SAND2022-11606C

16th International Conference on Greenhouse Gas Control Technologies, GHGT-16

23rd -27th October 2022, Lyon, France

Deep learning-based data assimilation in the latent space for real-time forecasting of geologic carbon storage

Hongkyu Yoon[a*], Jonghyun Harry Lee[b], Teeratorn Kadeethum[a]

[a]*Sandia National Laboratories, Albuquerque, NM 87125, USA*
[b]*University of Hawaii at Manoa, Honolulu, HI 96822, USA*

Abstract

This work presents a novel framework for deep learning (DL)-based CO_2 flow modeling of saturation and pressure distribution over time and data assimilation approach for real-time forecasting. In this work we develop and apply deep learning-based forward and inverse modeling framework to accelerate real-time forecasting. First, a combination of convolutional neural networks (CNNs), long-short term memory (LSTM), and deep neural networks (DNNs) is employed to construct DL-based surrogate modeling of multiphase CO_2 flow and pressure propagation over time in heterogeneous subsurface fields. Second, a variational autoencoder (VAE) in the data assimilation (DA) framework is developed for a real-time history matching of CO_2 and pressure plume development. The encoder in VAE works as a nonlinear dimension reduction method that determines a low-dimensional latent space "z" with encoded/compressed information, possibly performing better than traditional linear dimension reduction methods. For data assimilation, the latent space z, much smaller than the dimension of actual physical parameter space, is updated, which is conditioned on the observed data for effective sampling with better convergence on the low dimensional subspace. The decoder in VAE using the updated latent data is used to generate distribution of input parameters (e.g., permeability and porosity), which are used as input to CNN-LSTM-DNN forward model for fast prediction of CO_2 saturation and pressure plume development. With computationally efficient DL-based forward models the fast data assimilation can be completed in a timely fashion. Additionally, posterior covariance analysis gives us an ensemble generation of input parameters (e.g., permeability and porosity distribution) for computationally efficient uncertainty quantification with DL-based forward models. For demonstration purposes, two-dimensional (2D) example cases with various heterogeneous permeability fields are used to evaluate the feasibility of the proposed framework. The newly developed data assimilation framework demonstrates promising results of permeability estimation with uncertainty quantification and pressure prediction with the demonstration case. With full 3D reservoir scale problems, training time may take longer, but based on our experience with other 3D problems, the CNN-LSTM-DNN model can be trained very efficiently, the execution time of single ML forward model over transient solutions is very short (<< 1 second), and the latent dimension is typically small (e.g, 16-32). As a result, the computational efficiency of data assimilation framework will be similarly archived for a 3D reservoir model. Hence, the proposed data assimilation framework approach can be easily used to demonstrate in a realistic 3D problem for real-time forecasting.

Keywords: Deep learning; Geologic carbon storage; Data assimilation; Convolutional Neural network; Long short term memory; Variational AutoEncoder; Latent space

1. Introduction

This work presents a novel framework for deep learning-based CO_2 flow modeling of pressure and saturation distribution over time and data assimilation approach for real-time forecasting of pressure response. In traditional history matching/data assimilation approaches [1-3] the biggest challenging obstacle for real-time forecasting is that these approaches need to account for high dimensional problems which typically require the high computational burdens such as matrix calculations. Secondly, a number of observations are generally much smaller than unknown parameters such as permeability and porosity, which causes the inverse model to be underdetermined and ill-posed. Thirdly, primary model parameters such as permeability and porosity are uncertain, requiring us to account for the

uncertainty in the model and data. Hence, we often need to run a number of forward model runs to perform inverse modeling with uncertainty quantification such as ensemble data assimilation. Although many new algorithms and advanced hardware development enabled us to perform faster, it is still very challenging to achieve real-time forecasting.

In this work we overcome two main challenges by constructing ML-driven fast forward model(s) to resolve computational burdens of reservoir simulations and by reducing dimension of primary variables into low dimensional latent space to resolve the number of forward model runs required for inversion. First, combination of convolutional neural networks (CNNs), long-short term memory (LSTM), and deep neural networks (DNNs) is employed to construct deep learning (DL)-based surrogate modeling of predicting multiphase CO_2 flow and pressure propagation over time in heterogeneous subsurface fields. Second, a variational autoencoder (VAE) in the data assimilation (DA) framework is developed for a real-time history matching of CO_2 and pressure plume development. In contrast to linear dimensional methods such as principal component analysis [e.g., 24, 6], the encoder in VAE works as a nonlinear dimension reduction method that determines a low-dimensional latent space "z" with encoded/compressed information, possibly performing better than traditional linear dimension reduction methods. For data assimilation, the latent space z is updated conditioned on the observed data for effective sampling with better convergence on the low dimensional subspace. The decoder in VAE using the updated latent data is used to generate distribution of state variables (i.e., permeability and porosity), which are used as input to CNN-LSTM-DNN model for fast forecasting of CO_2 saturation and pressure plume development. With computationally efficient CNN-LSTM-DNN forward model the fast data assimilation can be completed in a timely fashion for real-time history matching of CO_2 operations and forecasting CO_2 and pressure plume development. In this report, we report the overview of VAE and data assimilation framework and two examples to demonstrate how the newly developed data assimilation framework is performed with DL-driven fast forward model constructed with synthetic data generated from the Clastic Shelf dataset.

