SPE-212821-MS

Reducing Simulation Time in a Huff-And-Puff Gas Injection Project in Complex Shale Reservoirs: Sequence-Based Proxy Multi-Porosity Reservoir Simulator

Cristhian Aranguren, University of Calgary; Carlos Rodríguez Araque, SierraCol Energy; Santiago Cuervo, Consultant; Alfonso Fragoso and Roberto Aguilera, University of Calgary

This paper was prepared for presentation at the SPE Canadian Energy Technology Conference and Exhibition held in Calgary, Alberta, Canada, 15-16 March 2023.

Abstract

The objective of this project is to explore cutting-edge sequence-based machine learning models commonly used in language processing to reproduce a multi-porosity reservoir simulator. The proposed method integrates advanced techniques to significantly reduce the numerical simulation time and improve the decision-making process for Huff and Puff (H-n-P) gas injection optimization in shale reservoirs. The proposed approach follows three crucial steps to predict an output sequence given an input sequence: 1) the simulation results should be validated against actual data, 2) train and validate a machine learning model using simulation results from either commercial or in-house numerical simulators, 3) exhaustive exploration of hyperparameter tuning and selection of machine learning techniques, such as sequence-to-sequence (Seq2Seq), Luong attention and ConvLSTM. The proxy model considers as input variables well control parameters such as injection and production periods, number of cycles and gas injection rates to estimate the proxy model results.

The multi-porosity proxy reservoir simulation model is a complementary tool that integrates numerical simulation and data-driven techniques. Although tuning the model typically demands significant time, it can speed up the simulation time up to 20,000X allowing for generating hundreds or even thousands of scenarios at the expense of accepting a reduction in the accuracy of the results in a matter of minutes. One of the most notable findings is that considering a small training dataset, the proxy model can reproduce the capabilities for predicting oil production in complex low and ultra-low permeability reservoirs with significantly reduced error, relative to the multi-porosity reservoir simulator. Finally, the possibility of reproducing a considerable number of scenarios in minutes opens the door to exploring different well control configurations such as injection and production periods, number of cycles and gas injection rates. The novelty of the proxy multi-porosity reservoir simulator is to notably accelerate the numerical simulation time by using techniques capable of solving sequence learning problems in which the output is dependent on previous outputs.

Introduction

Predicting well performance and ultimate recovery is one of the critical goals of reservoir simulation to identify opportunities to increase the recovery factor. When dealing with low and ultra-low permeability reservoirs, the recovery factor is extremely low, and it can range from 5 to 10% of the OOIP. Therefore, improved and enhanced oil recovery (IOR and EOR) techniques have been investigated extensively, concluding that H-n-P gas injection is one of the most promissory methods based on laboratory (Wan et al., 2015; Yang & Li, 2021) and numerical simulation (Lopez Jimenez, 2017; Alharthy et al., 2018; Lopez and Aguilera, 2019; Fragoso et al., 2019). However, characterizing and modelling these types of reservoirs is particularly challenging since they are composed by a quintuple-porosity system (Lopez and Aguilera, 2015, 2018) made out of adsorbed-porosity (ϕ_{ads_c}), organic porosity (ϕ_{org}), inorganic porosity (ϕ_m), natural fracture porosity (ϕ_2), and hydraulic fracture porosity (ϕ_{hyd}). These porosities are essential inputs that must be included in a physics-based numerical simulation model for estimating oil and gas recoveries. A fully implicit 3D multiphase modified black-oil finite-difference numerical formulation for quintuple porosity shales, termed GFREE-SIM, was developed by Lopez and Aguilera (2019). However, solving the 3D multiphase quintuple porosity formulation takes a significant amount of running time. Thus, a key objective of this paper is to investigate strategies that can be implemented in order to reduce the simulation time. These strategies include (1) using data driven techniques that can reproduce similar behavior as the numerical simulator at the expense of a modest reduction in accuracy. The results can be adopted for faster decision-making, and (2) using parallel computing in order to carry out multiple calculations simultaneously. There is no need to use parallel computation for the problem presented in this study.

Integrating physics-based models and data-driven techniques has been a relevant approach to accelerating the computational time, especially when dealing with optimization problems. For instance, Sarma et al. (2017) introduced a state-of-the-art application, which combines predictive capabilities of traditional physics-based models with data analytics models. This combination provides important advantages such as short development time, low development cost, and fast track analysis. In addition, numerous approaches to reproduce a digital twin for reservoir dynamics modeling have been proposed by several authors (Amini, 2014; Navratil et al., 2019; 2020; Chaki et al., 2020) using the well-known alternative of supervised learning, particularly employing deep learning and recurrent neural networks. Most of these applications use neural networks due to their ability to approximate any function through backpropagation and gradient descent. In contrast to recurrent neural networks, where each output is based on prior results, feed-forward neural networks are generally used. The latter, however, does not make physical sense because the output is independent of the previous outputs. On the other hand, treating these problems as sequence-based allows to predict a future value from a given input sequence.

As an example, Chaki et al. (2020) presented a proxy modeling comparison using deep neural networks (DNN) and recurrent neural networks (RNN) to minimize the computational cost associated with numerical simulation of a fully physics-based flow simulator. Chaki et al. (2020) concluded that DNN is faster but does not provide better quality results than RNN. A very important observation by these authors is that DNN can only predict within the interval for the time steps it was trained, whereas RNN can make future prediction without limitation. Similarly, Navratil et al. (2019; 2020), described an end-to-end deep surrogate model expanding on the sequence-to-sequence (Seq2Seq) approach considering as inputs (1) drilling actions, (2) distribution of rock properties in the reservoir and (3) well control and completions. Their effort achieved acceleration rates up to 40,000X with a considerable small margin of average error.

Hence, in the present study we consider as input and output a sequence of input information, and we deploy three architectures that integrate recurrent neural networks in an encoder-decoder format such as (1) Seq2Seq, (2) Luong Attention and (3) ConvLSTM. The proxy model is strongly dependent on the physics input in the numerical simulator, but it cannot be considered as a substitute of the fully implicit numerical model. When using the proposed proxy model, it is important ensure its quality through an exhaustive error analysis. Important contributions can be extracted from the proposed methodology, such as the application of multiple optimization methods including for example Reinforcement Learning.

Multi-porosity reservoir simulation

Shale reservoirs are complex rocks that store fluids in multiple porosity systems where fluid flow occurs through different mechanisms. Lopez (2017) and Lopez and Aguilera (2019) developed a fully implicit quintuple porosity formulation in GFREE-SIM to incorporate different storage mechanisms present in shales. These porosities are: (1) organic porosity, (2) inorganic porosity, (3) natural fracture porosity, (4) adsorbed porosity, and (5) hydraulic fracture porosity. In addition, GFREE-SIM considers gas dissolved in the solid kerogen.

GFREE-SIM is used for the physics-based part of the present study. GFREE-SIM is built based on the following assumptions and considerations (Lopez and Aguilera, 2019, Fragoso et al., 2019):

1. The shale reservoir has porosities in inorganic and organic matter, and in natural and hydraulic fractures.
2. There is viscous flow of gas and liquids in the fracture systems (natural and hydraulic fractures).
3. There is viscous flow of gas in both inorganic and organic matter.
4. There is viscous flow of liquids in both inorganic and organic matter.
5. Desorption occur only in wall surfaces of organic pores.
6. Desorption and diffusion of gas in the solid kerogen are included with the use of a desorption curve following the concept of 'gas evolution graphs' by Javadpour et al. (2007).
7. The gas dissolved in solid organic matter (solid kerogen) is handled by means of a fractional volume of solid kerogen (Vdiff) introduced by Lopez and Aguilera (2018).
8. There is modified black-oil modeling of the fluid phases (two hydrocarbon pseudo-components and one non-hydrocarbon component).
9. Porosity and permeability of fractures are stress dependent (Piedrahita et al., 2019).

