Stochastic pix2vid: a new spatiotemporal method for image-to-video synthesis in geologic CO_2 storage prediction

Misael M. Morales[1*], Carlos Torres-Verdín[1,2] and Michael J. Pyrcz[1,2]

[1]Hildebrand Department of Petroleum and Geosystems Engineering, The University of Texas at Austin, Austin, TX, USA.

[2]Jackson School of Geosciences, The University of Texas at Austin, Austin, TX, USA.

*Corresponding author(s). E-mail(s): misaelmorales@utexas.edu;

Abstract

Numerical simulation of multiphase flow in porous media is an important step in understanding the dynamic behavior of geologic $\mathbf{CO_2}$ storage (GCS). Scaling up GCS requires fast and accurate high-resolution modeling of the storage reservoir pressure building and saturation plume migration; however, such modeling is challenging due to the high computational costs of traditional physics-based simulations. Deep learning models trained with numerical simulation data can provide a fast and reliable alternative to expensive physics-based numerical simulations. We present a pix2vid neural network architecture for solving multiphase fluid flow problems with superior speed, accuracy, and efficiency. The pix2vid model is designed based on the principles of computer vision and video synthesis and is able to generate dynamic spatiotemporal predictions of fluid flow from static reservoir models. We apply the pix2vid model to a highly-complex $\mathbf{CO_2}$-water multiphase problem with a wide range of reservoir models in terms of porosity and permeability heterogeneity, facies distribution, and injection configurations. The pix2vid method is first-of-its-kind in static-to-dynamic prediction of reservoir behavior, where a single static input is mapped to its dynamic response. The pix2vid method provides superior performance in highly heterogeneous geologic formations and complex estimation such as gas saturation and pressure buildup plume determination. The trained model can serve as a general-purpose, static-to-dynamic alternative to traditional numerical reservoir simulation of 2D $\mathbf{CO_2}$ injection problems with significant speedups compared to traditional methods.

Keywords: Image-to-video synthesis, Spatiotemporal forecasting, Convolutional neural network, Recurrent neural network, Proxy model

1 Introduction

Geologic CO_2 sequestration (GCS) has emerged as a proven technology to reduce anthropogenic greenhouse gas emissions to the atmosphere [citation]. This has become increasingly popular worldwide due to the need to meet international climate protection agreements [citation]. However, there are several technical challenges associated with the modeling of large-scale GCS operations. In order to accurately forecast and monitor subsurface multiphase flow, physics-based high-fidelity numerical simulation is required. These numerical simulations are computationally intensive and time-consuming since they require iterative solutions of large-scale nonlinear systems of equations [citation]. Similarly, due to the large degree of uncertainty in subsurface data collection, inherent uncertainty in the spatial distribution of the properties of heterogeneous porous media require a robust probabilistic assessment for improved engineering decision-making [citation]. In order to capture the fine-scale multiphase flow behavior given an uncertain spatial distribution of subsurface properties, a large number of forward numerical simulation runs are required, leading to very high computational costs [citation]. To overcome this, machine learning techniques have emerged as candidate reduced-order models (ROMs) for efficient parameterization and prediction of subsurface flow and transport behavior [citation].

Recent advancements in computing power, specifically GPU-enabled neural network models, have accelerated the fields of forward and inverse modeling [citation]. Classical techniques are often hindered by the size of the models and data, specifically the volume, velocity, variety, value, and veracity encountered in big data [citation]. By analyzing extensive data sets, machine learning techniques can uncover complex latent patterns and relationships that may not be discernible through traditional methods [citation]. When combined with a reduced-order modeling framework, machine learning approaches can efficiently and accurately exploit latent or salient features hidden in the data, removing redundancies or noise, and decreasing the order of the problem significantly [citation]. These approaches can often be divided into two main categories, namely purely data-driven mapping operators or physics-informed neural networks (PINNs). Typically, the training process for PINNs is done by the minimization of the (physical) loss from the residual of the governing partial differential equations (PDEs) that govern the system along with the losses associated with the initial and boundary conditions [citation]. However, over variants of PINNs such as physics-guided or physics-constrained neural networks have also proven useful for subsurface energy resource engineering applications [citation]. On the other hand, data-driven mapping operators, or proxy models, are neural network architectures trained with labeled data that produce a mapping from input features to output parameters [citation]. This procedure requires significant amounts of training data but can be applied to a wide variety of settings and conditions [citation] but suffer from lack of generalization and struggle to provide accurate predictions away from the domain of the training data. For both approaches, typically, spatial relationships are captured through convolutional neural networks (CNNs) and the temporal relationships through recurrent neural networks (RNNs) [citation], but recent advancements in transformer-based architectures are showing improved performance compared to the aforementioned techniques [citation]. In general, efficient compression of the input features into a representative latent space is proven as an effective approach for spatial and temporal parameterization of the forward or inverse problem.

