**Title:** **Utilisation of Artificial Intelligence based Time-Series Prediction to validate Carbon Containment in Injection Well in Illinois Basin**

**Authors:** Munish Kumara and Kanna Swaminathana

**Affiliations:** aSingapore University of Social Sciences, School of Business, Clementi Rd, #463号, Singapore 599494

bCalifornia Resources Corporation, 900 Old River Rd, Bakersfield, CA 93311, United States

Corresponding author. Email: munishkumar001@suss.edu.sg

**Abstract:**

**One-Sentence Summary:** Applying machine learning and predicative analytics via time series injection information and monitoring data on a carbon capture well to predict well injection rate deltas.

**Keywords (minimum 6):** low-carbon, time-series, neural network, LSTM, carbon capture, injection pressure, monitoring, plume migration

# Introduction

There is no doubt that human activities are one of the main reasons for the increase in amounts of greenhouse gases (GHG) and global warming. Primarily, CO2 levels have increased due to the rapid pace of industrialisation and population growth, which has only picked up pace since the 1960s [1].

The burning of fossil fuels (coal, oil, and natural gas) has resulted in carbon dioxide (CO2), methane (CH4) and nitrous oxide (N2O) generation. A report generated by the United States Environmental Protection Agency (EPA) evaluated the amount of CO2 generated in 2020 to be ~3.11 million metric tons [2]. Such significant quantities, if not better managed, will lead to irreversible environmental consequences.

Carbon Capture Utilisation and Storage (CCUS) is one promising method to deal with the produced CO2 from anthropogenic sources. The idea is to capture CO2 from an emission point source and subsequent sequester it via injection into a suitable geological formation, with the explicit aim to store the CO2 safely, in a state of permanance. CCUS is now being viewed as a key technology to assist in reaching global anthropogenic climate change goals. Typically, CO2 storage is carried out in one of three ways - via injection into virgin saline aquifers; into depleted oil and gas fields; or if used for Enhanced Oil Recovery (EOR) processes. These methods all have different project drivers, risks, and commercial implications. However, what these 3 methods have in common is the requirement that (a) there be a good understanding of the subsurface geological properties and (b) there be some ability to monitor and even predict CO2 behaviour at the well scale, be in during the injection phase or during the shut-in phase.

The uncertainties present in (b) especially necessitate pilot development and validation of the technology. The Illinois Basin - Decatur Project is one such study which aimed to demonstrate the capacity, injectivity, and containment of carbon storage in the Mount Simon Sandstone, which is the main carbon storage resource in the Illinois Basin and the Midwest Region. Injection began in 2009 into the Mount Simon Sandstone at a depth of approximately 7,000 feet (2,100 meters), and continued for a 3-year period, during which over 1 million tonnes of CO2 was injected

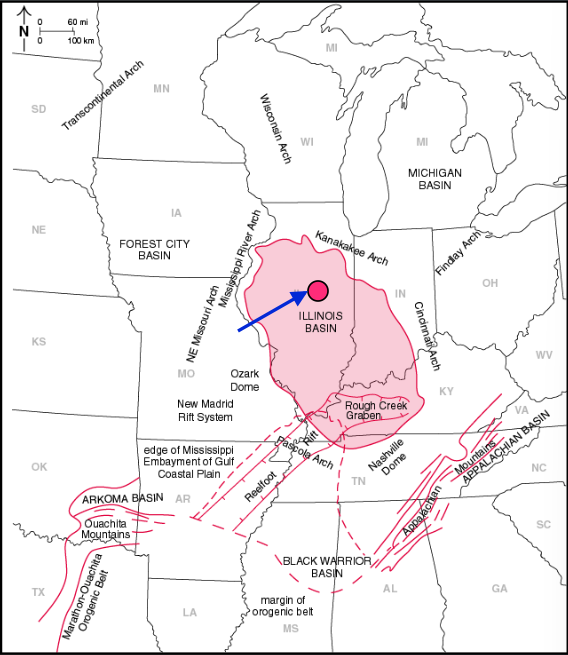


Figure 1: Location map of the Illinois Basin – Decatur Project (IBDP). Image taken from

This paper aims to use time series injection information and monitoring data on a carbon capture well to predict carbon capture well injection rates deltas (D) which is the difference in the injection rate (IR) at time t and time (t-1) i.e.

|  |  |  |
| --- | --- | --- |
|  |  | Equation 1 |

Correlating the change in injection rate to the behaviour of other parameters in the well can be used to provide a checkpoint against carbon migration from the well or other losses during the process. Utilisation of a machine learning (ML) method to predict injection rate deltas based on monitoring well data can be used to validate carbon containment throughout the injection of the well as well.