2. ML Model Structure and Description

2.1. *ML Forward model (CNN-LSTMs-DNN)*

ML model architecture: A DL-driven reduced order model is constructed based on Convolutional neural network (CNN)-Long Short Term Memory (LSTM)-Dense Neural Network (DNN) architecture for pressure prediction. This ML architecture has been originally developed to predict pressure and CO_2 saturation distribution in three-dimensional (3-D) Clastic Shelf data generated as a part of US DOE SMART-CS Initiative project [5] based on the previous study [6-7]. Here we developed a 2D version of CNN-LSTM-DNN model to make a pressure prediction. The model architecture includes two separate inputs consisting of (1) static and/or spatial data such as permeability, porosity, and final cumulative injection amounts at four well locations and (2) temporal or dynamic data such as cumulative injection amount over time per well. For static input, 2D distribution of permeability, porosity, and final cumulative injection amount are used to construct an input dimension with three features (i.e., n_features = 3) and for dynamic input, cumulative injection amount per well is used.

For static input, six 2D CNN layers as an encoder is used to project the state variables into a low-dimensional subspace which is concatenated with dynamic data to construct time-series input into one LSTM layer to assimilate the time-series dynamics in the low dimensional space, and a deep neural network with four dense layers as decoder to project the solution back into high dimensional space to match the dimension of model prediction. Each CNN layer block has the same architecture including 2D CNN followed by batch normalization and maxpooling layer. The number of LSTM unit is 512. The model architecture was optimized, however, other key hyperparameters are manually evaluated to find relatively optimal values. For CNN and dense layers, rectified linear unit (Relu) and leaky Relu activation functions are used, respectively. In particular, a TimeDistributed layer wrapper available in Keras is employed to allow each dense layer to maintain the same model parameters for every temporal slice (i.e., every time step in the current model).

Loss functions: Custom loss functions are constructed based on known physics and domain expert knowledge. In addition to a reconstruction mean square error (MSE) the pressure model employs loss equations to account for the initial pressure distribution and 3×3 pressure distribution around each injection well. Each additional loss term has its own weight that can be considered an optimizer.

2.2. *Nonlinear dimension reduction via Variational Autoencoder*

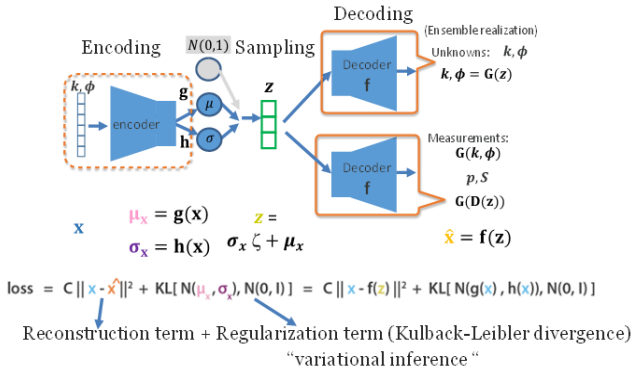

Figure 1. Schematic of variational autoencoder

VAE model description and architecture: Variational autoencoders (VAEs) are a variant of AEs that tackles the problem of the latent space irregularity by making the encoder return a *desirable* distribution over the latent space using variational (i.e., deterministic optimization in terms of the mean of distribution) techniques. The idea is to set a parameterized family of distribution as a target in the latent space by minimizing a given approximation error metric as in the loss function shown in the Figure 1. Once trained, input data are encoded in a low-dimensional latent space such that the outputs returned by the encoder are enforced to follow a well-known distribution, for example, a standard normal distribution. Then with the trained decoder, we simply feed samples from a standard normal distribution to generate meaningful input parameters such as permeability fields.

2.3. *Data Assimilation Framework*

The developed method utilizes one of the current efficient data assimilation methods such as Bayesian Inversion with Gaussian Prior and the posterior probability is maximized using the Gauss-Newton (GN) method. At each GN iteration, one has to construct a Jacobian matrix (**J**), which can be computed using the adjoint-based method. The schematic of our data assimilation approach is shown in Figure 2.

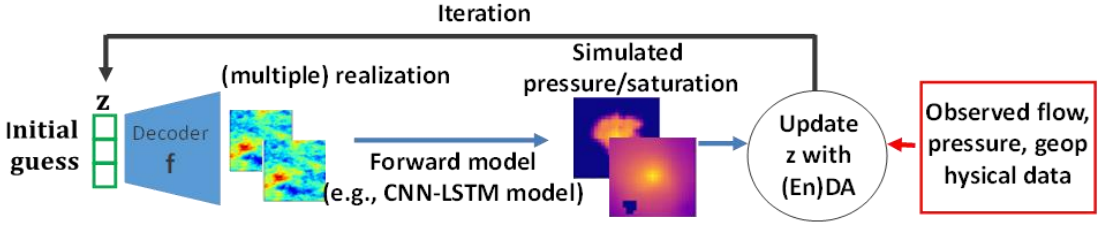

Figure 2. Schematic of data assimilation framework. Latent space "**z**" obtained by VAE, is updated in (En)DA-based methods with various measured data.