The GFREE-SIM numerical formulations as well as several examples demonstrating the above flow mechanisms have been published by Lopez and Aguilera (2019). Reservoir and fluid properties are presented in Table 1 and Table 2 respectively.

Table 1—Reservoir parameters used in GFREE-SIM for validation of primary recovery and H&P gas injection. Initial simulation uses the dual porosity option. The dissolved gas in solid kerogen, adsorbed gas, and organic porosity are switched off in this case (data from Lopez and Aguilera, 2019).

Parameter	Symbol	Value	Units
Number of cells in x-direction	N_x	21	-
Number of cells in y-direction	N_y	11	-
Number of cells in z-direction	N_z	10	-
Reservoir length	N_x	250	ft
Reservoir width	N_y	550	ft
Reservoir thickness	h	250	ft
Formation top		8,250	ft
Matrix porosity	ϕ_m	0.07	fraction
Matrix permeability	K_m	0.001	md
Natural fracture porosity	ϕ_2	0.009	fraction
Natural fracture permeability	k_2	0.9	md

Parameter	Symbol	Value	Units
Hydraulic fracture permeability	$k_{h}f$	2,000	md
Hydraulic fracture width	$w_{h}f$	0.01	ft
Hydraulic fracture half-length	$x_{h}f$	275	ft
Skin factor	S	0	-

Table 2—PVT properties used in GFREE-SIM for validation of primary recovery and H&P gas injection modeling (data from Lopez and Aguilera, 2019).

Pressure	Gas Oil Ratio (Solution)	Oil Formation Volume Factor	Viscosity	Gas Formation Volume Factor	Gas viscosity
psi	**scf/STB**	**RB/STB**	**cP**	**RB/scf**	**cP**
14.7	1	1.062	1.040	0.166666	0.0080
264.7	90.5	1.150	0.975	0.012093	0.0096
514.7	180	1.207	0.910	0.006274	0.0112
1,014.7	371	1.295	0.830	0.003197	0.0140
2,014.7	636	1.435	0.695	0.001614	0.0189
2,514.7	775	1.500	0.641	0.001294	0.0208
3,014.7	930	1.565	0.594	0.001080	0.0228
4,014.7	1,270	1.695	0.510	0.000811	0.0268
5,014.7	1,618	1.827	0.449	0.000649	0.0309
9,014.7	1,618	1.500	0.600	0.000386	0.0470

In addition, a validation case using data from a H-n-P gas injection pilot horizontal well in the Eagle Ford Shale was published by Fragoso et al. (2019), obtaining a good level of certainty from the history match of real data (Figure 1). The pilot well produced under natural driving mechanisms for the first 29 months, Subsequently the well was subjected to seven natural gas injection cycles, which substantially increased the recovery factor by 1.84 times after approximately 80 months. Their study includes the effects of adsorption/desorption and diffusion of gas from solid kerogen, stress-dependent properties of the hydraulic fractures and positive skin factor stemming from precipitation and deposition of asphaltenes, which causes formation damage.

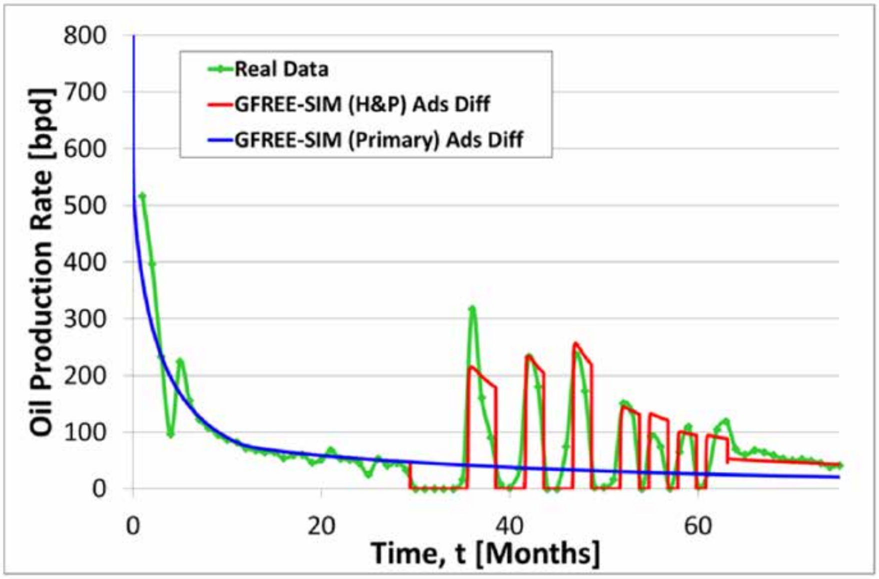

Figure 1—History match results of Eagle Ford pilot using GFREE-SIM. Ads and Diff indicate that the model includes adsorption and diffusion. Taken from Fragoso et al. (2019).

Proxy model

A proxy multi-porosity reservoir simulation model is proposed in this study utilizing deep learning techniques capable of approximating to the response of the physics-based model. The input a strategy considers variations in gas injection rate, injection period, production period and number of cycles during simulation of a H-n-P gas injection project. Several numerical simulation cases evaluating alternate injection/production schedules, skin factors and injection rates are performed. H-n-P results are collected and transformed to a regular daily time-series as shown in Figure 2. The positive skin factor is included as a discrete variable, which increases the formation damage due to asphaltene deposition.

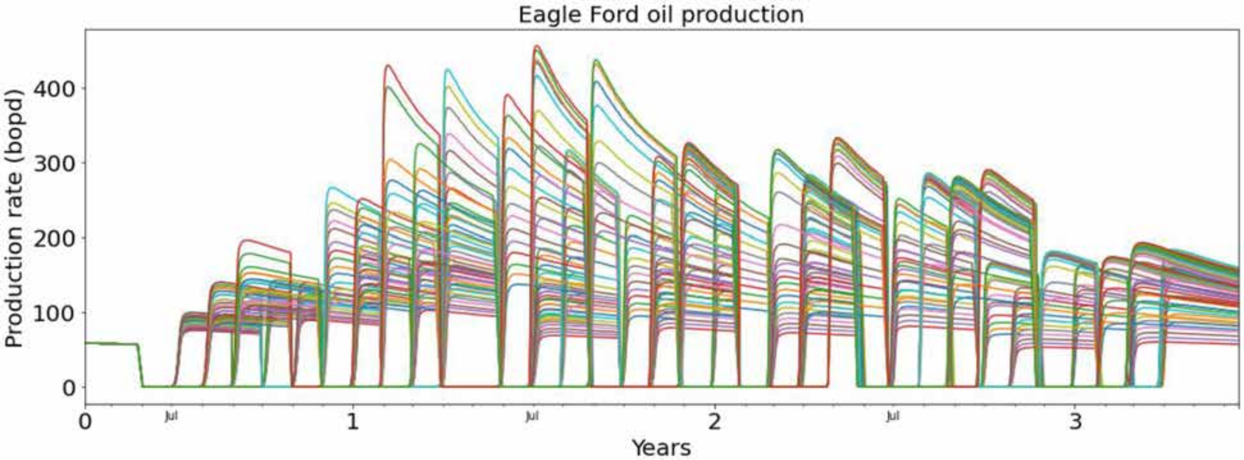

Figure 2—Example of daily oil production rate stemming from simulation results that consider different controls in the gas injection rate. The data are an example of the information used in the proposed Seq2Seq- based proxy multi-porosity reservoir simulator.

