The Pennsylvania State University
The Graduate School

PHYSICS-INFORMED DEEP LEARNING FOR PREDICTION OF CO_2 STORAGE SITE RESPONSE

A Thesis in
Computer Science and Engineering
by
Sumedha Prathipati

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2021

The thesis of Sumedha Prathipati was reviewed and approved by the following:

Daniel Kifer
Professor of Computer Science and Engineering
Thesis Advisor

Clyde Lee Giles
David Reese Professor at the College of Information Sciences and Technology

Parisa Shokouhi
Associate Professor of Engineering Science and Mechanics

Chitaranjan Das
Distinguished Professor of Computer Science and Engineering
Head of the Department of Computer Science and Engineering

Abstract

The accurate prediction of CO_2 saturation and pressure is crucial for effective management of carbon storage sites. As the extensive pre-injection tests of such carbon sites are not strategically feasible, the reservoir behavior can be replicated by numerically simulating the CO_2 flow. Prior knowledge of the carbon storage site response to injection of CO_2 and the underlying geological formations of the reservoir are required in order to develop an economic strategy to monitor the storage projects. In recent years, these accurate simulations have been modeled by data-driven machine learning methods. Although such models are computationally inexpensive, these may overfit the data and not take the underlying physical laws into account.

Here, a physics-informed deep learning method which incorporates flow equations is proposed to predict a storage site response to injection of CO_2 gas. This approach is demonstrated using a 3D synthetic dataset. The model approximates the evolution of CO_2 saturation, pressure and water production rate with respect to space and time, given the initial permeability, porosity and CO_2 injection rate. First, a data-driven Long Short-Term Memory (LSTM) model is developed as the baseline. The physics-informed LSTM model is built on this by adding constraints to the cost function which are defined by the governing physics equations for a two-phase flow system. The proposed approach can be incorporated in carbon storage management to accurately predict the storage site response to CO_2 injection.

Table of Contents

List of Figures

List of Tables

List of Symbols

K	Permeability, p. 6
ϕ	Porosity, p. 6
q_g	Injection rate, p. 6
p	Pressure, p. 6
S_g	Gas saturation, p. 6
q_w	Water production rate, p. 6
μ_w	Viscosity of water, p. 7
μ_g	Viscosity of gas, p. 7
ρ_w	Mass density of water, p. 7
ρ_g	Mass density of gas, p. 7
∇	Gradient with respect to the spatial coordinates x, y, z, p .7
$\cdot$	Inner product, p. 7
t	Time step, p. 8

Acknowledgments

This work was completed as part of the Science-informed Machine learning to Accelerate Real Time decision making for Carbon Storage (SMART-CS) initiative (edx.netl.doe.gov /SMART). Support for this initiative was provided by the U.S. Department of Energy's (DOE) Office of Fossil Energy's Carbon Storage Research program through the National Energy Technology Laboratory (NETL). This work is a part of the paper, titled 'Physics-informed deep learning for prediction of CO_2 storage site response', published in the Journal of Contaminant Hydrology, 2021. The findings and conclusions of this work are those of the author and do not necessarily reflect the view of the U.S. Department of Energy or the Pennsylvania State University's College of Engineering.

I would like to express my deepest gratitude to my advisor Professor Daniel Kifer for his continuous support and guidance throughout my Masters study and research. His suggestions and valuable insights carried me through all the stages of my thesis. I am indebted to Professor Parisa Shokouhi for her constant assistance and counsel, without which this project would not have been possible. I would also like to convey my sincere thanks to Professor Lee Giles for his insightful advice and recommendations.

Further, I extend my thanks to Professor Seyyed Hosseini and my labmate Vikas Kumar for their support throughout the project. I am grateful to the National Energy Technology Laboratory for providing this opportunity and their constant direction. My thesis would not be complete without the love and encouragement of my family, Murali Babu, Aparna and Sreemanth. I would also like to show my deep appreciation to my friends at Penn State for constantly motivating me to give my best each day. I have learnt a lot by interacting and exchanging ideas with several faculty members and students, for which I will be forever grateful for.

Chapter 1

Introduction

It is challenging to develop an effective strategy to monitor the CO_2 storage sites because it requires an extensive knowledge of the site response to injection of CO_2 gas and the underlying physical laws governing the geological formations. [1] Furthermore, there is very limited experience related to monitoring large CO_2 storage sites. To accommodate for the expensive pre-injection field tests, the state-of-the-art practice is to numerically simulate the reservoir to predict the CO_2 saturation, pressure and water production rates. However, the management of real-time CO_2 storage sites is restricted by large-scale dynamic simulations of fluid flow and the uncertainties in the associated controlling parameters [1]. One approach to overcome this is to incorporate data-driven modeling with these numerical simulations. This will enable an approach to rapidly examine the CO_2 site behavior in various operational scenarios and also aids in analyzing the impact of several operational parameters on different site outcomes.

In recent times, data-driven machine learning models were developed using synthetic data generated by numerical simulations to forecast the carbon storage site behavior [1]. These models include conditional deep convolutional generative adversarial network (cDC-GAN) for predicting CO_2 plume in heterogeneous formations [2], artificial neural networks for forecasting future reservoir performance [3], multi-adaptive regression spline and random forest to estimate cumulative capacity of a CO_2 storage site [4], and multi-variate regression analysis to evaluate CO_2 storage efficiency in aquifer-caprock systems [5]. The data-driven models approximate the dynamic system within the bounds of the training data, which limits the development of a reasonably generalizable model. Although this can be overcome by running extensive simulations to train the model, it remains unfeasible. Consequently, this could overfit the model and may be inconsistent

with the governing physical laws.

To incorporate the physical principles with machine learning models, physics-informed deep learning is proposed as a supervised learning strategy which respects the partial differential equations (PDEs) [6]. This is done by adding physical consistency constraints to the loss function in order to penalize the deviation from the governing physical equations [7]. Including domain knowledge in the deep learning architecture results in more accurate predictions even when the training data is inadequate [7]. For this reason, physics-informed deep learning is increasingly used for forecasting in different domains. One such application is to estimate hydraulic conductivity in saturated and unsaturated flows by enforcing the Darcy's law to minimize the PDE residual in the simulation domain [8]. From this work, it is inferred that the accuracy of estimations for sparsely observed functions is improved by incorporating physics constraints to the model. Therefore, it is possible to train deep neural networks when there are few available direct estimations of the target functions.

Long Short-Term Memory (LSTM) models are popularly used for sequence prediction problems. LSTM has the ability to learn the temporal relationships in time-series data [9]. The LSTM network comprises of memory blocks, which store and fetch information over periods of time. The memory blocks use recurrently connected cells to learn dependencies between data at two time frames, and then transfer this inference to the next time frame [9]. Hence, this architecture is crucial when dealing with data which has temporal dependencies and evolves with time. The numerically simulated data captures the behavior of the reservoir over time and spatial coordinates, making this model an appropriate choice for approximating the storage site output variables.