As shown in Figure 2, we use VAE and its decoder to map the permeability **k** field to the latent vector z whose dimension is much smaller than the original dimension of **k** field while ensuring a good approximation accuracy as below. With l Gauss Newton iterations from $m^0 = m_{prior}$

$$m^{l+1} = m^0 + C_{prior}J\left(JC_{prior}J^T + C_{obs}\right)^{-1}\left(y - G\left(m^l\right) + J\left(m^l - m^0\right)\right)$$

With (nonlinear) dimension reduction G of dim (m) to the dimension of the latent space with $dim\ (z)$, $y = G(D(z))$ where $dim\ (z) \ll dim\ (m)$ and the latent space variable is updated as

$$z^{l+1} = z^l + \alpha\left(J_z^T C_{obs}^{-1} J_z + C_{prior_z}^{-1}\right)^{-1}\left(y - G\left(D\left(z^l\right)\right) - C_{prior_z}^{-1} z^l\right)$$

with step length (learning rate α). And the posterior covariance is given as

$$C_{posterior_z} = \left(J_z^T C_{obs}^{-1} J_z + C_{prior_z}^{-1}\right)^{-1}.$$

In this work, Jacobian matrix can be easily approximated with a finite difference method with the number of unknowns (+ 1) forward model runs. As seen in the equations above, the unique feature of our framework is that the data assimilation is performed in small nonlinear latent space of unknown parameters with dim(z). Hence, it requires only "**dim(z)**" forward model executions at each iteration instead of dim(m) or dim(obs). Finally, with VAE it constructs the prior**.**

3. Results

3.1. *CNN-LSTM-DNN Forward Model Data*

A total of 70 cases from the Clastic Shelf data are used to train, validate, and test the CNN-LSTM-DNN model. The best trained model was selected based on the loss of the validation set which was not used directly for model training. Figure 3 shows key results of model prediction of pressure for cases P50-1 (case#48) and P50-16 (case#63) and parity plots of pressure fields for two P50 and two P75 cases.

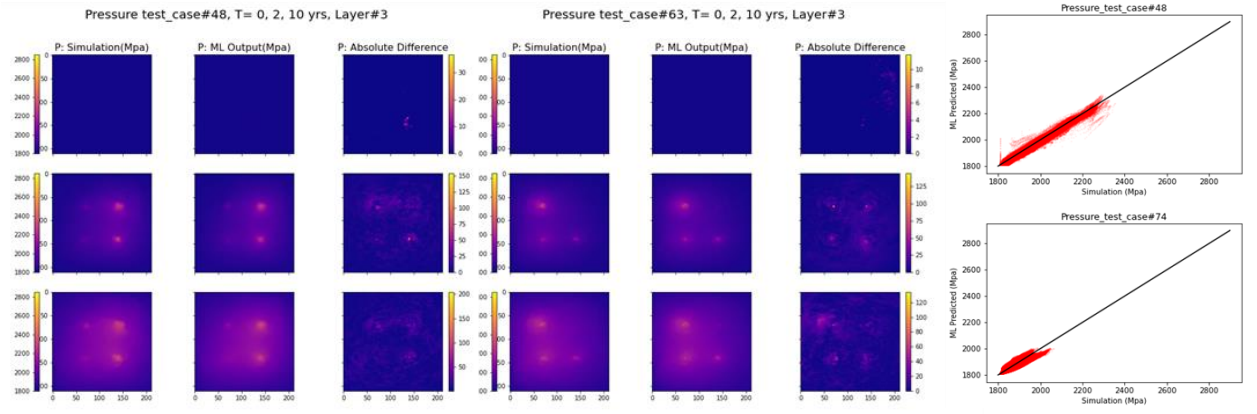

Figure 3. Selected pressure field cases (right, middle) and parity plots (left) of CNN-LSTM-DNN model for pressure prediction. Simulation values on the x-axis are from the Clastic Shelf model data (i.e., truth values in this synthetic cases). Case number represents permeability and porosity fields corresponding to the probability of the distribution.

3.2. *CNN-LSTM-DNN Forward Model Data*

As shown in Figure 2, the decoder is used as a generative model to construct permeability fields from the latent space (**z**). Figure 4 shows the reconstructed permeability fields through the latent space, which resemble training data very well. As described, the training data is encoded to learn how to construct the latent space, which is then used to reconstruct the permeability field. In this case the latent space dimension was 32 that is the same for the data assimilation case study in the following section.

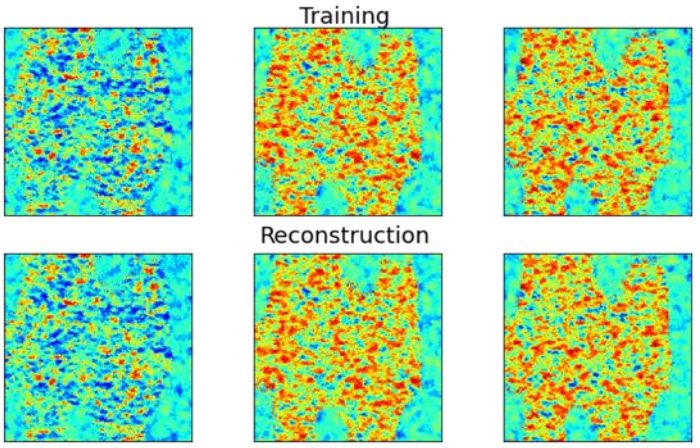

Figure 4. Performance of the decoder to construct 2D permeability fields from training Clastic Shelf permeability fields.