# Literature Review

Various authors have tried numerous methods to forecast future trends based on past data. Work by De Gooijer and Hyndman [3], for instance, reviewed a series of time-series forecast models over a 25 year period, from 1985 to 2005. Their review highlighted various models being developed and applied in a myriad of scenarios related to finance, statistics and manufacturing, and include methods such as (a) exponential smoothing [4, 5], (b) Autoregressive Integrated Moving Average (ARIMA) [6], (c) seasonal models [7], (d) state space and structural models and the Kalman filter [8], (e) nonlinear models [9], (f) long-range dependence models, including the family of Autoregressive Fractionally Integrated Moving Average (ARFIMA) models [10], (g) Autoregressive Conditional Heteroskedasticity/Generalized Autoregressive Conditional Heteroskedasticity (ARCH/GARCH) models [11], and (h) count data forecasting [12].

In reservoir engineering, the prediction of hydrocarbon or water rates from geological formations is a time-series forecasting problem, with empirical solutions developed by Arps [13], referred to as decline curve analysis (DCA) technique. The method is based on a curve-fit principle, where one would attempt to fit either exponential, hyperbolic or harmonic curve to historical flow production rate as a function of time. Equation 2 shows the general form of the equation

|  |  |  |
| --- | --- | --- |
|  |  | Equation 2 |

where qi is the initial rate (bbls/day), Di is the initial decline rate (units) and b is the degree of curvature of the line. An exponential curve fit would have b = 0, a hyperbolic curve would have 0 < b < 1 and a harmonic curve would have b = 1. The fitted curve is then used to predict future production rates and cumulative production. This method was originally designed to work with high porosity-permeability reservoirs, and tends to overestimate hydrocarbon recovery from unconventional (low permeability) reservoirs. Thus, various authors have tried to expand on this work [14, 15, 16, 17], but for all intends and purposes, they are mostly just variations of the initial DCA method developed by Arps. For an effective DCA forecast, domain knowledge is key, but inherently the process is one of trial and error, and thus it is not uncommon for DCA results to have low-best-high estimates.

With the advent of big data, fast computing and cheap memory, applying a machine learning (ML) and artificial intelligence (AI) solution for time-series forecasting seems a natural evolution. ML solutions were first introduced to the petroleum industry in the early 2000s. Application of ML and AI include addressing reservoir history matching, forecasting of oil, gas and water production (flow rate prediction), pattern recognition in well logs and well tests analysis, production enhancement and prediction of failures, among others [18, 19, 20].

However, in the domain of CCUS and injector production performance, there is little to no available data related to time series and forecasting. Rather, literature appears to be mostly concentrated in predicting carbon emissions [21], leakage [21], CO2 absorption and adsorption [22], property prediction and process simulation [21], simulation of transportation, and geological behaviours as it relates to uncertainty analysis, sequestration, utilisation and EOR processes [25, 26, 27]. Work by Iskander et al perhaps most closely approaches this, where Long Short-Term Memory (LSTM) networks were utilised to forecast oil, water and CO2 production at future infill well locations, for both single phase and 3-phase fluid models. Data was in the form of a synthetic PUNQ-S3 reservoir model, combined with real-world observations from 8 production wells, which recorded daily production volumes over a decade from 2004-2014.

While the LSTM network additionally considers well-to-well interactions, it is built off the synthetic and thus suffers from limited data. There

We aim to utilise ML and AI methods, and in particular LSTM, but focusing on the prediction of CCUS injection well performance, using the open-source information from the IBDP. We will demonstrate how our developed LSTM model shows a correlation between the change in injection rate to the behaviour of other dynamic parameters.

While the primary purpose of the model is as a checkpoint against carbon migration, either at the well location or from other losses during the injection process, we view the model as another means for engineers to perform scenario based de-risking of exploration plays, via modelling variation in well and storage parameters to validate CO2 containment. The model will also aid in the understanding of the injection process and potentially can be used to “right size” well operations and optimise costs.

## Challenges with XXXXX

# Scope and Methods

## Datasets

During this three-year period, a substantial amount of data was collected.

