

¹ Stochastic pix2vid: A new spatiotemporal deep learning
² method for image-to-video synthesis in geologic CO₂
³ storage prediction

⁴ Misael M. Morales^{1*}, Carlos Torres-Verdín^{1,2}, and Michael J. Pyrcz^{1,2}

⁵ 1. Hildebrand Department of Petroleum and Geosystems Engineering, The University of Texas at Austin

⁶ 2. Jackson School of Geosciences, The University of Texas at Austin

⁷ *Corresponding author; email: misaelmorales@utexas.edu

⁸ **Abstract**

⁹ Numerical simulation of multiphase flow in porous media is an important step in understanding
¹⁰ the dynamic behavior of geologic CO₂ storage (GCS). Scaling up GCS requires fast and accurate
¹¹ high-resolution modeling of the storage reservoir pressure and saturation plume migration; how-
¹² ever, such modeling is challenging due to the high computational costs of traditional physics-based
¹³ simulations. Deep learning models trained with numerical simulation data can provide a fast and
¹⁴ reliable alternative to expensive physics-based numerical simulations. We propose a Stochastic
¹⁵ pix2vid neural network architecture for solving multiphase fluid flow problems with superior speed,
¹⁶ accuracy, and efficiency. The Stochastic pix2vid model is designed based on the principles of com-
¹⁷ puter vision and video synthesis and is able to generate dynamic spatiotemporal predictions of
¹⁸ fluid flow from static reservoir models, closely mimicking the performance of traditional numerical
¹⁹ simulation. We apply the Stochastic pix2vid model to a highly-complex CO₂-water multiphase
²⁰ problem with a wide range of reservoir models in terms of porosity and permeability heterogeneity,
²¹ facies distribution, and injection configurations. The Stochastic pix2vid method is first-of-its-kind
²² in static-to-dynamic prediction of reservoir behavior, where a single static input is mapped to its
²³ dynamic response. The Stochastic pix2vid method provides superior performance in highly hetero-
²⁴ geneous geologic formations and complex estimation such as CO₂ saturation and pressure buildup
²⁵ plume determination. The trained model can serve as a general-purpose, static-to-dynamic (image-
²⁶ to-video) alternative to traditional numerical reservoir simulation of 2D CO₂ injection problems
²⁷ with up to 6,500× speedup compared to traditional numerical simulation.

28 **Keywords:** Image-to-video synthesis, Spatiotemporal prediction, Convolutional neural net-
29 work, Recurrent neural network, Proxy model

30 1 Introduction

31 Geologic CO₂ sequestration (GCS) has emerged as a potential technology solution to reduce an-
32 thropogenic greenhouse gas emissions to the atmosphere [1–3], and has become increasingly popular
33 worldwide due to the need to meet international climate protection agreements [4–6]. Modeling
34 injected CO₂ movement in the subsurface over and beyond the life of the project is a critical com-
35 ponent to support optimum GCS project decision making for safe and secure CO₂ sequestration. A
36 schematic of typical GCS operations is shown in Figure 1, including storage in depleted oil and gas
37 reservoir and deep saline formations, and CO₂ enhanced oil and coalbed methane recovery [7–9].
38 However, there are several technical challenges associated with the subsurface modeling to support
39 GCS operations. To accurately forecast and monitor subsurface multiphase flow, physics-based
40 high-fidelity numerical simulations are required. These numerical simulations are computationally
41 intensive and time-consuming since they require iterative solutions of nonlinear systems of equa-
42 tions applied over large volumes of the subsurface at sufficient resolution to represent heterogeneity
43 [10–13]. Also, due to the large degree of uncertainty in subsurface data, and the spatial distri-
44 bution of the properties of heterogeneous porous media between the sparsely sampled data, GCS
45 operations require a robust probabilistic-based uncertainty assessment for improved engineering
46 decision-making [14–16]. In order to capture the fine-scale multiphase flow behavior given an un-
47 certain spatial distribution of subsurface properties, a large number of numerical simulations are
48 required, leading to very high computational costs and delayed feedback unable to support timely
49 decision making [17, 18].