4. Demonstration with Clastic Shelf data

4.1. *Case Description*

A demonstration case with Clastic Shelf data is developed to perform forecasting of the pressure distribution due to CO_2 injection by updating the permeability field as new monitoring data become available over time. For performance comparison, case P50-1 is chosen as a testing case and the layer #3 in the Clastic Shelf model where the reservoir top layer is located is used in this demonstration case. Overall, Figure 5 shows two-dimensional (2D) permeability, porosity, pressure, and saturation distributions. Permeability (& porosity) and CO_2 injection rates at four injection wells are input and observed data include pressure data at four injection (INJ) and two monitoring wells (MW) (monthly data up to 2 yrs, followed by yearly data to 10 yrs).

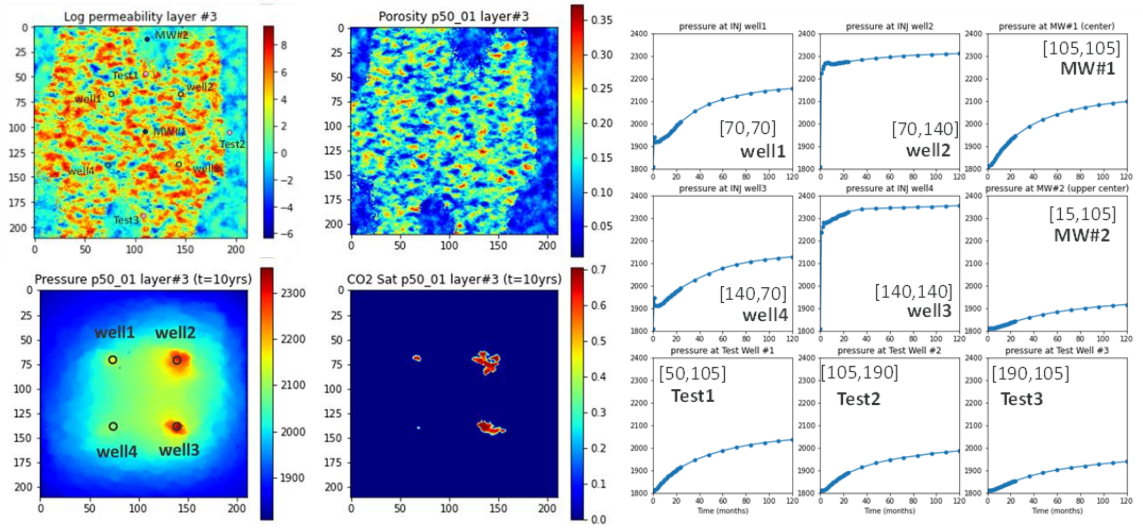

Figure 5. (left) Permeability, porosity, pressure, and saturation distributions in the layer #3 (top layer of the reservoir formation) of case P50-1 (case#48) in the Clastic Shelf data. Pressure and saturation distribution at the last time step (10 yrs) are shown with four injection wells. (right) Pressure profiles at four injection, two monitoring, and three testing wells. Two numbers in a blanket indicate the location of each well.

For the demonstration purpose, the observed pressure data at four injection (INJ) and two monitoring wells (MWs) are used to calibrate the permeability field during calibration stage up to Obs=1.5, 2, 5, and 10 yrs (Figure 5). For the following forecasting period up to 10 yrs, performance of the calibrated model with an updated permeability field is compared against the observed data to evaluate the data assimilation framework developed in this work. For an example of Obs=1.5 yrs, pressure data up to 1.5 yrs are used first for model calibration and then data assimilation with pressure data up to 2, 5, and 10 yrs are performed. For performance comparison, case P50-1 is used as testing case with two monitoring wells data. Pressure change over time at three testing locations (see Figure 5), 2D pressure fields in the entire domain at selected times, update of permeability fields and uncertainty quantification are also reported.

4.2. *Demonstration Results*

Pressure data observed at four injection wells only and two additional monitoring wells up to varying years (Obs = 1.5, 2, 5, and 10 yrs) are used for data assimilation in this section. Estimated permeability fields with different observed pressure data (Obs=1.5, 2, 5, 10 yrs) and corresponding standard deviation fields of 1,000 permeability fields generated by the post covariance analysis are shown in Figure 6. These stochastic realizations can be used to construct the uncertainty bound of predicted pressure (results not shown here). Predicted pressure data at three testing locations over time are presented in Figure 7. 2D pressure fields predicted by the calibrated model at selected times (1.5, 2, 5, and 10 yrs) are presented in Figure 8.