Data splitting

The number of simulations is specified by a combination of actions between the injection period and injection rate, in which lower and upper bounds are defined as well as variations within that range. The cases configurations are listed in Table 3. The model is trained with 80% of the total number of simulations runs, and 20% is removed randomly and selected for validation purposes as shown in Figure 3. The split configuration allows to run a more unbiased evaluation and to tune the hyperparameters, which is essential to assess the prediction performance on unseen data. Ideally, the proxy model is used to generate a thousand scenarios based on some given inputs, and it is crucial to make sure that the prediction obtained from the model approximate closely to the numerical simulation response making sure that the results are physically realistic.

Table 3—H-n-P gas injection scenarios.

Injection periods	Production periods	Cycles	Gas injection rate	Skin
1	2	12	Lower bound: 1 MMscfd Upper bound: 2.5 MMscfd One case every 100 Mscfd	5, 8, 20
1	3	9		
2	2	9		
2	3	7		
3	2	7		
3	3	6		

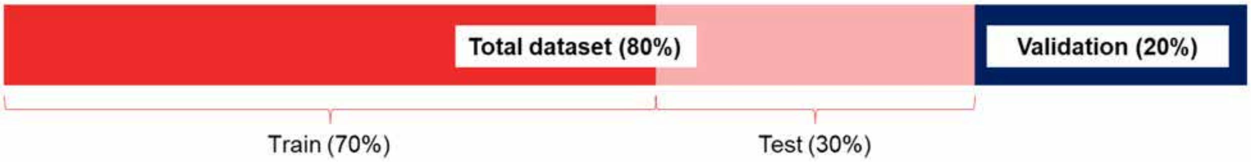

Figure 3—Split dataset showing training, testing and validation.

Data transformation

The proxy multi-porosity reservoir simulation model learns from a set of training simulations to replicate the output (oil production) of the numerical simulation model based on given input features. Defining and transforming the input variables needs to be done carefully in order to increase the efficiency of the data-driven model and to accomplish better predictivity performance. Initially, injection/production schedules, skin factor and injection rates are defined as inputs. Each of the input variables have a daily frequency, meaning that each observation is computed sequentially day by day. However, the results did not initially match to simulated scenarios. Therefore, data transformation strategies were implemented.

1. A case name is generated as follows: "Inj_" + "pro_" + "gas injection rate" + "number of cycles". For the case of the first configuration given in Table 3, the **"Code"** is "Inj1-pro2-1000000-12." The non-numerical labels are set to numerical variables using Label Encoder.
2. A new variable is created **"Actions",** which consists of string information following these considerations: If the well is producing, the variable is assigned to be "P" whereas if the well is injecting, the variable is given by "I" + gas rate, e.g., "I100000" referring to injecting 1 MMscfd. A Label Encoder transformer is used to convert categorical data into numerical labels. The new input features allow to include well controls.
3. It was noticed that the feature "gas injection rates" had a significant impact on the predicted results. Originally, a gas injection rate of zero was assigned in GFREE-SIM when the well was set for primary production. Subsequently, the value of the gas injection rate is given whenever the well starts injecting gas (huff) until it is opened for production (puff) when it is set to zero again. After several attempts, it was found beneficial for a better ML performance to alter the "gas injection rates" variable as a feature engineering operation by calculating the inverse procedure, e.g., when the well is producing, the rate of the injection is considered, while the gas injection rate is set to zero when the well is injecting.

Finally, this variable is named **"gas injection rates - inverse."** It means that future oil production will have a strong positive correlation with the gas rate that is injected before opening for production.

After running a sensitivity analysis, both categorical variables (**Code** and **Actions**) were eliminated as they did not have any significant impact on the proxy model. Thus, we decided to keep only five input variables: (1) injection period, (2) production period, (3) H-n-P cycles, (4) skin and (5) gas injection rate - inverse. The data is scaled using MinMaxScaler in the Python library Scikit-learn from a feature range between -1 and 1.

Truncation

The sequence of the input and output is divided (truncated) into smaller portions by sliding windows, and then these samples are stored in 3D arrays that contain the number of samples, timesteps and input variables. The output includes the number of samples to predict and the timesteps. According to our configuration, the first timestep, refers to $t = 1$ and the sequence of actions for prediction is as follows: $t = 1$, $t = 2$,…, $t = 30$ days of production. For the next sample, it slides one timestep towards the validation data, and will follow the same process from $t = 2$ and predicts as follows: $t = 2$, $t = 3$,…, $t = 31$.

Sequence-to-sequence model

The main objective is to find a recurrent neural network capable of predicting daily oil production based on a sequence of well control actions such as injection periods, production periods, number of cycles and gas injection rate. Initially, we employed a machine learning technique commonly used in language processing known as Seq2Seq following a similar process to the one shown in Aranguren et al. (2022), in which the model was used to predict future declining production rates. The Seq2Seq model has two main components: encoders and decoders, which are made of a series of LSTM units. The input receives the sequence of input information, rejects the output, and encloses all the information into a fixed-length context vector, which is finally transmitted to the decoder. The decoder produces a sequence of output data subjected to the information received from the context vector.

Aranguren et al., (2022) predict future oil rates by following a sequence series of historical data. For example, their model slides windows of the input data with a predetermined length of time steps (for example, 60 days oil rates) and an output of the next time steps in order to forecast the next step (e.g., 10 days oil rates). On the other hand, for the present study, the model takes the sequence of actions of one timestep (e.g., as injection periods, production periods, number of cycles and gas injection rate including the formation skin factor) to predict the oil production in the following 30 days. The target values during the training process are log transformed in order to handle negative unphysically realistic values using $Log(1 + x)$ whereas when predicting it needs to be converted back $exp(x) — 1$. Results obtained in the Seq2Seq model are shown in Figure 4 and compare the numerical simulation response in light green with the machine learning prediction in blue dashed line. The index is the position of an element indicating that all the actual (true) and predicted values for every scenario are collected in a single data frame. Every hump in Figure 4 specifies each scenario; for simplicity, we only display the first 21 cases. Figure 5 compares the cumulative production given by the numerical simulator versus output results from the Seq2Seq model using only the validation dataset, which is the ML responses on unseen data.

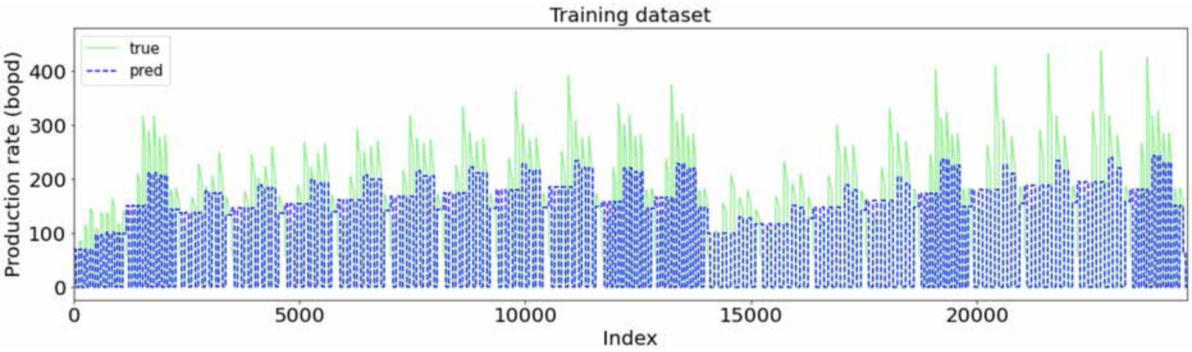

Figure 4—Predicted Seq2Seq daily oil production (training dataset) vs results from numerical simulation.