50 To overcome this, machine learning techniques have emerged as candidate proxy models due to
51 their ability to perform dimensionality reduction for efficient problem parameterization and model
52 complicated systems to calculate fast predictions of subsurface flow and transport behavior for
53 real-time feedback on the impact of geological and engineering controls on CO₂ behavior in the
54 subsurface over time [19–21]. Dimensionality reduction techniques are supervised or unsupervised
55 machine learning methods that compress (or encode) the data, X , into a lower-dimensional latent

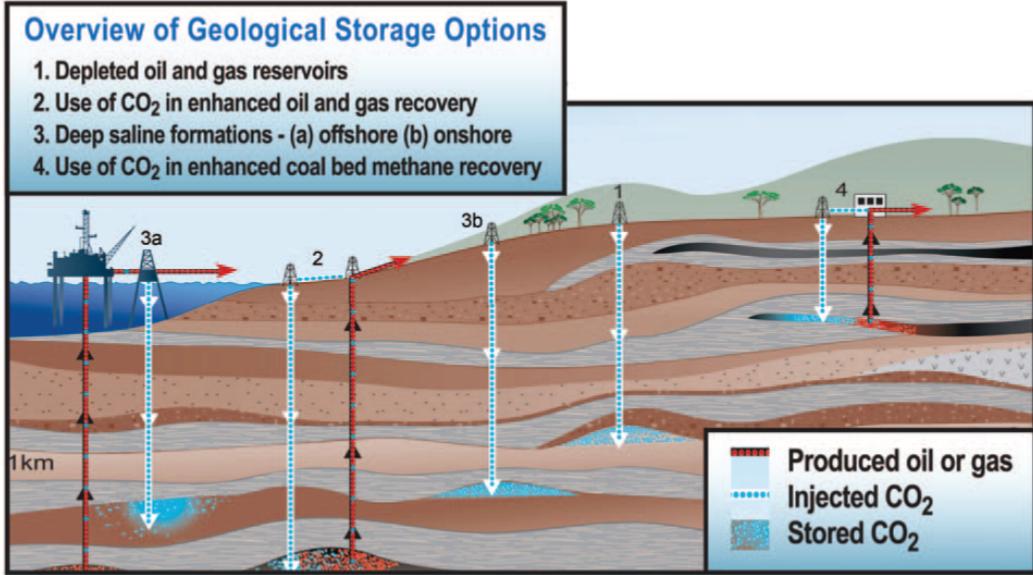


Figure 1: Types of geologic CO₂ storage operations and the geologic formations that can be used for sequestration. *Modified from the Carbon Dioxide Cooperative Research Center (CO2CRC), <http://www.co2crc.com.au/about/co2crc>*

56 feature representation, z , and decompress (or decode) the latent representation either: (1) back to
 57 the original data space, \hat{X} (unsupervised, AutoEncoder), or (2) to a new response feature space,
 58 y (supervised, Encoder-Decoder) [22–24], as shown in Figure 2. The recent advancements in deep
 59 learning algorithms and in computing architecture and power, enable GPU-enabled neural network
 60 models that have accelerated the fields of forward and inverse modeling [25, 26]. Classical statistical
 61 modeling methods are often hindered by the size of the models and their conditioning to big data,
 62 i.e., that is data with volume, velocity, variety, value, and veracity [27, 28], and fail to generalize
 63 beyond fit-for-purpose frameworks [29, 30]. By analyzing big data sets, machine learning techniques
 64 can uncover complex patterns and relationships in lower-dimensional, latent feature representations
 65 that may not be discernible through traditional statistical and geostatistical methods [31–33]. When
 66 combined with a latent space modeling framework, machine learning approaches efficiently and
 67 accurately exploit hidden patterns and features in the data, remove redundancies or noise, and
 68 decrease the mathematical and computational complexity of the problem significantly [34, 35].