A number of machine learning-based proxy (or surrogate) models have been developed to estimate the reservoir behavior in subsurface energy resource applications. Most techniques rely on the concept of image translation, or pix2pix, where a target image is predicted from an input image [citation]. Maldonado and Pyrcz [citation] developed a convolutional U-net model to predict pressure and saturation states given an uncertain geologic realization. This work is an example of image-to-image static forecasting, where the time state is given as an input, and the proxy model will predict a single response state of pressure and saturation at the given time. Wen and Benson [citation] developed a Fourier Neural Operator (FNO) architecture to predict image-to-image response states of pressure and saturation from an uncertain geologic realization and was further extended for multi-scale and nested domains [citation]. Moreover, numerous other proxy models have been developed for subsurface applications using more complex architectures such as generative adversarial networks (GANs) [citation] and transformers [citation]. However, most of these formulations are presented as an even-determined or sometimes over-determined estimation problem, with equal or greater number of input features compared to the output parameters since they are based on the pix2pix, or image-to-image formulation.

Moving beyond image-to-image predictions, Kim and Durlofsky [citation] developed a convolutional-recurrent proxy for image-to-series forecasting and discussed its advantages for closed-loop reservoir management under geologic uncertainty. This method moves beyond the image-to-image forecasting and exploits a spatiotemporal latent space in the encoder-decoder neural network architecture to obtain well flow rates and pressures over time from a static geologic realization. The image-to-series formulation can still be an even- or over-determined estimation problem, where we have equal or more inputs than outputs. Furthermore, Tang et al. [citation] and Jiang and Durlofsky [citation] developed a recurrent residual U-net (R-U-net) proxy for the prediction of dynamic pressure- and saturation-over-time from uncertain geologic realizations. This method aim to obtain dynamic response states over time from a single static input. This proxy is formulated as a more interesting under-determined estimation problem, where the number of input features is a fraction of the number of output parameters. However, the recurrent R-U-net proxy is limited by the fact that only the latent space receives spatiotemporal processing, while the model reconstruction is done via time-distributed deconvolutions, treating time as an additional "spatial" dimension, and not fully exploiting the spatiotemporal relations in the data and latent space as an image-to-video forecasting formulation.

The problem of image-to-video forecasting, also known as video synthesis, has been approached previously by researchers in the field of computer vision. Iliadis et al. [citation] were the first to develop a deep learning-based framework for video compressive sensing to reconstruct a video sequence from a single measured frame using a deep fully-connected neural network, or artificial neural network (ANN). Despite excellent accuracy in the video predictions, this method is still limited by time-distributed fully-connected layers in the encoder and decoder portions of the network, thus not exploiting the spatiotemporal relationships in the data. Xu and Ren [citation] developed a three-part encoder-recurrent-decoder network for video reconstruction from the estimated motion fields of the encoded frames. The implementation is similar to that of Tang et al. [citation] and Jiang and Durlofsky [citation] in that it applies a recurrent update in the latent space but relies on time-distributed deconvolutions for the video frames reconstruction. Dorkenwald et al. [citation] developed a conditional invertible neural network (cINN) as a bijective mapping between image and video domains using a dynamic latent representation. The cINN architecture allowed for video-to-image and image-to-video predictions, proving possible the generation of video frames from a static input image. Finally, Holynski et al. [citation] implemented the idea of Eulerian motion fields to define the moving portions of an image and thus were the able to accurately reconstruct a series of video frames from a static image using a spatiotemporal latent space parameterization. These advancements in the field of computer vision and video compressed sensing serve as a foundation for our image-to-video spatiotemporal proxy model.