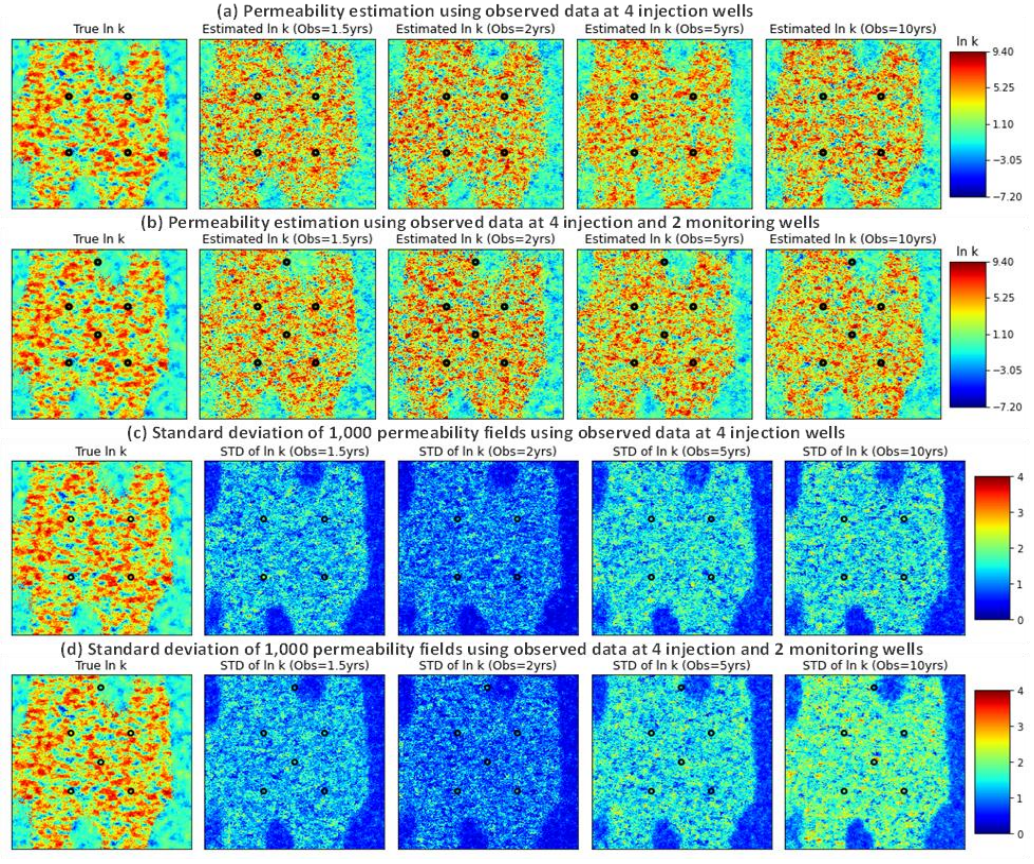

Figure 6. (a-b) Estimated permeability fields using observed pressure data at four injection wells only (Scenario #1) and two additional monitoring wells (Scenario #2) using different observed periods (Obs = 1.5, 2, 5, and 10 yrs). The locations of injection and monitoring wells are shown as circles. (c-d) Standard deviation fields of 1,000 permeability realizations generated based on the calibrated permeability fields using post covariance analysis are plotted together with the true permeability field (case P50-1). Stochastic realizations can be used to estimate the uncertainty bound of pressure prediction as well.