69 Supervised machine learning approaches applied to the subsurface are divided into two main
 70 categories, namely purely data-driven models or physics-informed models. Data-driven proxy mod-
 71 els are neural network architectures trained with labeled data that produce a mapping from input

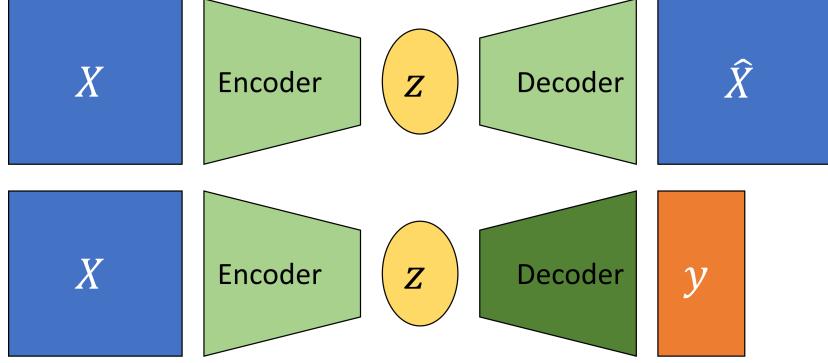


Figure 2: Dimensionality reduction model structures. Unsupervised AutoEncoder structure (top), and supervised Encoder-Decoder structure (bottom).

72 predictor feature to output response features [36, 37]. On the other hand, the training process
 73 to match training data for PINNs is regularized with the minimization of the (physical) loss from
 74 the residual of the governing partial differential equations (PDEs) along with the losses associ-
 75 ated with the initial and boundary conditions [38, 39]. However, other variants of PINNs such as
 76 physics-guided or physics-constrained neural networks where the PDE loss is not embedded in the
 77 training step, instead the models have specific architectures or parameters to mimic the physics
 78 in the system, have proven useful for subsurface energy resource engineering applications [40–42].
 79 One disadvantage of machine learning techniques is that they require significant amounts of train-
 80 ing data, but once trained these prediction models suffer from lack of generalization, i.e., inability
 81 to provide accurate predictions away from the training data beyond which they have been specif-
 82 ically trained [43, 44]. For both data-driven and physics-informed approaches, typically, spatial
 83 relationships are modeled through convolutional neural networks (CNNs) [45, 46] and the tempo-
 84 ral relationships through recurrent neural networks (RNNs) [47, 48], but recent advancements in
 85 transformer-based architectures improve performance compared to the CNN and RNN methods for
 86 spatial and temporal latent feature representations [49–51].

87 A number of machine learning-based proxy models have been developed to estimate the reservoir
 88 behavior in subsurface energy resource applications. Most techniques rely on the concept of image
 89 translation, or pix2pix, where a target image(s) is predicted from an input image(s) [52–55], as
 90 shown in Figure 3. Maldonado-Cruz and Pyrcz [56] develop a convolutional U-Net model to predict
 91 pressure and saturation states given an uncertain geologic realization. This work is an example
 92 of image-to-image static forecasting, where the time state is given as an input, and the proxy