In this work, we propose a novel image-to-video spatiotemporal proxy model for the prediction of dynamic reservoir behavior over time from an uncertain static geologic realization. In this work, we apply the spatiotemporal proxy to a large-scale GCS operation. Our model exploits the spatial and temporal structures in latent space to dynamically reconstruct the time-dependent pressure and saturation states from a static geologic realization. The encoder portion of the network receives as inputs the static geologic realization with channels representing the porosity, permeability, and facies distributions, and the location of CO_2 injection well(s). The uncertain geologic realizations are generated from a wide array of possible geologic scenarios (e.g., fluvial, turbidite, and deepwater lobe systems), and the number and location of CO_2 injection wells is also considered uncertain. The model then reconstructs the dynamic pressure and saturation distributions using a spatiotemporal decoder network with convolutional long short-term memory (ConvLSTM) layers, which are concatenated with the residuals of the spatial latent parameterizations from the encoder network. Thus, it is not an encoder-recurrent-decoder architecture, but instead a fully spatiotemporal convolutional-recurrent image-to-video model. Our proxy model shows significant advantages compared to image-to-image and encoder-recurrent-decoder models in terms of computational efficiency and prediction accuracy and can be used as a replacement for high-fidelity simulations (HFS) in GCS projects as an image-to-video mapping operator.

In the methodology section, we discuss the proposed spatiotemporal proxy model architecture as well as the geologic modeling and numerical reservoir simulation steps required to generate the training data. In the results and discussion sections, we evaluate the training and performance of the proposed proxy model and compare its efficiency and accuracy to high-fidelity numerical simulations using a 2D synthetic case for large-scale GCS operations.

2 Methodology

This section describes the governing equations, reservoir model and simulation specifications, model architecture, and training strategy of the pix2vid model.

2.1 Governing equations

For the CO_2-water multiphase flow problem, the general form of the mass accumulation for component $\kappa = CO_2$ or water is given by [citation]:

$$\frac{\partial M^k}{\partial t} = -\nabla \bullet F^\kappa + q^\kappa. \tag{1}$$

For each component κ, the mass accumulation term M^κ is summed over all phases p,

$$M^k = \phi \sum_p S_p \rho_p X_p^\kappa \tag{2}$$

where ϕ is the porosity, S_p is the saturation of phase p, ρ_p is the density of phase p, and X_p^κ is the mass fraction of component κ present in phase p. For each component κ, there is also the advective mass flux $F^\kappa|_{adv}$ obtained by summing over all phases p,

$$F^\kappa|_{adv} = \sum_p X_p^\kappa F_p \tag{3}$$

where each individual phase flux F_p is given by Darcy's equation:

$$F_p = \rho_p u_p = -k \frac{k_{r,p}\rho_p}{\mu_p}(\nabla P_p - \rho_p g). \tag{4}$$

Here, u_p is the Darcy velocity of phase p, k is the absolute permeability, $k_{r,p}$ is the relative permeability of phase p, μ_p is the viscosity of phase p, and g is the gravitational acceleration constant. The fluid pressure of phase p,

$$P_p = P + P_c \tag{5}$$

is given by the sum of the reference phase pressure P and the capillary pressure P_c. The numerical simulation does not include molecular diffusion or hydrodynamic dispersion for practical purposes.

2.2 Reservoir Model and Simulation

We use SGeMS [citation] to construct an ensemble of realizations that is representative of various potential geologic scenarios for CO_2 storage in deep geological formations. Using sequential Gaussian co-simulation [citation], we generate a set of 1,000 random porosity (ϕ) and permeability (k) distributions with a wide range of values, as shown in Figure 1. Facies distributions are obtained from a library of deepwater fluvial training images [citation]. These encompass a wide range of possible geologic scenarios including marked point (lobe, ellipse, and bar), FluvSim (channel, channel-levee, and channel-levee-splay), surface based (compensational cycles of lobes), and bank retreat (channel complex). To generate consistent porosity and permeability distributions with the facies-based geologic scenarios, we conditionally multiply the original porosity and permeability distributions with the facies distributions. The resulting fluvial distributions are shown in Figure 2.