Accordingly, a set of two-phase flow equations are added as constraints to the LSTM model. The LSTM model takes the initial geological properties of the reservoir (permeability and porosity distributions) and the CO_2 injection rates as the input to predict the CO_2 saturation, pressure fields and water production rates. Four LSTM architectures are developed, with the first two being data-driven models and the remaining being physics-informed models. These models include Single-output training, Multi-output training, Physics-informed on a supervised dataset and Physics-informed on an additional unsupervised dataset. These unsupervised datapoints are constructed by generating equally spaced temporal and spatial coordinates between two neighboring

time or position values in the original training dataset. Therefore, this dataset does not contain the simulated values of the target variables (gas saturation, pressure and water production rate). The data-driven models are used as a baseline, with respect to which the performance of the physics-informed models is compared. The quantitative and qualitative prediction results are discussed for each of these models.

Chapter 2

Dataset

The dynamic reservoir simulations are generated using a commercial software by CMG (Computer Modeling Group Ltd.) [1]. The synthetic dataset used here is from a 3D 'toy reservoir' of depth 7500 ft. The reservoir model consists a 25x25x3 simulation grid with three layers along the x, y and z spatial directions. Permeability and porosity are heterogeneous and taken from an offshore Gulf of Mexico geological model. Three realizations of permeability and porosity represent three different geological conditions, categorized as P10, P50 and P90. The dataset comprises a total of 27 realizations, which are obtained by generating nine different simulations for each of the three geological conditions. Each simulation is run by varying only the injection rate and not modifying any of the other parameters for 72 time steps. The reservoir model includes one water production well situated at grid block (23, 23) and one gas injection well at grid block (13, 13).

In summary, the dataset consists of 27 simulations with each simulation containing values of the permeability, porosity and CO_2 injection rate (inputs) along with the CO_2 saturation, pressure and water production rate (outputs). Each of these variables are reported on the 25x25x3 (x,y,z) grid points across 72 time steps ranging over 2161 days.

Realization	**Permeability**	**Porosity**	**Injection rate (MMscf)**
$P10$	heterogeneous	heterogeneous	$1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6$
$P50$	heterogeneous	heterogeneous	$1, 2, 3, 4, 5, 6, 7, 8, 9$
$P90$	heterogeneous	heterogeneous	$1, 2, 3, 4, 5, 6, 7, 8, 9$

Table 2.1. Overview of simulations

Table 2.1 illustrates the injection rates used for each of the 27 simulations, categorized into each of their respective realizations. The test data includes randomly selected 3

simulations corresponding to each of the different realizations. The training data is the remaining 24 simulations (8 simulations from each realization). For training and hyperparameter selection, k-fold (8-fold) cross-validation is used with each validation fold comprising of 3 simulations for different realizations. The hyperparameters associated with the lowest cross-validation loss on these 24 simulations are evaluated on the test data [1].

2.1 Input/Output

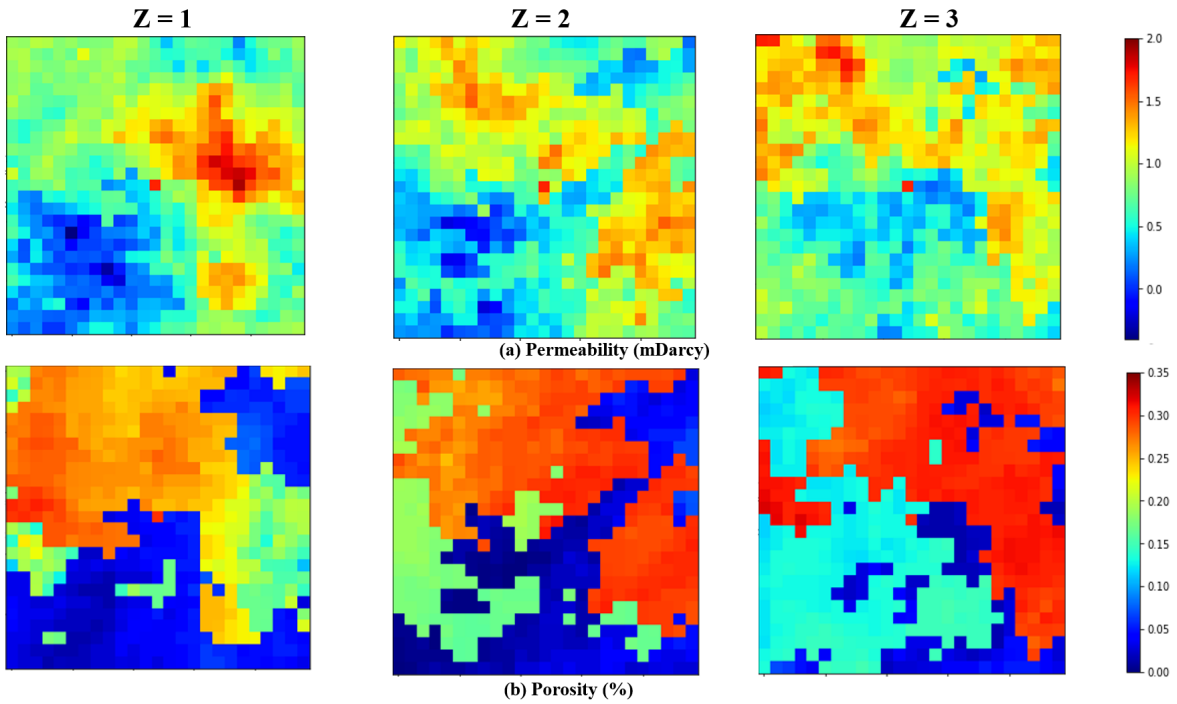

Figure 2.1. The (a) permeability and (b) porosity distribution (log scale) over 25*25 grid across z axes. Data shown corresponds to P10 geological conditions.

The input to the model is given as:

1. Time (t): Each of the 27 realizations have 72 time steps. These intervals are equally spaced and the values are normalized in the range of $[0, 1]$ by dividing with the maximum time step value.

2. Position (x, y, z): The 3D spatial coordinates are defined by $x \in [1, 25], y \in [1, 25]$ and $z \in [1, 3]$. These values are also normalized between $[0, 1]$ by dividing with the upper bound value for each of the dimensions.

3. Permeability K(x, y, z): The permeability at a given (x, y, z) coordinate is passed as input to the model. These values are normalized between 0 and 1 by dividing with the maximum permeability value in the dataset.

4. Porosity ϕ(x, y, z): Porosity is correlated with permeability and is heterogeneous. These values are also normalized between 0 and 1 using the maximum porosity value.

   Figure 2.1 denotes the permeability and porosity distribution (log scale) over the 25x25 grid corresponding to the P10 geological conditions. Appendix A depicts the permeability and porosity distributions for the P50 and P90 geological conditions.