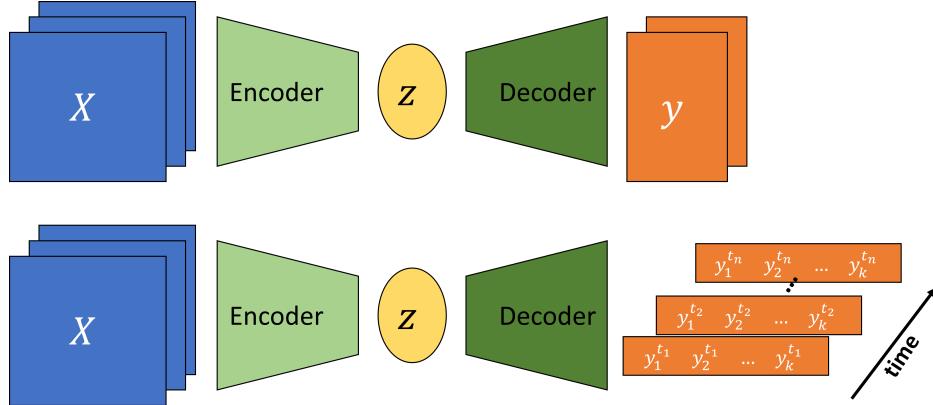


Figure 3: Image-to-image (pix2pix) (top) and image-to-timeseries (bottom) Encoder-Decoder structures.

model will predict a single response state of pressure and saturation at the given time. Wen et al. [57] develop a Fourier Neural Operator (FNO) architecture to predict image-to-image response states of pressure and saturation from an uncertain geologic realization and is further extended for multi-scale and nested domains [58]. These methods are based on a pix2pix, or image-to-image prediction, where a specific timestep is used as an input feature to predict the relationship between the geologic model and the reservoir response at that specific timestep. This implies that pix2pix or image-to-image methods are formulated as an even-determined or sometimes over-determined estimation problem, where the number of input features is equal to or greater than the number of output features. Moreover, numerous other proxy models have been developed for subsurface applications using more complex architectures such as generative adversarial networks (GANs) [59] and transformers [60, 61]. Despite showing consistent results and significant speedups compared to traditional numerical simulation, pix2pix models do not capture the spatiotemporal relationships and dynamic response of the subsurface system.

Moving beyond image-to-image predictions, Kim and Durlofsky [62] develop a convolutional-recurrent proxy for pix2time, or image-to-timeseries, forecasting and discuss 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 an encoder-recurrent neural network architecture to obtain hydrocarbon production forecasts. The image-to-series formulation can still be an even- or over-determined estimation problem, where we have equal or more inputs than outputs, as shown in Figure 3. Furthermore, Tang et al. [63, 64] and Jiang and Durlof-

sky [18] develop a recurrent residual U-net (R-U-net) proxy for the prediction of dynamic pressure-
and saturation-over-time from uncertain geologic realizations using an encoder-recurrent-decoder
architecture. These methods aim to obtain dynamic response states over time from a single static
image. This type of proxy model is formulated to resolve the more complex under-determined
estimation problem (compared to even- or over-determined), where the number of input features is
a fraction of the number of output features. 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 [65–69]. Iliadis et al. [70] are one of the
first to develop a deep learning-based framework for video compressive sensing to reconstruct a
video sequence from a single image 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 [71] develop 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 Jiang and Durlofsky [18] and Tang et al.
[63, 64] in that it applies a recurrent update in the latent space but relies on time-distributed
deconvolutions for the video frames reconstruction to exploit spatiotemporal relationships in the
data. Dorkenwald et al. [72] develop a conditional invertible neural network (cINN) as a bijective
mapping between image and video domains using a dynamic latent representation. The cINN
architecture allows for video-to-image and image-to-video predictions, demonstrating possible the
generation of video frames from a static input image. Finally, Holynski et al. [73] implemented the
idea of Eulerian motion fields to define the moving portions of the image 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 are the foundation
for our image-to-video proxy model.