The conditioned fluvial porosity and permeability distributions simulated for the problem of geologic CO_2 storage using MRST [citation]. Specifically, the MRST-CO_2lab module is used as an automatic-differentiation framework for the compositional simulation of the two-phase CO_2-water problem. The reservoir is initialized as a fully water saturated zone (i.e., aquifer) with an initial pressure of 4,000 psi. The reservoir has constant isothermal conditions and pressure boundary conditions and represents a large-scale geologic CO_2 storage project with negligible dip, such as found in the Illinois Basin and parts of the North Sea and Gulf Coast.

The model has dimensions of 1km-1km-100m in the x-, y-, and z-directions, respectively. We use 64 uniform grid cells in the x- and y-directions. The grid design is sufficiently refined to resolve the pressure and saturation plumes in highly heterogeneous reservoirs while remaining computationally tractable for the purpose of training deep learning models. A random number of injection wells, $w \in [1, 3]$, are placed randomly along the reservoir for each of the 1,000 realizations. Each injection well has a constant radius of 0.1m and a single and continuous perforation that injects pure supercritical CO_2 at a constant rate such that the total injection rate of the w wells is 0.5 megatons per year.

The numerical simulation is run for 5 years, monitored monthly, for a total of 60 timesteps. At each grid cell and for each time step, we resolve the implicit pressure, explicit saturation (IMPES) formulation of Eq. (1) to obtain the corresponding dynamic pressure and saturation distributions over time (videos) from the static geologic realizations of porosity and permeability conditioned to the fluvial facies (images).

2.3 Proxy Model Architecture

The pix2vid model is designed as an image-to-video data mapping operator from the static realizations of geologic distributions of porosity, permeability and facies as well as the injector well(s) distribution, to the dynamic responses of pressure and saturation over time. A single model is trained to predict both pressure and saturation distributions over time as a multi-channel output.

Let m represent a geologic model realization of porosity, permeability, facies, and injector well(s) distributions, such that $m = \{\phi, k, facies, w\}$. The dynamic respones of pressure and saturation over time are given by $d = f(m)$, such that $d = \{P(t), S(t)\}$ and f is the physics-based numerical reservoir simulation mapping operator. Our aim is to replace f with a more efficient data mapping operator by training the stochastic pix2vid model. For this purpose, we exploit the latent structures in space and time of the static inputs and dynamic outputs through a spatiotemporal encoder-decoder architecture. The encoder portion of the network is comprised of sequential convolutional layers to compress the spatial features of the model realizations into a latent parameterization z_m, given by $z_m = Enc(m)$. In their compressed representation, these features represent the salient characteristics of the geologic distributions. The decoder portion of the network is designed as a series of recursive residual convolutional-recurrent layers, such that the latent space z_m is recursively decoded into the dynamic distributions of pressure and saturation over time. The previous timestep latent representations, z_d^t, are used in the subsequent timestep to refine the outputs and reduce systematic error propagation in time. Thus, the full architecture is represented as

$$d = Dec^t([Enc(m); z_d^t]) \tag{6}$$

The encoder portion compresses the geologic realizations, m, into a latent representation z_m through the usage of separable convolutions [citation]. This type of convolution learns the parameters for each channel in the image separately, avoiding mixing of variables or loss of resolution. This is especially important when dealing with Gaussian-distributed permeability and porosity in combination with binomial-distributed facies and binary well(s) location distributions. Each separable convolution layer is regularized with an l_1-norm weight of 1×10^{-6}. Moreover, we use a squeeze-and-excite layer to improve channel interdependence, also avoiding mixing and loss of resolution [citation]. Each squeeze-and-excite layer will provide the optimal network weights for each channel independent of the other channels, adding content aware mechanisms to weight each channel adaptively. Furthermore, by applying instance normalization as opposed to the more common batch normalization, we achieve channel-independent normalization of the convolved features [citation]. Parametric rectified linear units (PReLU) is used as the activation function, where at each minibatch iteration, the network learns the optimal leaky slope for activation in each layer. Through 3 convolutional encoding layers, we obtain the latent parameterizations z_m^1, z_m^2, and z_m^3, as shown in Table 1.