5. Injection rate q_g: The CO_2 injection rate at well location (x_g, y_g, z_g) and time t is given as an input to the model. These values are normalized between 0 and 1 with respect to the maximum injection rate value in the dataset.

The output variables are:

1. Pressure $p(x, y, z, t)$: Pressure evolves with respect to space and time. These values are normalized between 0 and 1 by dividing with the maximum pressure value.

2. Gas saturation $S_g(x, y, z, t)$: CO_2 saturation is also a function of both time and space. The saturation values are already within the $[0, 1]$ range and are not normalized.

3. Water production rate $q_w(t)$: The water production rate is only a function of time because the well position (x_w, y_w, z_w) is fixed. These values are also normalized between 0 and 1. The water production rate is predicted at the well-location (irrespective of the input locations) by the model because it is zero elsewhere.

Appendix B depicts the evolution of the output variables across space and time for the three different geological conditions.

2.2 Simplified two-phase flow equations

A set of two-phase flow equations are used as physical constraints in the loss function. Under the assumption that CO_2 and brine are separate immiscible phases, mass balance coupled with Darcy's law gives the following partial differential equations [10]:

$$-\phi\frac{\partial}{\partial t}(S_g) - \nabla \cdot \left(\frac{1-S_g}{\mu_w}K(\nabla p - \rho_w g)\right) - q_w = 0 \tag{2.1}$$

$$\phi\frac{\partial}{\partial t}(S_g) - \nabla \cdot \left(\frac{S_g}{\mu_g}K(\nabla p - \rho_g g)\right) - q_g = 0 \tag{2.2}$$

Here, μ_w is the viscosity of water, μ_g is the viscosity of gas, ρ_w is the mass density of water and ρ_g is the mass density of CO_2. ∂t is the gradient with respect to time t, ∇ is the gradient with respect to the spatial coordinates x, y, z and $\cdot$ is the inner product.

These equations are derived assuming that the fluid properties like density and viscosity are constants and the formation is isotropic [1]. The parameters which are not model inputs or outputs such as the viscosity and mass density of water and CO_2 are treated as trainable parameters in the model. This ensures that there are no inconsistencies in the units of these parameters.

Chapter 3

Methods

First, two data-driven models (single-output LSTM and multi-output LSTM) are developed as the baselines. The physical laws given in Equations 2.1 and 2.2 are incorporated into the model as part of the physics-informed deep learning approach. The physics-informed LSTM is trained on the original supervised dataset and later on an unsupervised dataset with interpolated points. The performance of these models in predicting the CO_2 saturation, pressure and water production rate variables is compared.

LSTM is a deep learning time series model that carries over the inference between states [1]. The prediction at time t depends on the prediction at time $t-1$. Therefore, LSTM is an ideal choice for this dataset because the output variables vary with respect to space and time, as described in Section 2.1.

3.1 Single-output LSTM

For the single-output LSTM model, a separate LSTM is trained for each of the output variables (p, S_g, q_w) using MSE (Mean Squared Error) as the loss function. The input to the LSTM at each time step is a history of features at the previous three timesteps $t-3, t-2, t-1$ along with the features at the current timestep t. The earliest input is replicated at the beginning of the time series where a full history of features at the previous timesteps is not available [1]. The architecture to predict the pressure at time t (p_t) is given in Figure 3.1. The features at each timestep t are the permeability K_t, porosity ϕ_t, gas injection rate q_{g_t}, time t and the spatial coordinates x_t, y_t, z_t. The network architecture comprises of an input layer, two LSTM layers, three fully connected layers and one output layer with the ReLU activation function [1]. Each LSTM unit controls

the dependency of features between the successive timesteps. The same architecture is used for predicting each of the CO_2 saturation and water production rate variables.

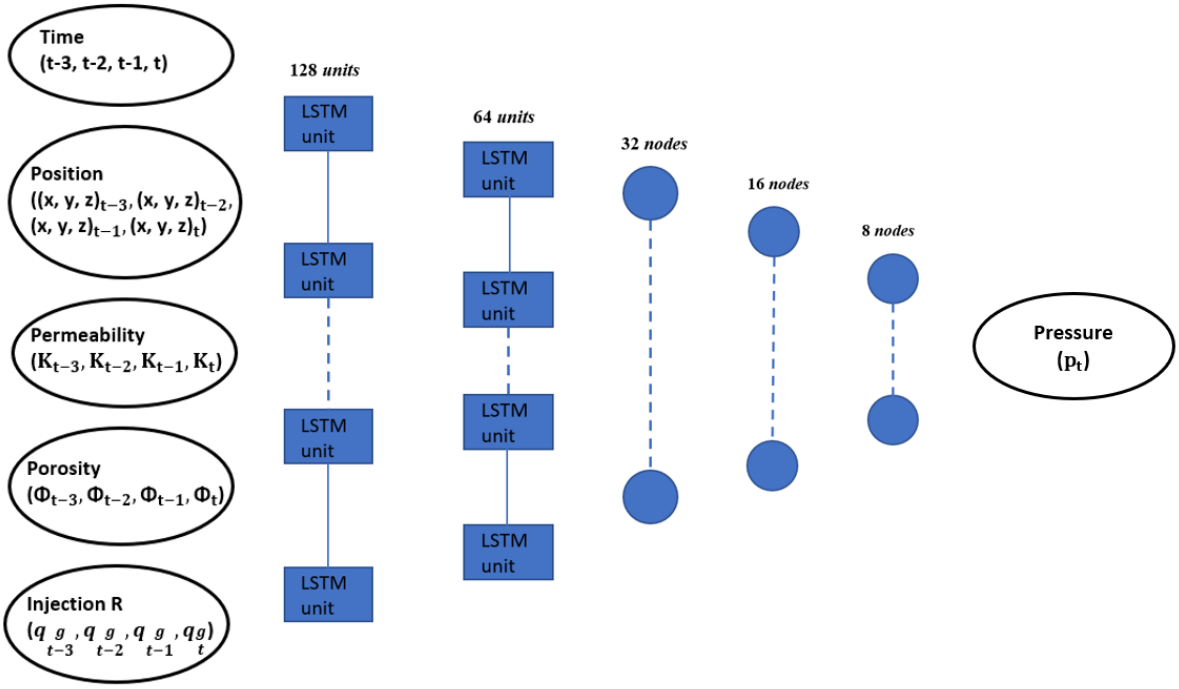

Figure 3.1. LSTM Architecture for prediction of pressure. The number of nodes in each layer is denoted above.

This architecture is decided after experimenting with a number of LSTM (3-layer, 2-layer and 1-layer variant) and Dense layers (2-layer, 3-layer and 4-layer variant). The length of the history window is also determined after examining several choices. The choices include a history of length 2 (i.e, at time t, the input included features at times t and $t-1$), length 3, and length 4. The history of length 4 is chosen according to the lowest validation loss [1]. Additionally, the model is tested using the ReLU activation function and the sigmoid activation function. ReLU is finalized because of its faster convergence.