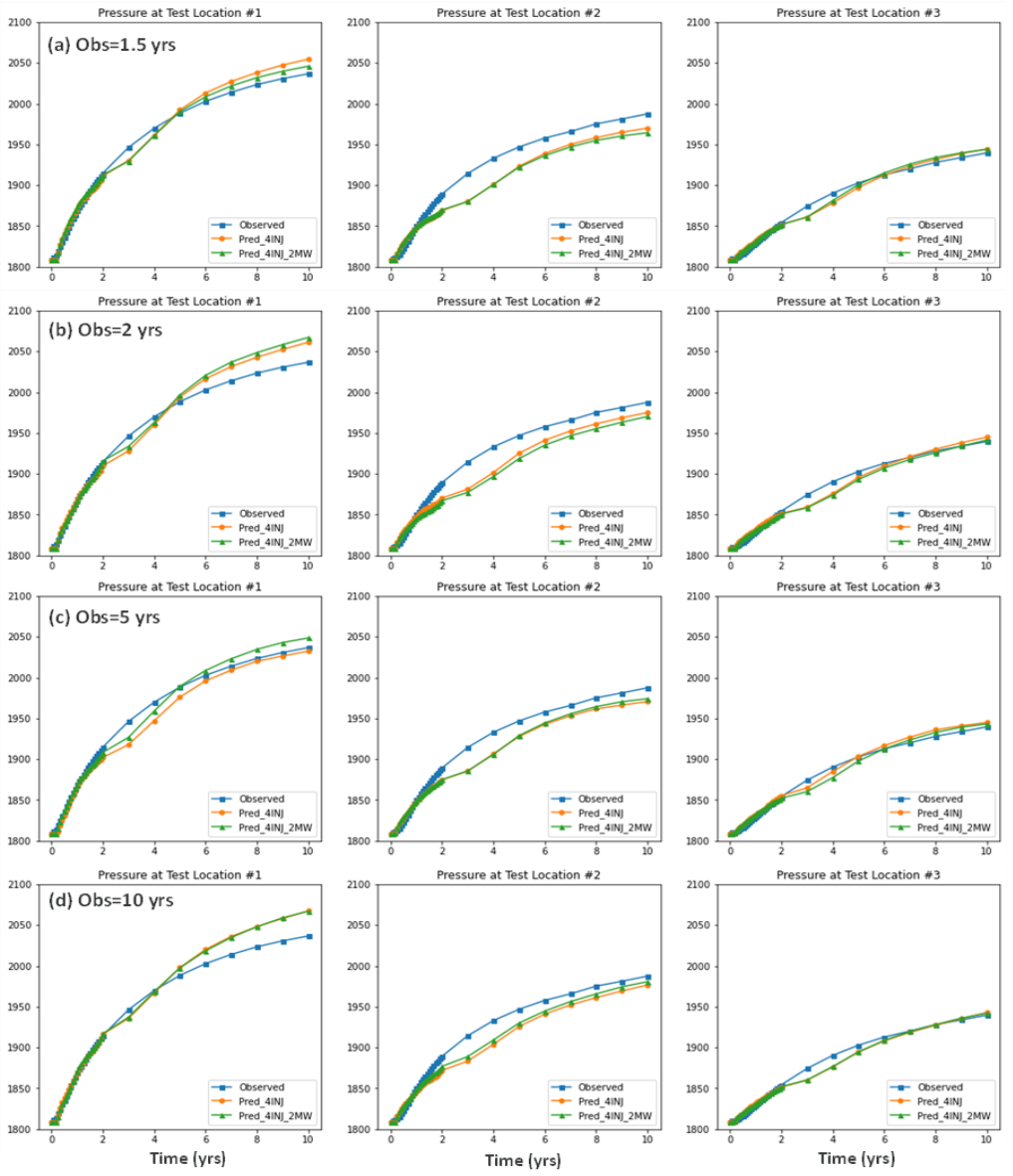

Figure 7. Comparison between observed and predicted pressure data over time at three testing locations for two Scenarios with varying observed periods (Obs = 1.5, 2, 5, and 10 yrs). Pressure data observed at four injection wells (4INJ) and two additional monitoring wells (4INJ_2MWs) are known. Varying observed data (Obs = 1.5, 2, 5, and 10 yrs) are used for data assimilation and the rest of pressure data is the ML based prediction using the permeability field in Figure 6.