Table 1 Encoder network architecture

Layer	Shape in	Shape out
row 1	data 1	data 2
row 2	data 4	data 5
row 3	data 7	data 8

The decoder portion of the pix2vid model extracts the spatiotemporal relationships from the latent representations of m to reconstruct the dynamic pressure and saturation responses over time, d. To accurately reconstruct the spatiotemporal structure from the static latent space z_m, we employ a series of convolutional-recurrent layers, namely a convolutional long-short term memory layer (ConvLSTM). Through 3 convolutional-recurrent layers, we obtain the dynamic prediction of d as follows:

1. *Spatiotemporal decoding of z_m^3*: The first ConvLSTM layer takes the smallest latent representation, z_m^3, and reconstructs the first decoded timestep z_d^3.
2. *Residual concatenation of z_m^2*: The first decoded timestep, z_d^3, is concatenated with the intermediate static encoding z_m^2 to retain multi-scale features and improve prediction performance and resolution.
3. *Intermediate spatiotemporal decoding*: The second ConvLSTM layer takes the combined intermediate latent representation, $[z_m^2, z_d^3]$ to predict the next spatiotemporal representation z_d^2.
4. *Residual concatenation of z_m^1*: The intermediate decoded timestep, z_d^2, is concatenated with the largest static encoding z_m^1.
5. *Final spatiotemporal decoding*: The third ConvLSTM layer takes the combined larger latent representation, $[z_m^1, z_d^2]$ to predict the full-scale dynamic output, d.

To enhance the performance of the spatiotemporal decoding, each ConvLSTM layer is followed by a batch normalization, activation, and a transpose convolutional layer, the latter for downscaling the latent space to twice its dimension. Spatial dropout is then applied, and the concatenated features are once more convolved and activated to obtain the layer prediction. Table 2 shows the architecture of the decoder network:

Table 2 Decoder network architecture

Layer	Shape in	Shape out
row 1	data 1	data 2
row 2	data 4	data 5
row 3	data 7	data 8

This process yields the first video frame prediction, d^1, from the latent representation of the geologic realizations z_m. Each subsequent video frame prediction is obtained by another set of residual concatenation of the previous timestep dynamic decoded representation. The static latent representation z_m is concatenated at each timestep with the previous dynamic decoded representation for each layer such that we have $[z_m, z_(d_t)^i]$, where i is the decoding layer number and t is the timestep. By recursively implementing spatiotemporal decoding to the latent representation z_m, we obtain the prediction of the dynamic response at times $[t_0, t_1, \dots, t_n]$ for each iteration $t = 1, \dots, n$. The complete decoder architecture is shown in Figure 3.

2.4 Training Strategy

The inputs to the stochastic pix2vid are the geologic realizations, comprised of the distributions of porosity, permeability, facies, and injection well(s) location, represented as a matrix m of dimensions $64 \times 64 \times 4$. The outputs are the results from the numerical reservoir simulation, namely pressure and saturation distributions over time, represented as a matrix d of dimensions $64 \times 64 \times 60 \times 2$. This yields an ill-posed and under-determined estimation problem, which is extremely difficult to resolve [citation]. To improve the training efficiency and performance, we subsample in time from 60 timesteps to 11. In other words, instead of monthly monitoring, we predict the dynamic outputs at the initial step and every 6 months afterward. We also perform min-max normalization so that the input and output features are in the range of $[0, 1]$, which greatly improves the performance of the nonlinear activation functions. Furthermore, we perform data augmentation by 90° rotation, making the network agnostic to orientation and effectively learning the flow physics in the system rather than memorizing spatial distribution patterns. The total amount of training data is therefore 2,000 realizations (after augmentation), which is split into 1,500 realizations for training and 500 realizations for testing. To improve model generalizability, at each epoch, each minibatch is split into 80/20 for training and validation sets, respectively.

A custom three-part loss function is used to accurately predict pixel-wise and perceptual information in the predictions. The mean squared error (MSE) is used to reconstruct the pixel-wise intensity values, while the mean absolute error (MAE) is used to optimize for the pressure and saturation plume edges. The third part is the structural similarity index metric (SSIM), which provides a perceptual image-to-image comparison of luminance, contrast, and structure. For optimal training, the aim is to minimize the MSE and MAE while maximizing the SSIM for the true versus predicted outputs, d and $\hat{d}$, such that the total loss is given by:

$$\mathcal{L} = \alpha(1-SSIM)+(1-\alpha)[\beta\cdot MSE+(1-\beta)MAE] \tag{7}$$

where α and β are weighting coefficients obtained empirically as 0.33 and 0.66, respectively.