The data-driven model is trained with a batch size of 250 for 500 epochs using Adam optimizer with a $1e-4$ learning rate. These hyperparameters are tuned according to the least cross validation loss over the training data. The Adam optimizer is used because of its adaptive learning rate for each parameter of the model. The best model, which is used for testing, is taken from the checkpoints saved according to the validation loss [1].

3.2 Multi-output LSTM

For the multi-output LSTM model, the LSTM is jointly trained for predicting all the three output variables (p, S_g, q_w) at once using MSE as the loss function. The input features and model architecture is similar to the one described earlier. The only difference is that the multi-output model has three nodes in the output layer. This is illustrated in Figure 3.2.

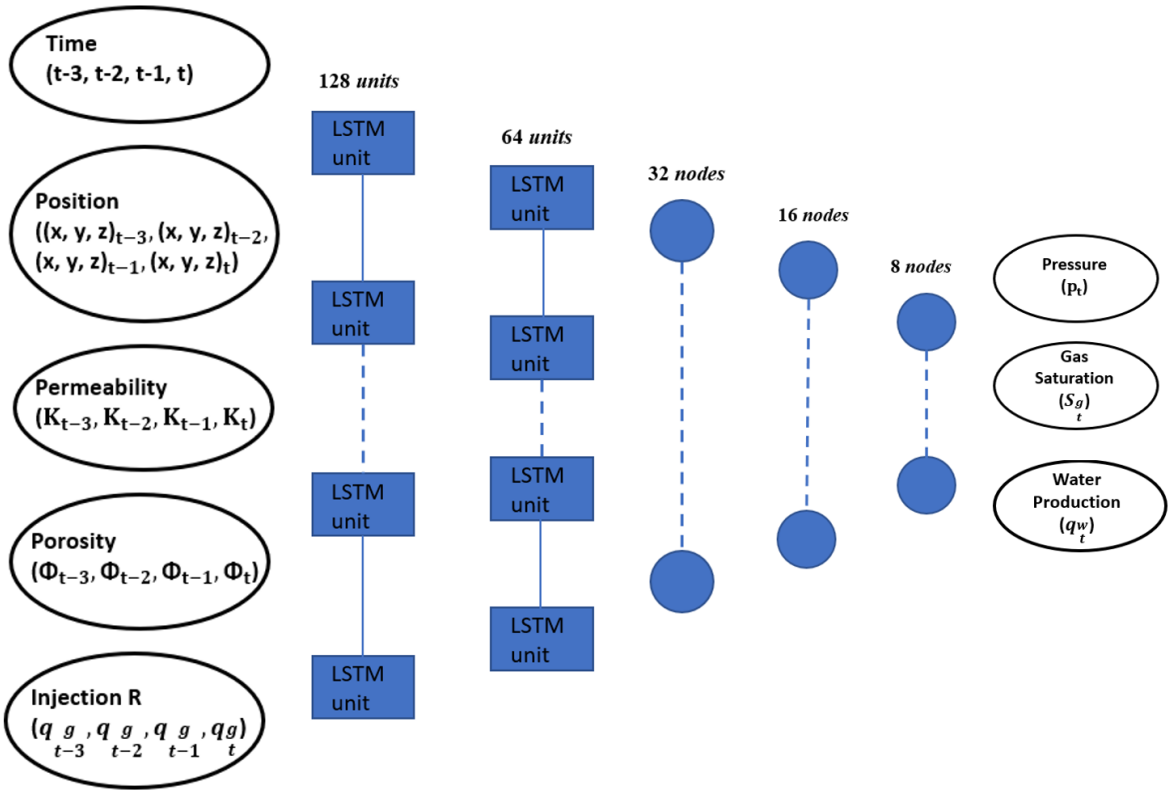

Figure 3.2. LSTM Architecture for joint prediction of pressure, gas saturation and water production rate. The number of nodes in each layer is denoted above.

This model is trained using Adam optimizer with a learning rate of 1e-4 for batches of size 250 each for 500 epochs. The model with the lowest validation loss is used as the final model for testing [1].

3.3 Physics-informed LSTM without interpolated points

The physics-inspired neural network (PINN) approach [7] [8] is followed for developing the physics-informed LSTM. The constraints corresponding to the physical laws are

added to the loss function to act as penalties when the governing equations are violated.

For instance, let $\widehat{p}$ be the neural network approximation for the true gas pressure p. Similarly, let $\widehat{S_g}$ and $\widehat{q_w}$ be the approximations for the true gas saturation S_g and true water production rate q_w respectively. Since the water production rate is predicted only at the well locations, it is 0 at the non-well locations. Therefore, $\widehat{q_w}$ is the model's prediction only at the well locations and is 0 otherwise. The above approximations are represented as:

$$\begin{aligned} p(x,y,z,t) &\simeq \widehat{p}(x,y,z,t,K,\phi,q_g;\theta) \\ S_g(x,y,z,t) &\simeq \widehat{S}_g(x,y,z,t,K,\phi,q_g;\theta) \\ q_w(x,y,z,t) &\simeq \widehat{q}_w(x,y,z,t,K,\phi,q_g;\theta) \end{aligned} \tag{3.1}$$

Now, let the input $u = (x, y, z, t, K, \phi, q_g)$. By inserting these neural network approximations into the simplified governing physical equations (Section 2.2), the functions $\widehat{f_1}(u)$ and $\widehat{f_2}(u)$ are given as:

$$\widehat{f}_1(u;\theta) = -\phi\frac{\partial}{\partial t}\widehat{S}_g(u;\theta) - \nabla\cdot\left(\frac{1-\widehat{S}_g(u;\theta)}{\mu_w}K(\nabla\widehat{p}(u;\theta)-\rho_w g)\right) - \widehat{q}_w(u;\theta) \tag{3.2}$$

$$\widehat{f}_2(u;\theta) = \phi\frac{\partial}{\partial t}\widehat{S}_g(u;\theta) - \nabla\cdot\left(\frac{\widehat{S}_g(u;\theta)}{\mu_g}K(\nabla\widehat{p}(u;\theta)-\rho_g g)\right) - q_g \tag{3.3}$$

Incorporating these terms into the loss function $L(\theta)$ of the PINN, it is therefore given by:

$$\begin{aligned} L(\theta) =& \frac{1}{N}\sum_{i=1}^{N}(\widehat{p}(u_i;\theta)-p(u_i))^2 + \frac{1}{N}\sum_{i=1}^{N}(\widehat{S}_g(u_i;\theta)-S_g(u_i))^2 \\ &+ \frac{1}{N}\sum_{i=1}^{N}(\widehat{q}_w(u_i;\theta)-q_w(u_i))^2 \\ &+ \frac{1}{N}\sum_{j=1}^{N}(\widehat{f}_1(u'_j;\theta))^2 + \frac{1}{N}\sum_{j=1}^{N}(\widehat{f}_2(u'_j;\theta))^2 \end{aligned} \tag{3.4}$$

Here, the first three terms are the standard MSE used to train the data-driven model for each of the pressure, gas saturation and water production rate variables and the last

two terms are penalties on violating the governing physics equations. N refers to the total number of samples in the dataset for which the true values of p, S_g and q_w are known (supervised dataset). Even though a history of features is passed as input to the LSTM, the gradient of the corresponding $\widehat{S_g}$ and $\widehat{p}$ functions in Equations 3.2 and 3.3 are computed only with respect to the current timestep t and current spatial coordinates x, y and z. Additionally, the network is unrolled to compute the second order gradients with respect to each layer in the LSTM [1].