We propose a novel image-to-video spatiotemporal proxy model, Stochastic pix2vid, for the pre-

143 diction of dynamic reservoir behavior over time from a subsurface uncertainty model suite of static
 144 geologic realizations. Our model exploits the spatial and temporal structures in latent space to dy-
 145 namically reconstruct the time-dependent pressure and multiphase saturation states from a static
 146 geologic realization. The model then reconstructs the dynamic pressure and saturation distributions
 147 using a spatiotemporal decoder network with convolutional long short-term memory (ConvLSTM)
 148 layers, which are concatenated with the residuals of the spatial latent parameterizations from the
 149 encoder network. Thus, it is not an encoder-recurrent-decoder architecture, but instead a fully spa-
 150 tiotemporal convolutional-recurrent image-to-video synthesis model. Our stochastic pix2vid model
 151 has significant advantages compared to image-to-image and encoder-recurrent-decoder models in
 152 terms of computational efficiency and prediction accuracy and can be used as a replacement for
 153 physics-based numerical reservoir simulations, or high-fidelity simulations (HFS), in GCS projects
 154 as an image-to-video mapping operator.

155 In the methodology section, we describe the governing equations of multiphase flow in GCS,
 156 and the proposed spatiotemporal proxy model architecture. In the results and discussion sections,
 157 we describe the geologic modeling and numerical reservoir simulation steps required to generate
 158 the training data, and evaluate the training and performance of the proposed proxy model and
 159 compare its efficiency and accuracy to high-fidelity numerical simulations using a 2D synthetic case
 160 for large-scale GCS operations.

161 **2 Methodology**

162 This section describes the governing equations, and the architecture and training strategy of the
 163 Stochastic pix2vid model.

164 **2.1 Governing equations**

165 For the CO₂-water multiphase flow problem, the general form of the mass accumulation for com-
 166 ponent $\kappa = \text{CO}_2$ or water is given by [74]:

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

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

$$M^\kappa = \phi \sum_p S_p \rho_p X_p^\kappa \quad (2)$$

168 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
169 mass fraction of component κ present in phase p . For each component κ , there is also the advective
170 mass flux $F^\kappa|_{adv}$ obtained by summing over all phases p ,

$$F^\kappa|_{adv} = \sum_p X_p^\kappa F_p \quad (3)$$

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

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

172 where u_p is the Darcy velocity of phase p , k is the absolute permeability, $k_{r,p}$ is the relative
173 permeability of phase p , μ_p is the viscosity of phase p , and g is the gravitational acceleration
174 constant. The relative permeability curves for the CO₂-water system are shown in Figure 4. The
175 fluid pressure of phase p ,

$$P_p = P + P_c \quad (5)$$

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

179 2.2 Proxy Model Architecture

180 Our proposed Stochastic pix2vid image-to-video data-driven method, is mapping operator from
181 the static realizations of geologic distributions of porosity, permeability and facies as well as the
182 injector well(s) distribution, to the dynamic responses of pressure and saturation distributions over
183 time.

184 Let m represent a geologic model realization of porosity, permeability, facies, and injector well(s)

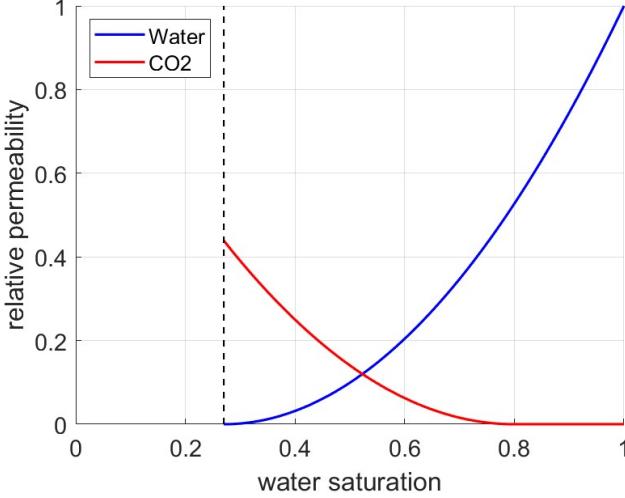


Figure 4: Relative permeability curves for the CO₂-water system. The residual saturations are 0.27 and 0.2 for water and CO₂, respectively.