The model is trained using the AdamW optimizer [citation]. This variant of the well-known adaptive momentum (Adam) optimizer [citation] includes an added method to decay weights for the adaptive estimation of first-order and second-order moments. We implement a learning rate of 1×10^{-3} with a weight decay term of 1×10^{-5}.

3 Results

This section describes the stochastic pix2vid model training performance and discusses the results for various training and testing realizations.

3.1 Training Performance

Using an NVIDIA Quadro M6000 GPU, we train for 100 epochs with a batch size of 50. The model has a total of 97,523,370 parameters, and the training time required is approximately 68 minutes for all 1,500 training realizations. The training and validation performance per epoch is shown in Figure 99. We observe minimal overfit in the validation set, corresponding to good model generalizability and prediction accuracy. Using physics-based numerical simulation, each realization requires approximately 30 seconds to obtain the dynamic pressure and saturation predictions from the static geologic models. Our stochastic pix2vid model obtains the same results in approximately 4.59 milliseconds, corresponding to a 6,500× speedup. The average MSE for the ensemble is 9.21×10^{-4} and 9.70×10^{-4} for training and testing, respectively. Similarly, the average SSIM for the ensemble is 98.97 and 97.91 for training and testing, respectively.

3.2 Prediction Results

The stochastic pix2vid model is capable of predicting dynamic reservoir response from a static geologic model as an image-to-video data mapping operator. Figure 99 shows the predicted pressure and saturation distributions along with the absolute difference to HFS for four training realizations, one from each of the different geologic scenarios. We observe reasonable agreement between the true and predicted CO_2 pressure and saturation plumes over time, with an average MSE of 0 and SSIM of 1.

Similarly, Figure 99 shows the pressure and saturation distributions predictions along with the absolute difference to HFS for four testing realizations from each of the possible geologic scenarios. We observe a similar performance, with an average MSE of 0 and SSIM of 1. This indicates that the stochastic pix2vid model has excellent generalization ability and achieves on par performance with HFS at a fraction of the computational cost.

These results implies that our stochastic pix2vid is capable of learning the spatiotemporal relationship between the static geologic models and the dynamic reservoir response. Thus, our image-to-video architecture can outperform current image-to-image and encoder-recurrent-decoder architectures for improved reservoir behavior prediction.

A comparison of true versus predicted results for pressure and saturation responses is shown in Figure 99. For the pressure and saturation predictions, the R^2 is approximately 1% and 1%, respectively, with narrow 95% prediction bands.

This shows that despite some minor inaccuracies in the plume front prediction, the overall shape and intensity of the CO_2 pressure and saturation plumes are accurately recovered, and the model can be used as a reliable replacement for expensive numerical reservoir simulations, especially in cases where large number of runs are required to obtain dynamic estimates (e.g., well placement and control optimization, history matching, uncertainty quantification).

3.3 Discussion

The results shown above suggest that our stochastic pix2vid is an efficient and accurate predictor of dynamic reservoir response from static geologic models, serving as a reasonable replacement for physics-based numerical reservoir simulation.

CO_2 saturation and pressure buildup fronts are important quantities for geologic CO_2 storage projects and are often used for regulatory oversight [citation], monitoring metrics or history matching purposes [citation]. The distance between the injection well(s) and the saturation fronts represents the maximum extent of the CO_2 plume. However, these are often very difficult to capture accurately with data-driven proxy models. Our stochastic pix2vid model shows greater absolute error on and around the plume fronts compared to within the plumes. However, the overall shape and intensity of the pressure and saturation plumes is very well captured for all realizations despite being highly heterogeneous.

By using GPU-enabled computations, we can significantly accelerate the training and prediction time of the pix2vid model. Each HFS run was performed on an Intel ®i9-10900KF processor with 10 cores. The 1,000 realizations are parallelized equally among all cores and the total simulation time accounting for parallelization for all realizations is about 8.33 hours. Dynamic prediction using the pix2vid model on an NVIDIA RTX 3080 GPU require a total of 4.6 seconds, or 0.001275 hours, with an accuracy of 99% and 98% for training and testing, respectively. This provides a sustainable argument for the usage of our stochastic pix2vid model as a replacement for HFS when computational time is a constraint.