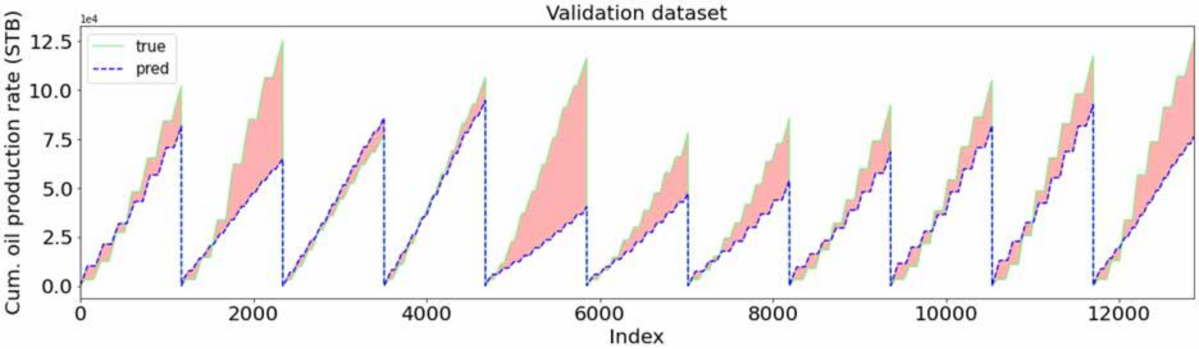

Figure 5—Predicted Seq2Seq cumulative oil production (validation dataset) vs results from numerical simulation. The light red color indicates the difference between them.

Luong Attention model

The fixed-length context vector's inability to contain all the necessary information when dealing with large input sequences is one of the Seq2Seq model shortcomings, which means the decoder will receive and make predictions on ineffective contextualization of the data. For that reason, mechanisms have been widely developed to tackle this problem by introducing special attention to particular input vectors of the input sequence assigning attention weights, which provide specific context information to the decoder. There are two types of attention mechanisms reported in the literature, such as Bahdanau Attention and Luong Attention mechanisms. The main difference between these two approaches is well described by Loye, (2019), who emphasizes that the models differ on how the alignment score is calculated and the position of the attention mechanisms presented in the decoder. For the present study, we implemented the application using the Luong Attention mechanism. Detailed information on this mechanism has been presented by Luong et al., (2015) and need not be repeated here. Similar to Seq2Seq shown in Figure 4, Figure 6 shows a representation of the training dataset cases contrasting the results from the Luong Attention and the numerical simulation results. Likewise, Figure 7 shows the cumulative oil production of some validation cases, highlighting in red the difference between the multi-porosity numerical simulator and the Luong Attention model.

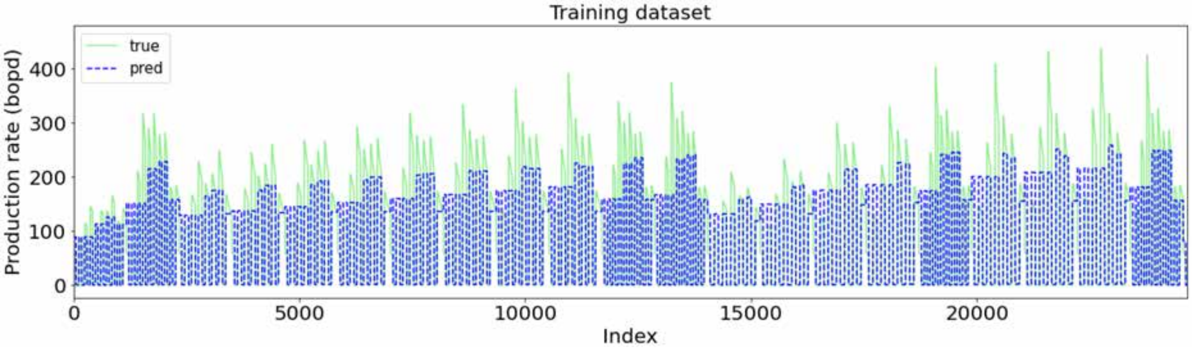

Figure 6—Predicted Luong Attention daily oil production (training dataset) vs results from numerical simulation.

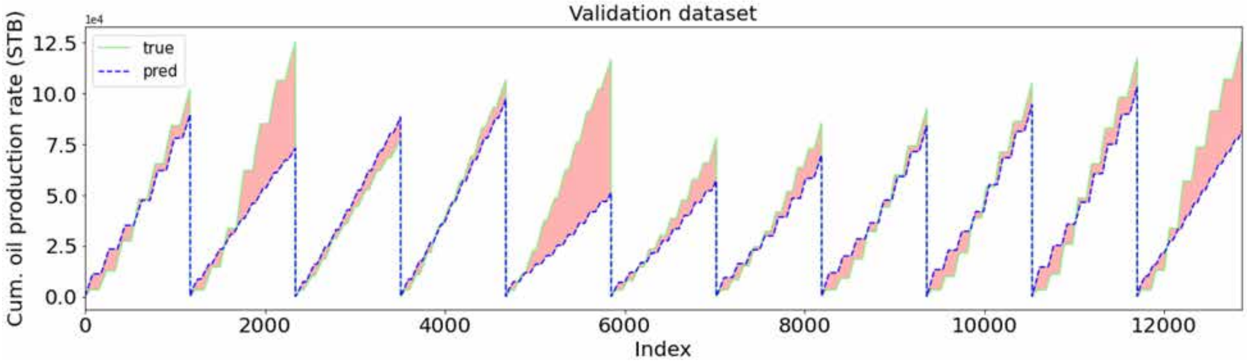

Figure 7—Predicted Luong Attention cumulative oil production (validation dataset) vs results from numerical simulation. The light red color indicates the difference between them.

Hyperparameter tuning and model selection

Figure 6 and Figure 7 show, at first glance, some improvement compared to the results obtained from the simpler Seq2Seq model shown in Figure 4 and Figure 5. We decided to put special emphasis on tuning hyperparameters such as the number of hidden LSTM units, dropout rate, and batch size during training. For this process, we run several experiments varying the previously mentioned H-n-P input parameters and identifying the impact of these variations on the Seq2Seq results. Figure 8 presents a summary containing 8 experiments, in which from 1 to 3 are the error measurements of the Seq2Seq model, and from 4 to 7 using the Luong Attention model. It is important to note that for each experiment, we considered the partition of the data in the training, testing and validation with the aim of evaluating the performance in each section. Some significant conclusions can be taken from the analysis:

1. Expanding the dropout size results in a significant increase in the training error while achieving relatively good results on the testing and validation datasets.
2. Increasing the number of LSTM units produces an overfitted model with poor results in the validation dataset. It proves the inability of the Seq2Seq model to deal with long input information. Undoubtedly, a lengthened context vector reduces the performance of the model.
3. From experiments 3 to 7, the idea is to demonstrate that even though reducing the batch size from 2,048 to 16 generates considerably lower error, it is at the expense of linear growth in the computation cost (training time).