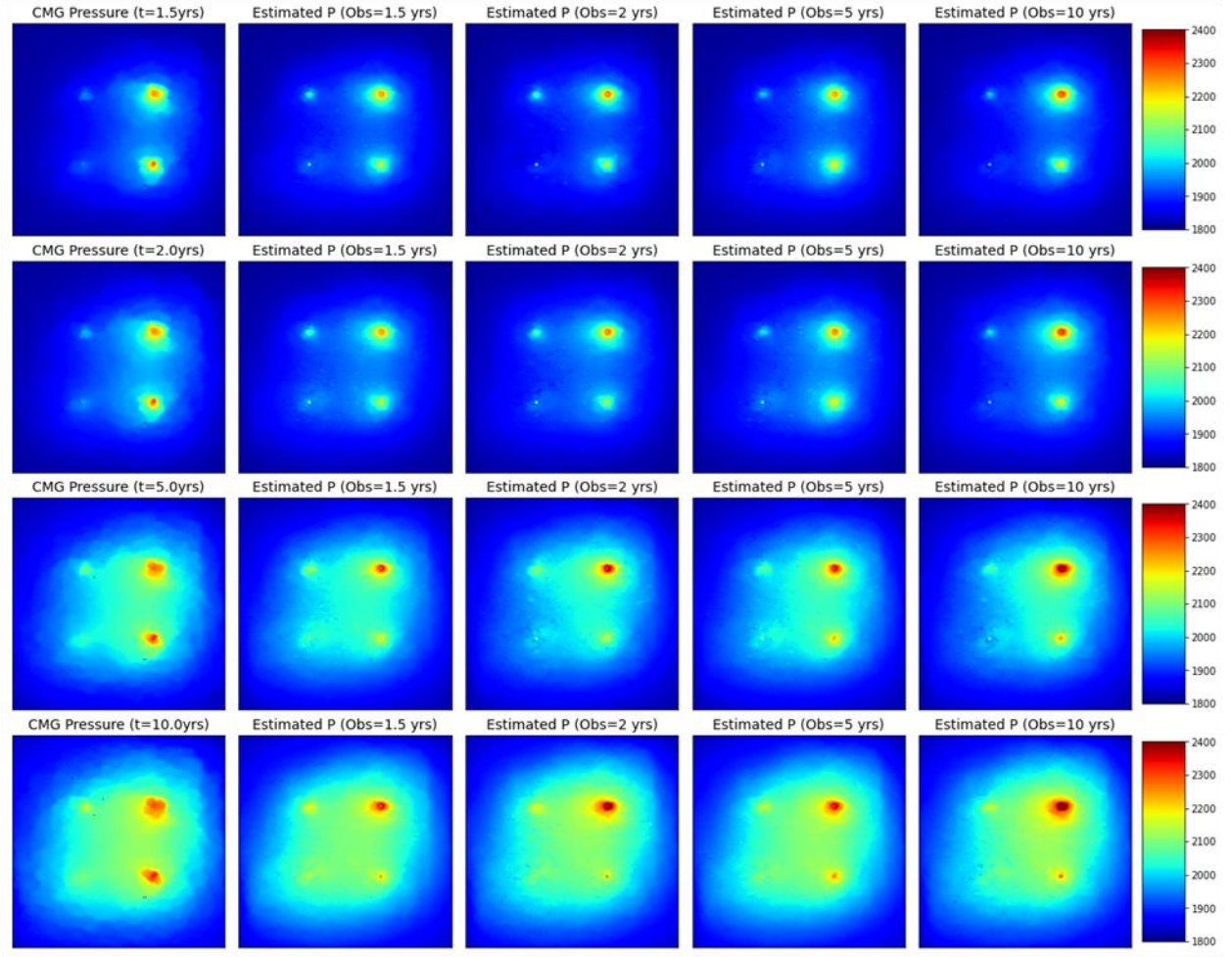

Figure 8. Comparison between observed (CMG data) and predicted pressure field at four different times. Observed pressure data at four injection wells and two monitoring wells up to Obs =1.5, 2, 5, 10 are used for data assimilation. For Obs =1.5 yrs, the rest of pressure fields after 1.5 yrs are ML-based predictions at 2, 5, and 10 yrs using the permeability field estimated (see Figure 6b).

Overall, Figure 6 shows that estimated permeability fields for all 8 cases (four different observed periods per each monitoring scenario) resemble the target permeability field very well. Although subtle changes in permeability fields (e.g., changes around wells) are shown locally, the visual comparison shows remarkable similarities. This indicates that pressure data at a limited number of wells may contain relatively good response over the model domain. In addition, the decoder developed in this work (see Figure 4) can generate a relatively high quality of permeability fields which can represent key features of the clastic shelf data. It is noted that the initial field was generated relatively uniformly using the latent variables with zero mean and unity of variance. Standard deviation fields (Figure 7c-d) that are computed based on 1,000 permeability realizations generated by the post covariance analysis show that both Scenarios for Obs=1.5, 2, and 5 yrs have relatively similar patterns until Obs=10 yrs when Scenario #2 with two additional MWs shows higher standard deviations than Scenario #1. The similar patterns of permeability estimation and standard deviations of stochastic fields using Obs=1.5, 2, and 5 yrs for both Scenarios indicate that pressure data between Obs=1.5 yrs and up to Obs=5 yrs may have similar information regarding the permeability and pressure response, but additional data after 5 yrs may contain more pressure response, particularly with pressure data at two monitoring wells (as in high standard deviation field using Obs=10yrs for Scenario #2).

Figure 8 shows the pressure profile predicted at three testing locations (see Figure 6 for testing locations) with the forward ML model using the permeability fields estimated using varying observed data (Obs=1.5, 2, 5, 10 yrs).

Overall, predicted pressure profiles at three testing locations for all cases match the observed data well. Particularly the monthly data up to 2 yrs are quite well predicted at all three testing locations while the pressure data after 2 yrs (i.e., yearly observed data) are predicted with varying accuracy. Comparison of results with Obs=1.5 and 10 yrs at testing locations #1 and #3 (Figure 8a&d) shows that pressure up to 2 yrs is predicted well at both locations, while pressure after 2 yrs is predicted differently at two locations. At test location #1, pressure prediction after 2 yrs tends to be better with Obs=1.5 yrs compared to Obs=10yrs, while a few yearly pressure data between 2 and 5 yrs are relatively predicted well with Obs=10yrs, but less accurately after 5 yrs with Obs=10 yrs. At testing location #3, pressure prediction after 2 yrs is relatively similar. This reveals the impact of local heterogeneity patterns on the pressure field. The relatively higher error at test location #2 is mostly due to the location of test #3 which is located outside the high permeability zone. Since pressure perturbation is caused by CO_2 injection at four different wells in the high permeability zone (i.e., lithoface), test #3 outside this zone is less sensitive to pressure data at additional monitoring wells and less accurately predicted.

Pressure fields predicted at 1.5, 2, 5, and 10 yrs using the estimated permeability fields (Figure 7) are presented in Figure 8. As expected in the estimated permeability fields, predicted pressure fields match the CMG data (i.e., truth in this synthetic case) very well. MAPE values in Table 1 shows a slightly different tendency over time. However, the scale of MAPE values are quite small compared to the range of pressure fields (~1800 to ~2400 psi). This clearly demonstrates that the data assimilation framework works very well to recover the pressure fields with relatively limited observed data.

Table 1. Mean absolute pressure error (MAPE) values for each case predicted at T=1.5, 2, 5, and 10 yrs with varying observed data (Obs=1.5, 2, 5, 10 yrs). Pressure fields in Figure 8 are used to compute MAPE values.