As described in Section 2.3, the stochastic pix2vid model takes the static geologic realizations, m, and compresses them into a latent space representation, z_m. Here we provide a visualization for a random selection of latent feature maps, along with their superposition on the porosity and facies distribution, as shown in Figure 99. This can be interpreted as an analog to the attention head mechanisms recently developed in transformer-based architectures [citation]. We observe that the latent feature maps are essentially learning the injection location(s) and direction of flow based on the geologic distributions. Thus, proving that the stochastic pix2vid model is learning multiphase flow physics and dynamic reservoir behavior appropriately.

To further demonstrate the effectiveness of our stochastic pix2vid model for geologic CO_2 storage operations, we plot the cumulative pixel-wise CO_2 saturation as a surrogate for the cumulative CO_2 volume injected. For all training and testing realizations, Figure 99 shows the sum of pixel-wise CO_2 saturation and the probability density function (PDF) of the true versus predicted saturations. We observe an R^2 of 96% for training and 95% for testing in the cumulative CO_2 saturation of true versus predicted results, and a conformable PDF for both training and testing.

4 Conclusions

In this study, we developed a deep learning-based spatiotemporal proxy model to provide flow predictions for a large-scale GCS operation. The key extension introduced is the use of a spatiotemporal convolutional-recurrent architecture for dynamic predictions of CO_2 pressure and saturation distributions over time from an uncertain static geologic realization. The framework was developed as an image-to-video prediction, which is a noteworthy under-determined estimation problem. Specifically, the implementation extends the architectures of current encoder-recurrent-decoder models and provides a fast and accurate proxy as a replacement for high-fidelity numerical reservoir simulation.

The encoder block is composed of separable convolutions, squeeze and excite layers, and instance normalization. These three special implementations allows for precise parameterization of the geologic realization into a latent representation, without mixing the effects of Gaussian distributed properties against binary of binomial distributed properties. Using recursive ConvLSTM layers, the recurrent decoder block recursively predicts each dynamic state, or frame, from the concatenation of the previous latent representation and the intermediate encoding parameterizations. Thus, this architecture presents the proxy as an image-to-video prediction formulation for dynamic reservoir states from a static geologic realization.

The spatiotemporal proxy was applied to a synthetic 2D GCS project with multiple uncertain geologic scenarios and random number and location of injection well(s). A total of 1,000 geologic models were obtained from a variety of possible geologic scenarios including fluvial, turbidite, and deepwater lobe systems. The spatial distribution of porosity, permeability and facies, and the spatial location of the injector well(s) were used as the input data. The proxy then predicts the dynamic reservoir response over time, namely the video frames, corresponding to the dynamic CO_2 pressure and saturation distributions, which are obtained offline for training using HFS. The total training time is 78 minutes on a single NVIDIA Quadro M6000 GPU, and predictions are obtained with 98-99% accuracy within approximately 4.6 milliseconds, compared to the approximate 30 seconds required for HFS – a $6,500\times$ speedup.

There are several possible directions that could be considered for future work. Firstly, an extension to 3D geologic models and their corresponding dynamic predictions is key to extending this method to real-world applications. Similarly, although the spatiotemporal convolutional-recurrent proxy was only trained for CO_2 sequestration, it should be applicable for a range of processes such as compositional, geothermal, or conventional oil and gas systems. The proxy could also be applied to several subsurface energy resource workflows such as optimization and history matching. Moreover, it would be interesting to extend the proxy from a data-driven mapping operator to a PINN by including the discretized form of the governing PDE in the loss function and minimizing the residuals. Another future direction would be to test the performance of the stochastic pix2vid model on unseen timesteps, either interpolating the training timesteps or extrapolating beyond the training timesteps. Furthermore, the proxy is robust to uncertain geology and variable number and placement of injector wells but could be extended to variable well controls and applied to robust optimization and closed-loop reservoir management workflows.

Acknowledgments. The authors thank the Digital Reservoir Characterization Technology (DIRECT) and Formation Evaluation (FE) Industry Affiliate Program at the University of Texas at Austin for supporting this work.

Declarations. The authors declare no conflict of interests.

References