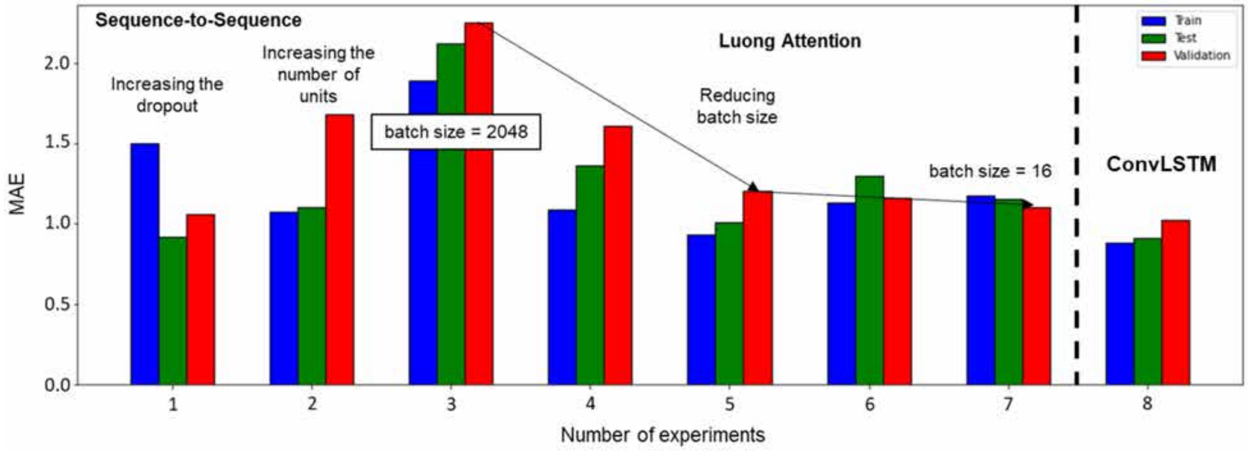

Figure 8—Model selection and hyperparameter tuning (scaled MAE).

After an extensive effort trying to find the right hyperparameters, we compared the training dataset (which is comprised of training and testing data) against results from the multi-porosity numerical simulator to provide a validation. Despite our attempts to validate the Luong Attention model vis-a-vis the multiporosity numerical simulator (GREE-SIM), results shown on Figure 9 still indicate that there is room for improvement, a topic that is addressed in the next section dealing with the **ConvLSTM model**. The black dashed line indicates the maximum cumulative oil production for the simulated case and is used as a reference in both training and validation.

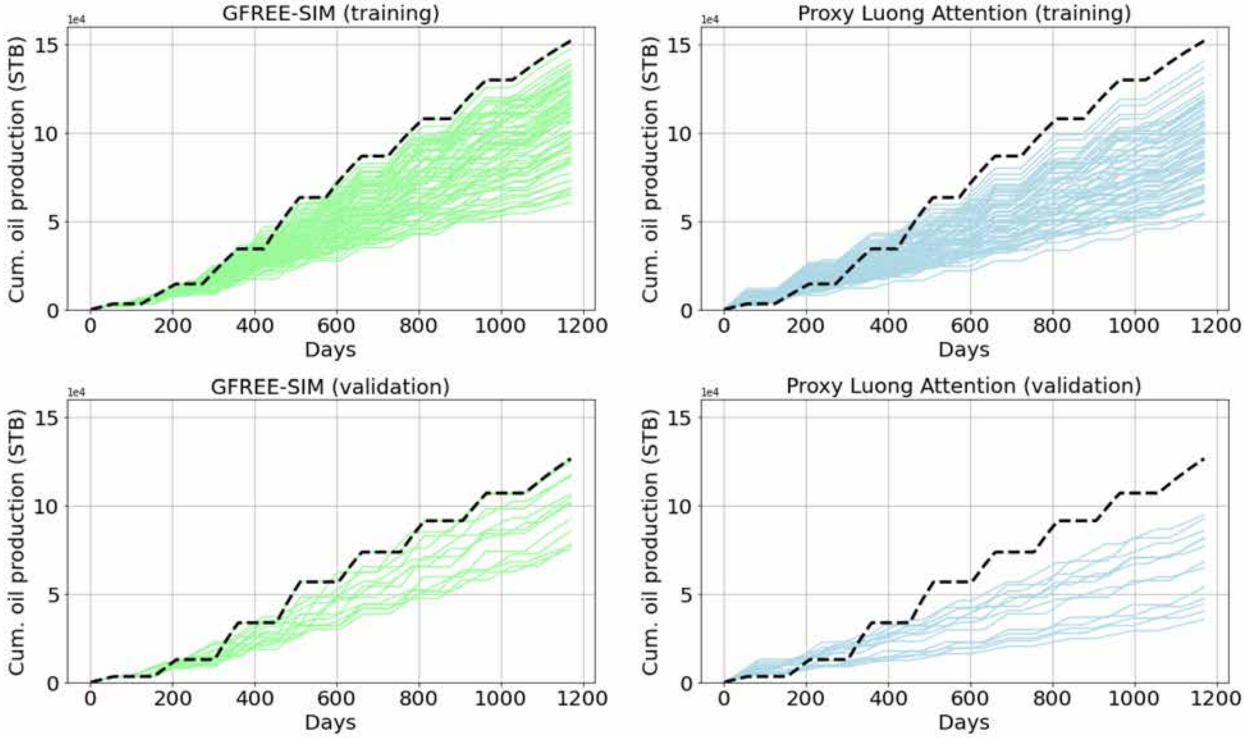

Figure 9—Cumulative oil production. Upper left-hand side: training dataset results from the multi-porosity numerical simulator (GFREE-SIM); upper right-hand side: training dataset results from the Luong model; lower left-hand side: validation dataset results from the multiporosity numerical simulator; lower right side: validation dataset results from the Luong model.

ConvLSTM model

To improve the proxy model's performance, we investigate the hybrid ConvLSTM model (Eq. 1) frequently used in spatio-temporal problems. Similar to the previous models, ConvLSTM is also a recurrent layer model. However, convolution operations are used instead of internal matrix multiplication. In (1, the symbol $\odot$ denotes the Hadamard product, * is the convolutional operator, *W, H, C,* and *b* represent the weight, hidden state, memory state, and bias respectively. Shi et al. (2015) first introduced the method by forming an encoding-forecasting problem through stacking multiple ConvLSTM layers as shown in Figure 10. Long term dependencies are addressed including the Constant Error Carousel (CEC), which instead of having a forget gate, includes an unchanged cell state that helps to solve the vanishing problem (Wang et al., 2015). As a result, ConvLSTM performs better when dealing with long-term dependencies as compared with the Seq2Seq and Luong Attention models.

$$
\begin{aligned}
i_t &= \sigma\left(W_{xi} * \chi_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \odot C_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_{xf} * \chi_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \odot C_{t-1} + b_f\right) \\
C_t &= f_t \circ C_{t-1} + i_t \odot \tanh\left(W_{xc} * \chi_t + W_{hc} * \mathcal{H}_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_{xo} * \chi_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \odot C_t + b_o\right) \\
\mathcal{H}_t &= o_t \odot \tanh\left(C_t\right)
\end{aligned}
\tag{1}
$$