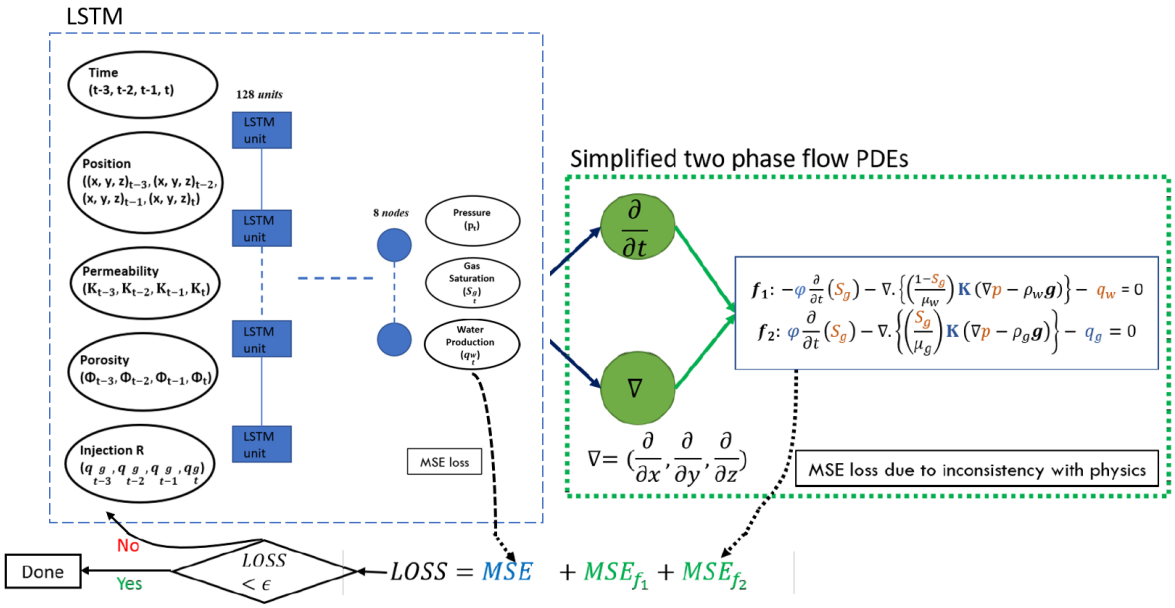

Figure 3.3. LSTM Architecture for joint prediction of pressure, gas saturation and water production rate incorporating physics constraints. The loss function is illustrated above.

Figure 3.3 depicts the LSTM architecture incorporating the governing physical equations. The trainable parameters in the equations take care of any problems associated with passing the normalized input features to the model. The gas injection rate and water production rate are passed to the physical laws only at the well locations. Since the well locations constitute a minority of the dataset, the LSTM may not see enough of these samples to learn the physics for nonzero q_g and q_w [1]. Oversampling is used to tackle this issue. This approach involves resampling the less frequent samples at the well locations to increase their proportion in the dataset with respect to the majority samples at the non-well locations.

This physics-informed model is trained using an Adam optimizer of learning rate 1e-5

with a batch size of 250 for 300 epochs. A batch should optimally contain 60% samples at the non-well locations for better convergence. The ratio of the non-well samples to the well samples in a batch is determined after experimenting with the respective 80:20, 70:30, 60:40 and 50:50 splits.

3.4 Physics-informed LSTM with interpolated points

Since the last two terms in Equation 3.4 are simply penalties on violating the governing equations, they do not require knowledge of the true p, S_g and q_w variables. Therefore, these functions can be evaluated over a set of arbitrary data points (unsupervised data) $\{u'_1, u'_2, \ldots, u'_M\}$. Here, M refers to the total number of points in the unsupervised dataset.

There are many possibilities for constructing the unsupervised dataset. A simple choice is to set $M = N$ and $u'_i = u_i$ for all i. This refers to using the same data as in the supervised dataset. However, it is effective to incorporate additional points such that $M > N$ when dealing with a limited training dataset [1]. One alternative approach to create the unsupervised dataset is by using equally-spaced linear interpolation of the spatial and time coordinates from the supervised dataset. Hence, the loss function for the unsupervised dataset only contains terms corresponding to the PDEs because there is no ground truth available.

Considering that the supervised data is given at equally spaced spatial and time coordinates, let a parameter c denote the number of unsupervised points required to be generated between any two neighboring position or time values. Accordingly, the interpolated timestep $\widehat{t_i}$ is given by:

$$\hat{t}_i = t_1 + \frac{t_2 - t_1}{c + 1} * i \tag{3.5}$$

Here, t_1 and t_2 denote two neighboring timestep values in the supervised dataset with $i \in [1, c-1]$. Similarly, c is used to generate new interpolated position values ($\widehat{x_i}, \widehat{y_i}$ and $\widehat{z_i}$). From experiments, it is determined that $c = 2$ is the optimal choice for generating the interpolated dataset.

The physics-informed LSTM with interpolated points is obtained by training the LSTM alternatively on the supervised and unsupervised dataset. The physics-based LSTM (Figure 3.3) is first trained on the supervised dataset until it achieves the lowest

validation loss. Later, the model is optimized using the interpolated dataset (including the supervised datapoints) using the physics MSE loss with a learning rate of 1e-5 for an additional 100 epochs until a saturation in the validation loss is noted.

Chapter 4

Results

The MSE values obtained for the different LSTM models are:

Model	STT	MTT	PI w/o interp.	PI with interp.
q_w	0.438	$9.249 * 10^{-2}$	$1.23 * 10^{-3}$	$4.255 * 10^{-4}$
p	$4.272 * 10^{-2}$	$1.533 * 10^{-2}$	$2.024 * 10^{-2}$	$1.063 * 10^{-7}$
S_g	$5.405 * 10^{-5}$	$1.09 * 10^{-5}$	$1.068 * 10^{-5}$	$4.175 * 10^{-6}$

Table 4.1. MSE values (original un-normalized scale) for STT (Single-output LSTM), MTT (Multi-output LSTM), PI w/o interp. (Physics-informed LSTM without interpolated points) and PI with interp. (Physics-informed LSTM with interpolated points) predicting water production rate (q_w), pressure (p), and gas saturation (S_g) variables.

