalifah, T. (2021). Direct microseismic event location and characterization from passive seismic data using convolutional neural networks. Geophysics, 86(4), KS99-KS116.

Zhong, Z., Sun, A., & Wu, X. (2020). Inversion of time-lapse seismic reservoir monitoring data using CycleGAN: A deep learning-based approach for estimating dynamic reservoir property changes. Journal of Geophysical Research: Solid Earth, 125(3), e2019JB018408.