	Scenario: Observed data at 4 INJs, 2 MWS			
	1.5 yrs	2 yrs	5 yrs	10 yrs
T = 1.5 yrs	0.070	0.165	0.041	0.058
T = 2 yrs	0.153	0.209	0.064	0.090
T = 5 yrs	0.210	0.287	0.091	0.182
T = 10 yrs	0.272	0.246	0.088	0.158

4.3. *Computational Times*

Overall, training of CNN-LSTM-DNN and VAE decoder models was performed using NVIDIA Quadro RTX 6000 and forward model prediction and data assimilation were performed using a laptop with 32G RAM - Intel(R) Core(TM) i9-10885H CPU @ 2.40GHz. It is noted that execution of trained ML models does not need GPU computing. Overall, training of CNN-LSTM-DNN takes ~ 10-15 min with 11 min for the forward model used in this demonstration. VAE training takes ~ 10min. Each run of CNN-LSTM-DNN and decoder of VAE models during data assimilation takes ~10 milliseconds. For data assimilation, a total number of iterations vary depending on the convergence, but it ranges between 20 and 42 for cases reported in this work. Overall, it takes 3-5 min for data assimilation.

5. **Recommended Next Steps**

The newly developed data assimilation framework with CNN-LSTM-DNN forward model and decoder of permeability generation demonstrates promising results of permeability estimation with uncertainty quantification and pressure prediction with the demonstration case. Although the demonstration case was performed with 2D examples, our framework is not limited to 2D, but readily applicable for 3D full reservoir scale problems. Here we demonstrate the framework with pressure data, however, it is straightforward to extend to a problem for both pressure and saturation. It is noted that pressure prediction at the injection wells has slightly higher errors in the ML model although this is quite common in ML-driven models. In the near future, we will evaluate a recently developed continuous conditional generative adversarial network (CcGAN) [8,9] which seems to work for challenging datasets well including pressure at the injection wells. One key aspect of joint data assimilation is to represent reservoir characteristics (channel preferential flow patterns & barriers) and faults/fractures-matrix systems in the inverse modeling toolbox through multiple datasets that will have different types of spatio-temporal information. This will be achieved by updating the fast ML-based forward modeling engine for coupled hydro-mechanical processes (e.g.,

poroelaticity) and data assimilation of both point measurement data and distributed monitoring data with the inverse modeling engine developed in this work. Generative ML models (VAE and GAN) can be utilized to represent joint fracture/fault and continuum-based reservoir models for data assimilation with uncertainty quantification.

Acknowledgements

This work was supported by US Department of Energy Office of Fossil Energy and Carbon Management, Science-Informed Machine Learning to Accelerate Real Time Decisions-Carbon Storage (SMART-CS) initiative. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] H. Yoon, D. B. Hart, S. McKenna. (2013). Parameter estimation and predictive uncertainty in stochastic inverse modeling of groundwater flow: Comparing null-space Monte Carlo and multiple starting point methods. Water Resources Research, 49(1): 536-553

[2] J. Lee, H. Yoon, P. K. Kitanidis, C. J. Werth, A. J. Valocchi. (2016). Scalable subsurface inverse modeling of huge data sets with an application to tracer concentration breakthrough data from magnetic resonance imaging. Water Resources Research, 52(7): 5213-5231

[3] H. Ghorbanidehno, A. Kokkinaki, J. Lee, E. Darve. (2020). Recent developments in fast and scalable inverse modeling and data assimilation methods in hydrology, Journal of Hydrology, 591: 125266

[4] J. Lee, P. K. Kitanidis. (2014). Large-scale hydraulic tomography and joint inversion of head and tracer data using the principal component geostatistical approach (PCGA), Water Resources Research, 50(7): 5410-5427

[5] US DOE SMART-CS Initiative project, https://edx.netl.doe.gov/smart/.

[6] Bosshart N.A., Azzolina, N.A., Ayash, S.C., Peck, W.D., Gorecki, C.D., Ge, J., Jiang, T., and Dotzenrod, N.W., 2018, Quantifying the effects of depositional environment on deep saline formation CO2 storage efficiency and rate: International Journal of Greenhouse Gas Control, v. 69, p. 8–19.

[7] Bosshart N.A., 2017, Ayash, S.C., Azzolina, N.A., Peck, W.D., Gorecki, C.D., Ge, J., Jiang, T., Burton-Kelly, M.E., Anderson, P.W., Dotzenrod, N.W., and Gorz, A.J., 2017, Optimizing and quantifying CO2 storage resource in saline formations and hydrocarbon reservoirs: Final report for U.S. Department of Energy National Energy Technology Laboratory Cooperative Agreement No. DE-FE0009114, EERC Publication 2017-EERC-06-18, Grand Forks, North Dakota, Energy & Environmental Research Center, June.

[8] T. Kadeethum, D. O'Malley, Y. Choi, H. Viswanathan, N. Bouklas, and H. Yoon. (2022). Continuous conditional generative adversarial networks for data-driven solutions of poroelasticity with heterogeneous material properties, Computers & Geosciences, 167, 105212, https://doi.org/10.1016/j.cageo.2022.105212.

[9] T. Kadeethum, F. Ballarin, Y. Choi, D. O'Malley, H. Yoon, and N. Bouklas. (2022). Non-intrusive reduced order modeling of natural convection in porous media using convolutional autoencoders: comparison with linear subspace techniques. Advances in Water Resources, 104098.