The multi-output LSTM performs better than the single-output model in predicting all the three output variables. Thus, it is inferred that the multi-output model assists in modeling the inter dependency between p, S_g and q_w. Further, the performance is improved when the governing physical equations are added as constraints to the loss function. This is evident by observing the performance of the physics-informed LSTM without interpolated points with respect to the data driven models. A significant improvement in predicting all the three output variables is noticed because of the improved supervision provided by the equations. Eventually, the physics-informed model with interpolated points gives the lowest MSE values for all the three outputs. These results indicate that training with interpolated space-time coordinates is constructive in enhancing the performance of the model.

Figure 4.1 depicts the comparison between the ground truth and predicted values for the pressure and gas saturation variables for the data driven and physics-informed models. It is observed that the relative error decreases gradually as physics is incorporated into

the deep learning architecture.

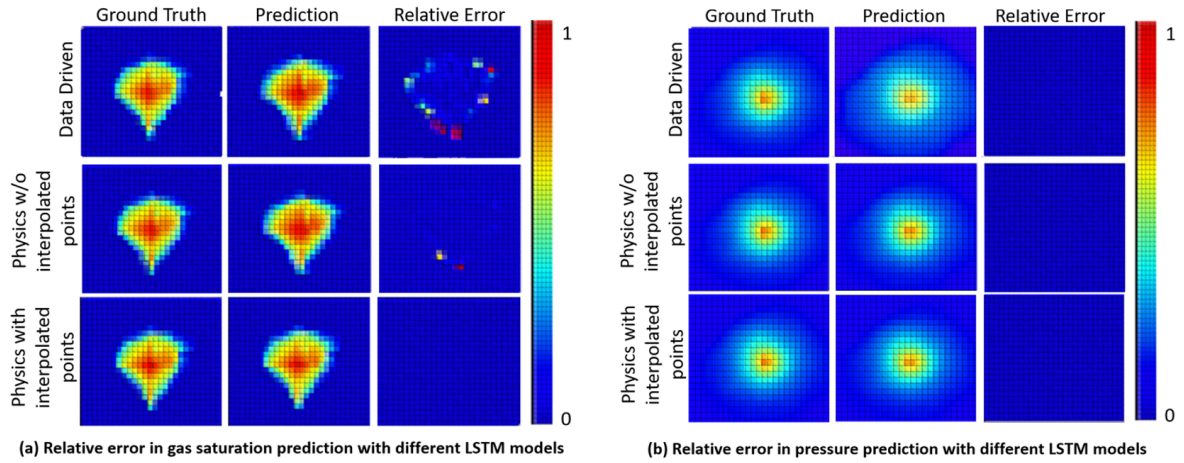

Figure 4.1. A comparison between the ground truth and predictions along with the corresponding relative error for pressure and gas saturation variables over 25*25 grid across z=1 axis. Data corresponds to P90 geological conditions. The figure displays the results of the multi-output LSTM (top), physics-informed LSTM without interpolated points (middle) and physics-informed LSTM with interpolated points (bottom).

Figure 4.2 shows the difference in the prediction of q_w by the physics-informed and data-driven models with respect to the ground truth for all the 72 timesteps. It can be analyzed that the deviation from the ground truth for both the $P10$ and $P90$ realizations is significantly reduced for the physics-informed models.

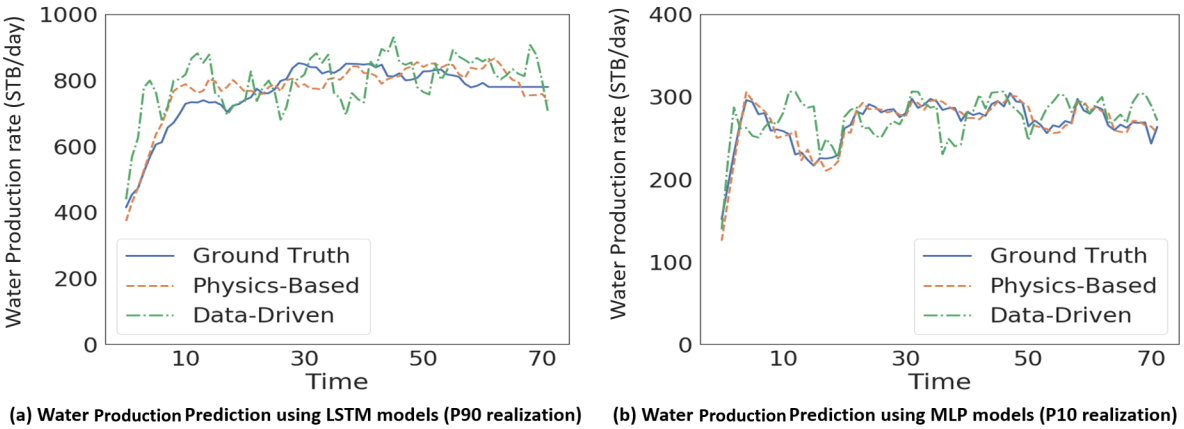

Figure 4.2. The predicted water production rate using multi-output data-driven LSTM and physics-informed LSTM without interpolated points compared to the ground truth.

Figure 4.3 displays the qualitative comparison of the predicted gas saturation variable using the physics-informed LSTM model for P90 geological conditions. The relative error

remains low across several timesteps. The plots between the relative error and each of the 72 timesteps imply that the error is stable throughout for both the gas saturation and pressure variables.