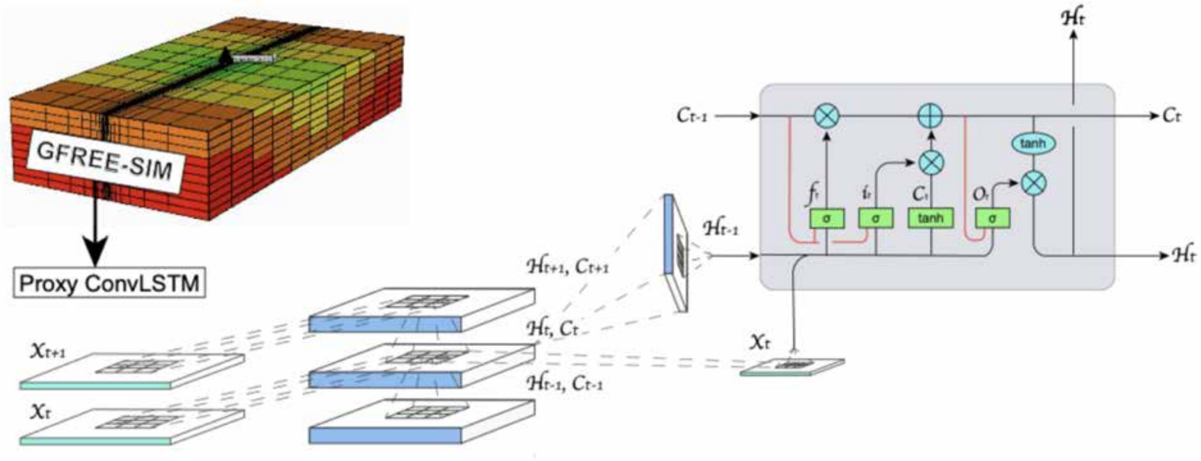

Figure 10—Multiple ConvLSTM architecture. Modified from Shi et al. (2015).

The main divergence of ConvLSTM with respect to the earlier models is in the input dimensions, which in this case have 5 dimensions: the number of samples, time steps, channels, rows, and columns. In order to understand the configuration, it is necessary to go back to the **truncation** section. The final shape obtained from the truncation for the input sequence has 3 dimensions: samples, timesteps and features and only 2 dimensions for the output: samples and timesteps. For instance, one timestep that follows a sequence of 5 variables predicts the subsequent 30 days of oil production. As stated before, the ConvLSTM requires 5 dimensions, which means that channels and rows are variables that need to be added and are equal to 1. Concerning hyperparameter tuning, several experiments were evaluated. Table 4 shows the final hyperparameter's configuration. Results from the training and validation datasets are presented in Figure 11. The black dashed line indicates the maximum cumulative oil production for the simulated case and is used as a reference in both training and validation.

Table 4—ConvLSTM hyperparameters.

Hyperparameters	Value
Kernel	1x1
Filter	32
Dropout	0.6
Epochs	100 with early stop @ 50
Number of hidden units	350
Activation function	elu
Kernel regularizer	L2 0.02

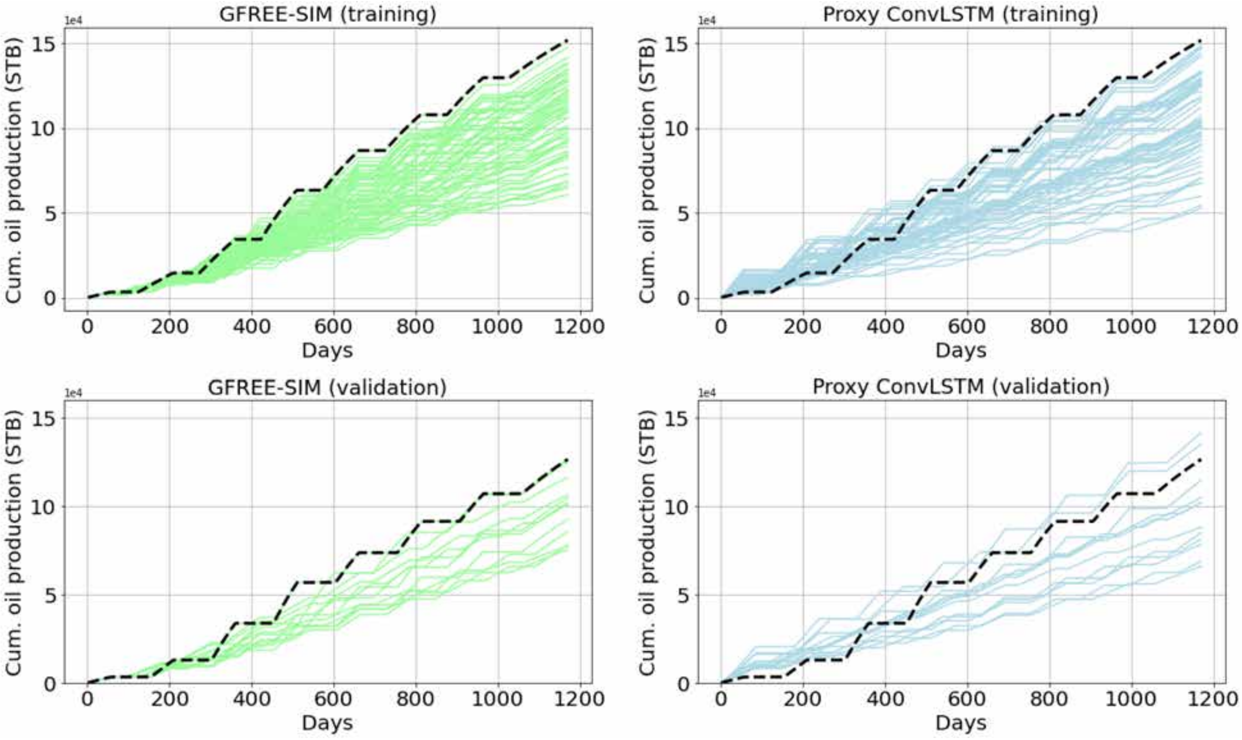

Figure 11—Cumulative oil production. Upper left-hand side: training dataset results from the multi-porosity numerical simulator (GFREE-SIM); upper right-hand side: training dataset results from the ConvLSTM model; lower left-hand side: validation dataset results from the multiporosity numerical simulator; lower right-hand side: validation dataset results from the ConvLSTM model.

Results

Model validation

The proxy multi-porosity reservoir simulation model has been extensively validated with 80% training and testing sets as well as 20% for validation of the entire number of cases. The objective is to identify how the proxy model performs on unseen data (cases that have not been integrated in the training process), which in turn leads to a beneficial hyperparameter selection (tune up) procedure. The validation was performed in three models: Seq2Seq, Luong Attention and ConvLSTM.

The performance of the model is determined through the mean absolute error (MAE) in STB, which measures the error between the proxy model predictions with respect to the multi-porosity numerical simulator results; this is shown in Table 5 Smaller error between paired observations is achieved using the ConvLSTM. Although the Luong Attention model did not accomplish the best performance, it is remarkable to note the considerable improvement over the Seq2Seq as a consequence of dealing properly with large input sequences by adding an attention layer.

Table 5—MAE results comparing three models against numerical simulation results.

Proxy Model	MAE train	MAE test	MAE validation
Seq2Seq	47.41513	53.17095	56.39743
Luong Attention	29.42098	28.8994	27.51325
ConvLSTM	22.89125	20.70289	22.12897

Figure 12 summarizes the value distribution for both sets (training-testing) and validation using the ConvLSTM model. For the most part, the value distribution results confirm that the ConvLSTM model represents similar values to those results obtained from the multi-porosity numerical simulations. According to this analysis, there is an acceptable error vis-a-vis the simulator outputs, indicating that the proxy machine learning model is capable of predicting daily oil production, achieving the specific purpose of this study.