## Methodology

1. *Data Collection and Preparation*

**Table 1:** XXXXX

1. *Cleaning Data*
2. *Machine Leaning Model Selection -* Long Short-Term Memory (LSTM)
3. *Model Training and Hyperparameter Optimization*

We initially utilized 20 different algorithms but finally opted for 3 based on the ranked results of the mean absolute error (MAE), the mean squared error (MSE), the root mean squared error (RMSE) and the coefficient of determination (R2):

where yi is the predicted value, xi is the true value, is the mean of xi and n is the total number of data points. Note that a good result is one that has a low MAE, MSE and RMSE, but a high R2 (since this is essentially a measure of “goodness of fit”).

The models were then tuned by varying “hyperparameters”. A hyperparameter is a characteristic of a model that is external to the model and whose value cannot be estimated from data. The hyperparameters are optimized via a search algorithm with the goal of minimize the overall error metric.

To account for idiosyncrasies in the data (noise, patterns, outliers, etc.), k-fold cross-validation was run to validate the stability of the model.

1. *Testing, Deployment, and Evaluation of model against our blind datasets*

# Results and Discussion

|  |  |
| --- | --- |
|  |  |

Figure 3: Correlation Matrix – Pairwise Correlation

We observed that for the TORIS database, “lithology code” and “depositional system” show collinearity, as did the “Temperature-Pressure” pair. For the GOM database, “Pressure-Depth” and “Gas Oil Ratio-Formation Volume Factor” were collinear. The relationship between these variable pairs aligns with conventional reservoir engineering concepts. As keeping both parameters will not add additional information to the predictive model and may potentially result in an overfit within the model, a single element from each of the variable pairs is eliminated to allow for a more stable model.

The final input variables were determined and summarized in Table 1 for the TORIS dataset and Table 2 for the GOM dataset.

The TORIS dataset comprises 10 categorical data types, and 14 numerical data types. The total data size is 389 values per column, for a total of 9,336 data points which was approximately ~9.6% of the original data base.

Table 1: Statistics of the TORIS Database Input Variables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 0.25 percentile | 0.5 percentile | 0.75 percentile | max |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

*Initial Machine Leaning Model*

*Tuning and Optimizing the Model*

*Evaluation against the “Hold-out” Data and the “Double Blind” Set*

# Discussion & Recommendations

Below we summarise our key findings and learnings in this of developing the ML model for RF prediction.

*Data Preparation*

When performing the initial data-analysis, it is essential to combine both data analysis techniques with domain knowledge. In both datasets, many values are missing. This can be dealt with by dropping missing variables. However, a better alternative would be to use domain knowledge to try to supplement the missing data. The use of domain knowledge and experience is critical because correlation and causality are not the same. Taking the example of “STOIIP vs Porosity” and “Gross Pay vs Porosity”, a correlation relationship is geologically reasonable in the former, but not in the latter.

*Optimization*

In an optimalization problem, the end goal is the minimization of a metric/ cost function. In our case, as we are dealing with regressors, the minimization must be applied to error functions like MAE, MSE and RMSE. The solution is best achieved via a gradient descent algorithm where the minimum of the cost function is achieved iteratively, largely because as the negative of the gradient is followed over time, it would theoretically reach a point where it will no longer be possible to decrease the cost function any further. In other words, when the number of iterations increases, the solution moves towards the minima which is defined by an optimal input hyperparameter set. The challenge is than finding that sweet spot of iterations and hyperparameters, and accepting the tradeoffs that occur, especially as it related to training time.

*Parametric Feature Importance*

When looking at the ranked feature importance, we observed that categoric data was not considered to be critical in any of the selected ML algorithms. The effects of categoric data, like trapping type, lithology or diagenetic overprint might be important characteristics in certain reservoirs. We think that the underrepresentation is due to the plethora of possible individual values. For example, in TORIS, there are 54 unique depositional environments (eolian, lacustrine, shelf, reefs, pinnacles etc). To the ML model, there are insufficient training instance of each to observe a strong trending behaviour. To increase the importance of categoric data, the categories should be grouped and simplified. A recommendation is for a skilled geologist or geophysicist with some background in data science to look at helping to simplify some of the categoric input parameters such that only 3 to 5 unique instances present.

*Conventional vs ML*

The developed ML models outperformed conventional empirical correlations such as Arps et al and Guthrie and Greenberger. The spread of datapoints was much wider with using both the Arps et al. (R2=0.21 and 0.006 in the TORIS and GOM model respectively) and Guthrie and Greenberger correlations (R2=0.15 and 0.007 in the TORIS and GOM model respectively). Further work would be to investigate the machine learning model and its comparative performance on other fields.