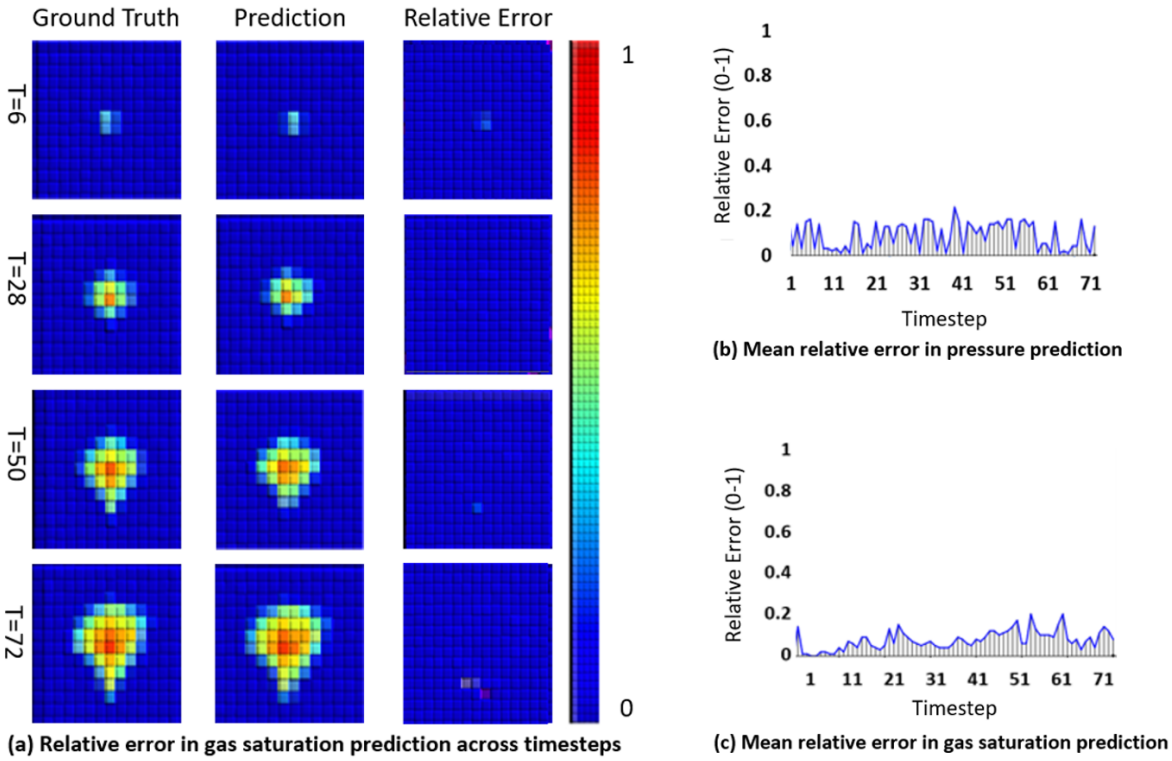

Figure 4.3. (a) A comparison between the ground truth and predictions along with the corresponding relative error for gas saturation variable over 25*25 grid across several timesteps and z=1 axis. Data corresponds to P90 geological conditions. The graphs in (b) and (c) plot the mean relative error for pressure and gas saturation variables with respect to all 72 timesteps.

Therefore, the physics-informed LSTM models accurately predict the p, S_g and q_w variables across space and time. The relative error between the ground truth and predicted variables remains low in comparison with the data driven LSTM models. These results demonstrate that adding physical laws as constraints to the cost function enhances the model performance.

Chapter 5

Conclusions

The data-driven and physics-informed LSTM models are developed to predict the evolution of pressure, CO_2 saturation and water production rate with respect to space and time for a carbon storage site. The LSTM model uses input history and carries over a state, proving beneficial for the simulated dataset. The data-driven models perform comparably with the multi-output model generating slightly better predictions than the single-output model. On the other hand, the physics-informed models perform significantly better than the data-driven models. The main finding is that incorporating a set of simplified flow equations in the loss function of a deep learning model provides much accurate predictions. Further, training the model on an interpolated dataset improves the performance of the physics-informed model.

The LSTM was trained on a small synthetic 'toy dataset', which is comparatively smaller than the realistic carbon site datasets. The feasibility of this proposed approach is to be explored on such larger datasets in the future. One shortcoming is that such physics-informed models take a considerably large amount of training time because of the modified network architecture. This can be addressed by techniques like feature reduction, using multiple GPUs and optimizing the data loading process when dealing with large datasets [1]. However, even though the training times are longer, the testing time is largely unaffected because it involves computations with pre-assigned model weights. Therefore, the advantages of having an accurate model may outweigh the overhead training times depending on the application. In conclusion, the physics-informed deep learning approach can be employed in CO_2 storage managements to effectively estimate the site response to CO_2 injection.

Appendix A
Permeability/Porosity Realizations

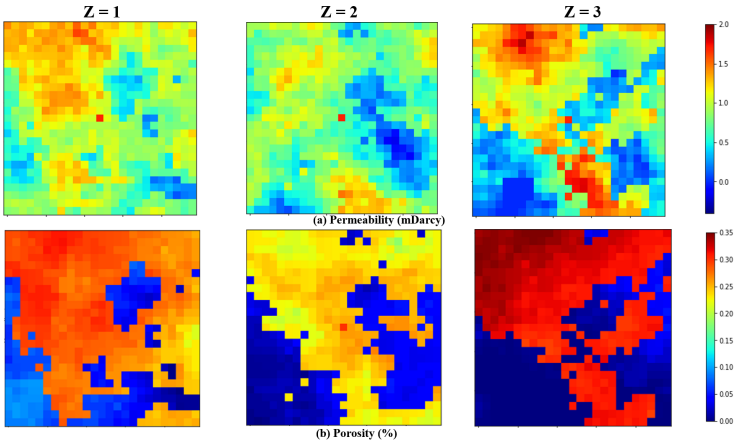

Figure A.1. The (a) permeability and (b) porosity distribution (log scale) over 25*25 grid across z axes. Data shown corresponds to P50 geological conditions.

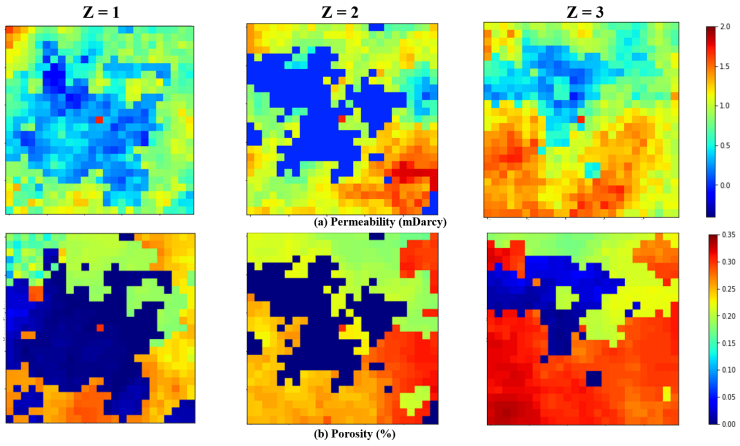

Figure A.2. The (a) permeability and (b) porosity distribution (log scale) over 25*25 grid across z axes. Data shown corresponds to P90 geological conditions.