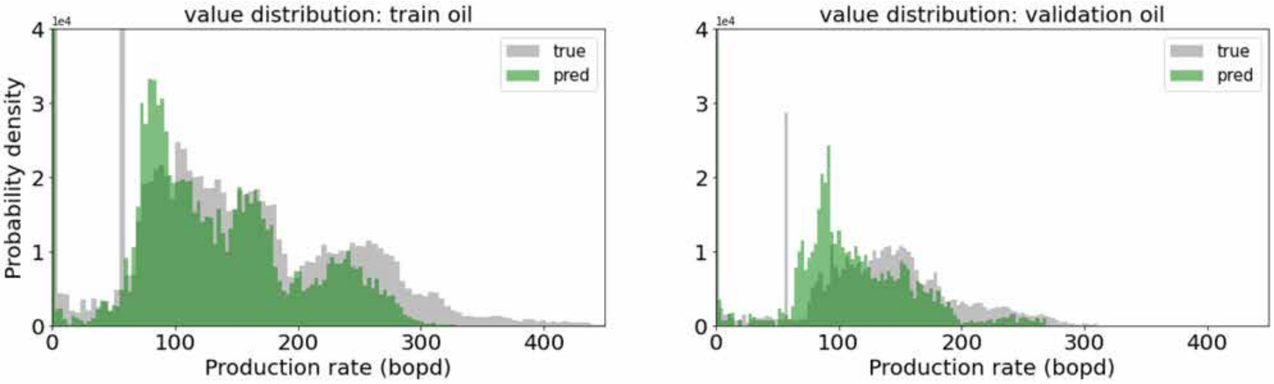

Figure 12—Value distribution. (Training and validation).

Time performance

The acceleration performance varies depending on the well control scenarios. As a reference, the numerical simulation was conducted on an Intel(R) Core (TM) i7-10750H CPU @ 2.60GHz, with 8.00 GB installed RAM. It was established that the larger the injection periods and gas injection rates, the larger was the simulation time. Table 6 compares the prediction time of three different injection/production periods and one gas injection rate (e.g., 1.7 MMscfd) using the multiporosity numerical simulator and the proxy model. As noticed, increasing the injection period can substantially increase the simulation time from 590 minutes to over 10,000 minutes considering the same injection rate. In addition, Figure 13 demonstrates that increasing the injection rate results in a significant growth in the simulation time (the x-axis refers to 1 for 1 MMscfd, and it moves increasingly to 16, indicating the maximum gas injection rate of 2.5 MMscfd). For the case of 3 months injection/production respectively, the simulation time is too long, and it was unworthy to continue collecting the run time for those specific scenarios. Therefore, we decided to extrapolate it (red dashed line), assuming it follows a second polynomial order function accomplishing an R2 equal to 0.98.

Table 6—GFREE-SIM simulation time performance compared to the ConvLSTM proxy model.

Model	Well control scenarios	Simulation time (mins)	Acceleration
GFREE-SIM	Injection: 1 month Production: 2 months Gas injection rate: 1.7 MMscfd	590	
ConvLSTM		0.5	1,180X
GFREE-SIM	Injection: 2 months Production: 2 months Gas injection rate: 1.7 MMscfd	871	
ConvLSTM		0.5	1,742X
GFREE-SIM	Injection: 3 months Production: 2 months Gas injection rate: 1.7 MMscfd	10,337	
ConvLSTM		0.5	20,674X

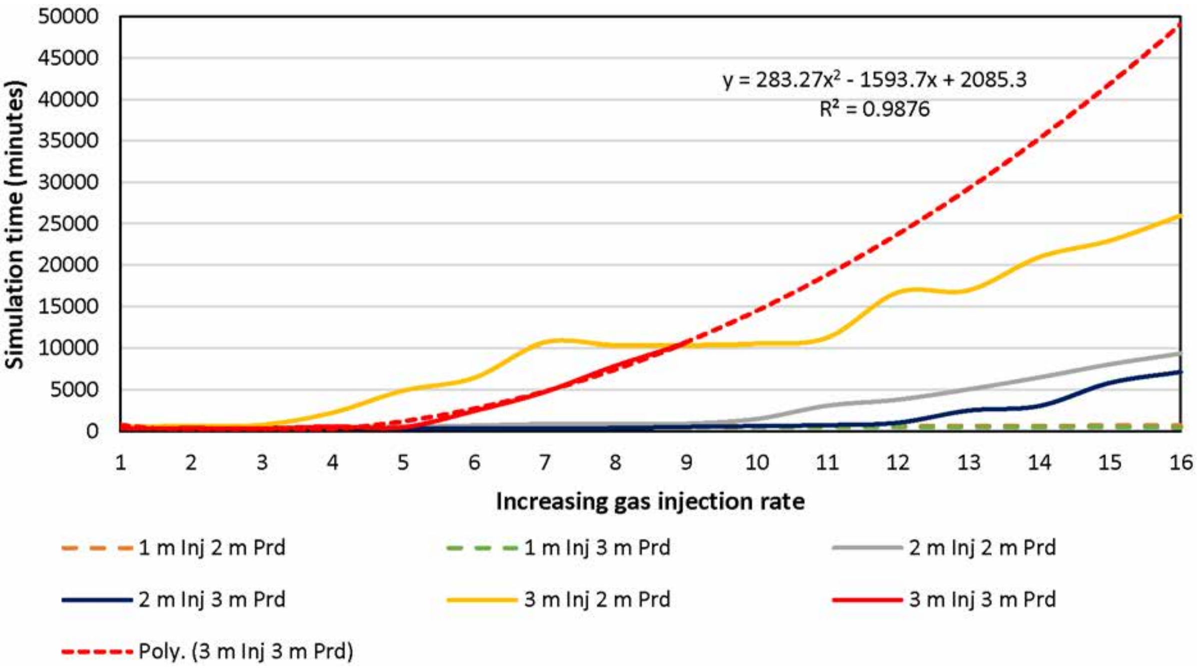

Figure 13—GFREE-SIM simulation time increasing gas injection rate.

Highest accelerations are the result of long simulations times rather than the time in which the proxy model predicts one case. In fact, the proxy model runs any case regardless of the injection/production periods and/or gas injection rate. For instance, the case when the injection period is 3 months and the production period is 2 months, there is an important acceleration of 20,674X in contrast the case with injection period equals to 1 and production period equals to 2, with an acceleration of 1,180X. The advantage of the proxy is that by reducing the simulation time to even a thousand times leads to a sharp reduction in computational cost to pave the way for optimization analysis.

Conclusions

In this study we present, step-by-step, the construction of a proxy multi-porosity numerical simulator with the aim of reducing the computational cost for H-n-P gas injection optimization in a shale reservoir. Furthermore, we present an analysis for hyperparameter tuning and determining the proper model selection. Conclusions are as follows:

1. Sacrificing some accuracy within reason, the proxy model can predict H-n-P results from a sequence of actions that include well control configurations such as injection/production periods, gas injection rate and the formation skin factor.
2. The ConvLSTM technique performed better than the Seq2Seq and Luong Attention models. The last two are well-known architectures commonly used in language processing. Luong Attention model performs better than Seq2Seq when dealing with large input sequences.
3. The computational cost of simulating all the cases in the multiporosity numerical simulator is extremely large, and therefore, it makes it unfeasible to run a complete optimization analysis. It is the combination of physics-based numerical simulation validated with actual field data and a suitable proxy model what can lead to reach a sensible and pragmatic optimization of H-n-P in a reasonable time frame.
4. Based on this study, the simulation time with the H-n-P proxy model can be accelerated more than 20,000X as compared with numerical simulation. This presumes the acceptance of a reasonable reduction in the accuracy of the proxy results. There are obviously some time variations depending on the well control configurations.