*Non-linearity*

Complementary to this work is the use of Artificial Neural Networks (ANN), which are a type of deep learning network complementary to ML. The key benefit of ANN is the ability to better handle non-linearities. Future work may investigate developing ANN networks to study the criticality of domain expertise and just how different RF predictions are if ANN is utilized instead of just ML.

# Conclusion

In this work, we have demonstrated that ML models form a good basis for estimating RF; however, applying general domain knowledge and sense-checking results is still very important. The use of any ML model is dependent on the purpose of the RF estimation and should be complementary (rather than contrasting) to conventional techniques (whether they be using analogs or empirical methods). Any ML model should not replace the need for decline based, or simulation-based estimates in fields with extensive production history. In such instances, ML can be used to determine ultimate RF potential. For early-stage RF estimate, ML models might perform better that Arps et al. correlation and Guthrie and Greenberger correlation, especially when data is sparse.

# Reference

|  |  |
| --- | --- |
| [1] | R. Lindsey, “Climate Change: Atmospheric carbon dioxide, NOAA Climate.gov,” [Online]. Available: https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide. [Accessed 25 April 2023]. |
| [2] | United States Environmental Protection Agency, “Sources of Greenhouse Gas Emissions,” 20 Jul 2022. [Online]. Available: from https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions. |
| [3] | J. G. D. Gooijer and R. J. Hyndman, “25 years of time series forecasting,” International Journal of Forecasting, vol. 22, no. 3, pp. 443-473, 2006. |
| [4] | E. S. Gardner, “Exponential smoothing: The state of the art,” Journal of Forecasting, vol. 4, no. 1, pp. 1-28, 1985. |
| [5] | R. D. Snyder, “Recursive Estimation of Dynamic Linear Models,” Journal of the Royal Statistical Society. Series B (Methodological), vol. 47, no. 2, p. 272–276, 1985. |
| [6] | R. H. Scumway and D. S. Stoffer, Time Series Regression and ARIMA Models, New York: Springer, 2000. |
| [7] | G. Huyot, K. Chiu, J. Higginson and N. Gait, “Analysis of Revisions in the Seasonal Adjustment of Data Using X-11-Arima Model-Based filters,” International Journal of Forecasting, vol. 2, no. 2, pp. 217-229, 1986. |
| [8] | S. H. Babbs and K. B. Nowman, “Kalman filtering of generalized Vasicek term structure models,” Journal of financial and quantitative analysis, vol. 34, no. 1, pp. 115-130, 1999. |
| [9] | M. P. F. P. H. &. S. N. R. Clements, “Forecasting economic and financial time-series with non-linear models,” International journal of forecasting, vol. 20, no. 2, pp. 169-183, 2004. |
| [10] | B. K. Ray, “Long-range forecasting of IBM product revenues using a seasonal fractionally differenced ARMA model,” International Journal of Forecasting, vol. 9, no. 2, pp. 255-269, 1993. |
| [11] | S. J. Taylor, “Forecasting the volatility of currency exchange rates,” International Journal of Forecasting, vol. 3, no. 1, pp. 159-170, 1987. |
| [12] | T. R. Willemain, C. N. Smart, J. H. Shockor and P. A. DeSautels, “Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method,” International Journal of Forecasting, vol. 10, no. 4, pp. 529-538, 1994. |
| [13] | J. J. Arps, “Analysis of Decline Curves,” Transactions of the AIME, Vols. SPE-945228-G, 1945. |
| [14] | A. L. L. a. T. P. Clark, “Production Forecasting with Logistic Growth Models,” in SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 2011. |
| [15] | A. Duong, “An Unconventional Rate Decline Approach for Tight and Fracture-Dominated Gas Wells,” in Canadian Unconventional Resources and International Petroleum Conference. , Calgary, Alberta, Canada, 2010. |
| [16] | D. R. J. A. P. A. D. a. T. A. B. Ilk, “Exponential vs. Hyperbolic Decline in Tight Gas Sands — Understanding the Origin and Implications for Reserve Estimates Using Arps' Decline Curves.,” in Paper presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 2008. |
| [17] | P. Valko, “Assigning value to stimulation in the Barnett Shale: a simultaneous analysis of 7000 plus production hystories and well completion records,” in SPE Hydraulic Fracturing Technology Conference , Woodlands, Texas. , 2009. |