Appendix B
Output Variables Visualization

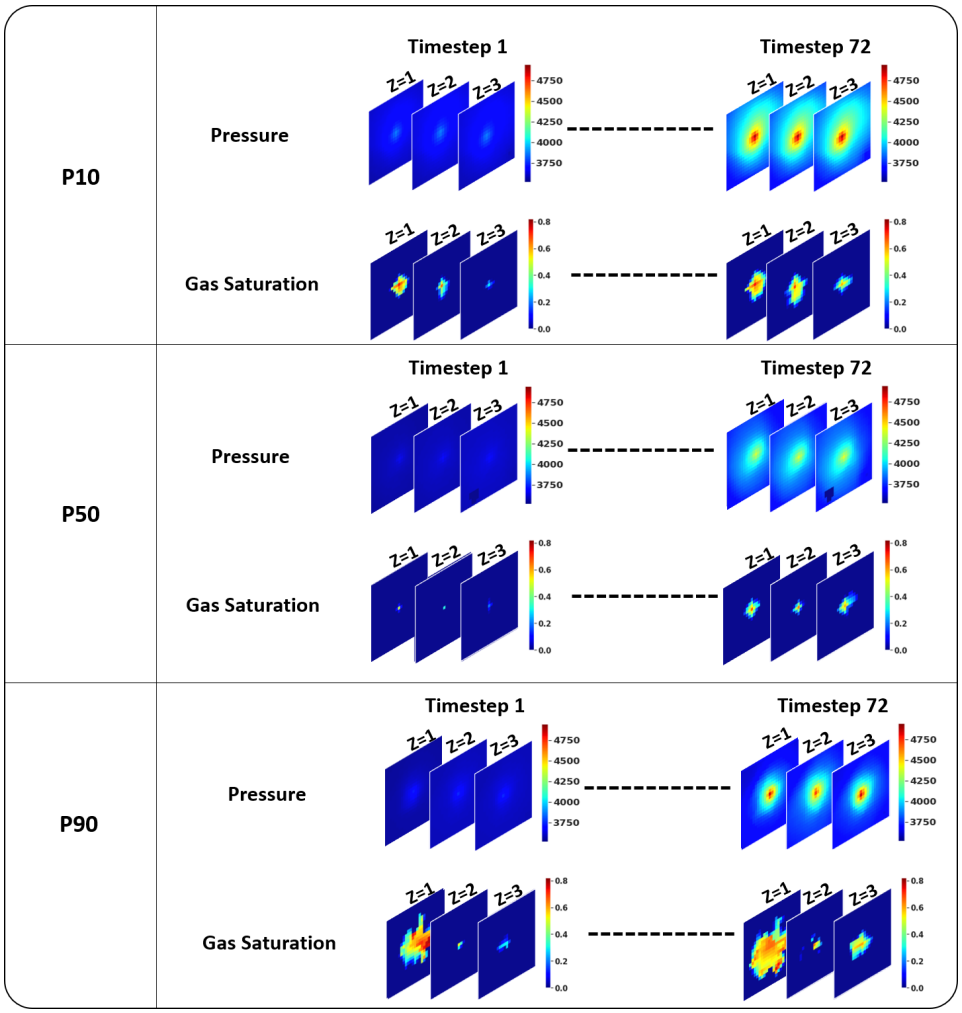

Figure B.1. The evolution of pressure and gas saturation fields over 25*25 field across several timesteps and z axes. Data corresponds to P10, P50 and P90 geological conditions.

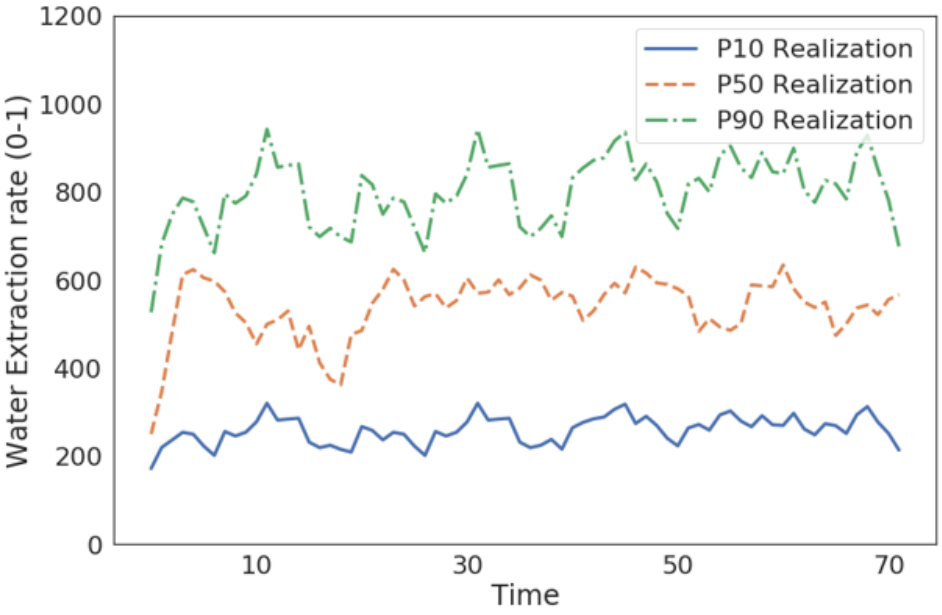

Figure B.2. The evolution of water production rate across the 72 timesteps. Data corresponds to P10, P50, and P90 realizations.

Bibliography

[1] Shokouhi, P., V. Kumar, S. Prathipati, S. A. Hosseini, C. L. Giles, and D. Kifer (2021) "Physics-informed deep learning for prediction of CO2 storage site response," *Journal of Contaminant Hydrology*, **241**, p. 103835.
URL https://www.sciencedirect.com/science/article/pii/S0169772221000747

[2] Zhong, Z., A. Y. Sun, and H. Jeong (2019) "Predicting co2 plume migration in heterogeneous formations using conditional deep convolutional generative adversarial network," *Water Resources Research*, **55**(7), pp. 5830–5851.

[3] Jeong, H., A. Y. Sun, J. Lee, and B. Min (2018) "A learning-based data-driven forecast approach for predicting future reservoir performance," *Advances in Water Resources*, **118**, pp. 95–109.

[4] Pawar, R. J., T. Chen, and Y. Lin (2019) *LANL ML Applications Overview*, *Tech. rep.*, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[5] Ganesh, P. R. and S. Mishra (2016) "Simplified physics model of CO2 plume extent in stratified aquifer-caprock systems," *Greenhouse Gases: Science and Technology*, **6**(1), pp. 70–82.

[6] Raissi, M., P. Perdikaris, and G. E. Karniadakis (2017) "Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations," *arXiv preprint arXiv:1711.10561*.

[7] Raissi, M., P. Perdikaris, and G. E. Karniadakis (2019) "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, **378**, pp. 686–707.

[8] Tartakovsky, A. M., C. O. Marrero, P. Perdikaris, G. D. Tartakovsky, and D. Barajas-Solano (2020) "Physics-Informed Deep Neural Networks for Learning Parameters and Constitutive Relationships in Subsurface Flow Problems," *Water Resources Research*, **56**(5).

[9] Yang, Y., J. Dong, X. Sun, E. Lima, Q. Mu, and X. Wang (2018) "A CFCC-LSTM Model for Sea Surface Temperature Prediction," *IEEE Geoscience and Remote Sensing Letters*, **15**(2), pp. 207–211.

[10] Ebigbo, A., H. Class, and R. Helmig (2007) "CO 2 leakage through an abandoned well: problem-oriented benchmarks," *Computational Geosciences*, **11**(2), pp. 103–115.