Acknowledgments

We acknowledge the contributions of former members of the GFREE research team (Geoscience (G); Formation Evaluation (F); Reservoir Drilling, Completion, and Stimulation (R); Reservoir Engineering (RE); and Economics and Externalities (EE)) and the Schulich School of Engineering at the University of Calgary. In addition, we want to extend our gratitude to the SPE data science mentorship program, Calgary, 2022, for their contributions to make this project possible.

List of equivalent terms

batch size = it is a hyperparameter that influences the speed and stability of the learning process by controlling how many training samples are passed through the network prior to the updating of the weights.

dropout rate = it is a hyperparameter that is used as a regularization technique in which some neurons are ignored at random, mainly used to minimize the loss function and reduce the variance of the model.

hidden units = it can be considered as a fixed-length context vector that receives and compiles the data from the encoder's output and feeds it to the decoder as input.

index = position of an element in a data frame.

truncate = the sequence of the input and output is divided into smaller portions by sliding windows (e.g., it takes the first sequence of input variables in one time step as the input of the model and outputs the sequence of 30 days of daily oil production).

Nomenclature

$\odot$	= hadamard product,
W	= weights
$\mathscr{H}$	= hidden state
C	= memory state
b	= bias

Acronyms

ANN	= artificial neural networks
CEC	= constant error carousel

DCA	= decline curve analysis
DL	= deep learning
EUR	= estimated ultimate recovery
H-n-P	= huff and puff
LSTM	= long short-term memory
MAE	= mean absolute error
MAPE	= mean absolute percentage error
ML	= machine learning
MSE	= mean square error
NPV	= net present value
RMSE	= root mean square error
RNN	= recurrent neural networks
Seq2Seq	= sequence to sequence

References

Alharthy, N., Teklu, T. W., Kazemi, H., Graves, R. M., Hawthorne, S. B., Braunberger, J., & Kurtoglu, B. 2018. Enhanced Oil Recovery in Liquid-Rich Shale Reservoirs: Laboratory to Field. *SPE Reservoir Evaluation & Engineering*, **21**(01), 137-159. doi:10.2118/175034-PA

Amini, S. 2014. *Developing a Grid-Based Surrogate Reservoir Model Using Developing a Grid-Based Surrogate Reservoir Model Using Artificial Intelligence*. PhD Thesis, Petroleum and Natural Gas Engineering Department. https://doi.org/10.33915/etd.5096

Aranguren, C., Fragoso, A., & Aguilera, R. 2022. *Sequence-to-Sequence (Seq2Seq) Long Short-Term Memory (LSTM) for Oil Production Forecast of Shale Reservoirs*. https://doi.org/10.15530/urtec-2022-3722179

Chaki, Soumi, Zagayevskiy, Yevgeniy, Shi, Xuebei, Wong, Terry, and Zainub Noor. 2020. *Machine Learning for Proxy Modeling of Dynamic Reservoir Systems: Deep Neural Network DNN and Recurrent Neural Network RNN Applications*. Paper presented at the International Petroleum Technology Conference, Dhahran, Kingdom of Saudi Arabia, January 2020. doi: https://doi-org.ezproxy.lib.ucalgary.ca/10.2523/IPTC-20118-MS

Fragoso, A., Lopez, B., Aguilera, R., & Noble, G. 2019. *Matching of Pilot Huff-and-Puff Gas Injection Project in the Eagle Ford Shale Using a 3D 3-Phase Multiporosity Numerical Simulation Model*. https://doi.org/10.2118/195822-MS

Javadpour, F., Fisher, D., and Unsworth, M. 2007. Nanoscale Gas Flow in Shale Gas Sediments. *Journal of Canadian Petroleum Technology*. Volume **46**, No 10, 55 – 61. October 2007. https://doi.org/10.2118/07-10-06.

Lopez Jimenez, B. A. 2017. *Characterization and Construction of 3D Numerical Simulators for Oil and Liquids-Rich Multi-Porosity Shale Reservoirs*. PhD Thesis, Chemical and Petroleum Engineering Department. http://dx.doi.org/10.11575/PRISM/25260

Lopez Jimenez, B. A. and Aguilera, R. 2019. Physics-Based Fluid Flow Modeling of Liquids-Rich Shale Reservoirs Using a 3D 3-Phase Multi-Porosity Numerical Simulation Model. *SPE Reservoir Evaluation and Engineering (2019)*. DOI: https://doi-org.ezproxy.lib.ucalgary.ca/10.2118/191459-PA

Lopez, B., & Aguilera, R. 2015. SPE-175115-MS *Physics-Based Approach for Shale Gas Numerical Simulation: Quintuple Porosity and Gas Diffusion from Solid Kerogen*. https://doi.org/10.2118/175115-MS

Lopez, B. and Aguilera, R. 2018. Petrophysical Quantification of Multiple Porosities in Shale-Petroleum Reservoirs with the Use of Modified Pickett Plots. *SPE Reservoir Evaluation and Engineering*. Volume **21**, No 01, 187 – 201, February 2018. https://doi.org/10.2118/171638-PA.

Loye, G. 2019. *Attention Mechanism*. https://blog.floydhub.com/attention-mechanism/

Luong, M.-T., Pham, H., & Manning, C. D. (2015). *Effective Approaches to Attention-based Neural Machine Translation*. https://doi.org/10.48550/arXiv.1508.04025

Navratil, J., de Paola, G., Nadukandi, P., Codas, A., & Ibanez-Llano, C. 2020. *An End-to-End Deep Sequential Surrogate Model for High Performance Reservoir Modeling: Enabling New Workflows*. https://doi.org/10.2118/201775-MS

Navratil, J., King, A., Rios, J., Kollias, G., Torrado, R., & Codas, A. 2019. *Accelerating Physics-Based Simulations Using Neural Network Proxies: An Application in Oil Reservoir Modeling*. https://doi.org/10.3389/fdata.2019.00033

Piedrahita, J., Lopez, B. and Aguilera, R. 2019. Generalized Methodology for Estimating Stress-Dependent Properties in a Tight Gas Reservoirs and Extension to Drill-Cuttings Data. SPE Reservoir Evaluation and Engineering. Volume **22**, No 01, 173– 189, February 2019. https://doi.org/10.2118/189972-PA.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & Woo, W. 2015. *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting*. http://arxiv.org/abs/1506.04214

Wan et al, 2015 Wan, T., Yu, Y., & Sheng, J. J. 2015. *Experimental and Numerical Study of the EOR Potential in Liquid-Rich Shales by Cyclic Gas Injection*. https://doi.org/10.1016/j.juogr.2015.08.004

Wang, X., Liu, Y., Sun, C., Wang, B., & Wang, X. 2015. Predicting polarities of tweets by composing word embeddings with long short-Term memory. ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, **1**, 1343–1353. https://doi.org/10.3115/v1/p15-1130

Yang Li, Yang, G., & Li, X. 2021. *Improved Fluids Characterization Model During Gas Huff-n-Puff EOR Processes in Unconventional Reservoirs*. https://doi.org/10.2118/200873-MS