185 distributions, such that $m = \{\phi, k, facies, w\}$. The dynamic responses of pressure and saturation
 186 over time are given by $d = f(m)$, such that $d = \{P(t), S(t)\}$ and f is the physics-based numerical
 187 reservoir simulation. Our aim is to replace f with a more efficient data-driven proxy by training the
 188 Stochastic pix2vid model, which is trained as a single model to predict both pressure and saturation
 189 distributions over time as a multi-channel output from the multi-channel input features, m . For
 190 this purpose, we exploit the latent structures in space and time of the static inputs and dynamic
 191 outputs through a spatiotemporal encoder-decoder architecture.

192 The encoder portion of the network is comprised of sequential convolutional layers to com-
 193 press the spatial features of the subsurface realizations into a latent parameterization z_m , given by
 194 $z_m = Enc(m)$. In their compressed representation, these features represent the salient character-
 195 istics of the geologic distributions. The decoder portion of the network is designed as a series of
 196 recursive residual convolutional-recurrent layers, such that the latent space z_m is recursively de-
 197 coded into the dynamic distributions of pressure and saturation over time. The previous timestep
 198 latent representations, z_d^t , are used in the subsequent timesteps of the decoder, such that the subse-
 199 quent timesteps will predict the current and previous timestep(s) jointly and iteratively, providing
 200 a reduction of systematic error in time as subsequent frames of the dynamic output video are
 201 predicted. The full architecture is represented as:

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

202 The encoder, $Enc(\cdot)$, compresses the geologic realizations, m , into a latent representation z_m
 203 through the use of depthwise separable convolutions [75]. This type of convolution learns the
 204 parameters for each channel in the image separately, avoiding mixing of variables or loss of reso-
 205 lution, as shown in Figure 5. This is especially important when dealing with discrete, non-smooth
 206 porosity and permeability spatial distributions due to discrete facies and binary well(s) location
 207 distributions. Each separable convolution layer is regularized with an l_1 -norm weight of 1×10^{-6} .
 208 Moreover, we use a Squeeze-and-Excite layer to improve channel interdependence, and to avoid
 209 mixing and loss of resolution [76]. Each Squeeze-and-Excite layer will provide the optimal network
 210 weights for each channel independent of the other channels by passing the feature maps through a
 211 global pooling layer (squeeze) and a dense layer with nonlinear activation (excite), to add content
 212 aware mechanism for re-weighting each channel adaptively, as shown in Figure 6. Furthermore, by
 213 applying instance normalization, as opposed to the more common batch normalization, we achieve
 214 channel-independent normalization of the convolved features [77]. Instance normalization is a spe-
 215 cial case of group normalization, where the numbers of channels per group is exactly 1, such that
 216 each channels gets its own normalization scheme, as shown in Figure 7. Parametric rectified linear
 217 units (PReLU) is used as the activation function, where at each minibatch iteration, the network
 218 learns the optimal leaky slope for activation in each layer, as shown in Figure 8. Finally, pooling
 219 and spatial dropout are applied to reduce in half the input dimension of each layer and to pro-
 220 vide a means of spatial regularization, respectively. Through 3 convolutional encoding layers with
 221 filter size 3×3 , we obtain the latent parameterizations z_m^1 , z_m^2 , and z_m^3 . Table 1 summarizes the
 222 architecture and dimensions of each layer.

223 Step 1: **Depthwise Separable encoding:** The first layer of Enc takes the geologic model realiza-
 224 tion, m , and computes the depthwise separable convolutional features channel-by-channel.

225 Step 2: **Squeeze-and-Excite encoding:** By taking the channel-wise global average of the feature
 226 space from Step 1, a fully-connected layer predicts the appropriate weighting coefficients
 227 to best parameterize the features.

228 Step 3: **Instance Normalization of the feature space:** Feature normalization is applied on a
 229 channel-by-channels basis for each batch of the encoded feature space, avoiding mixing and
 230 blurring.

231 Step 4: **Activation, Pooling, and Spatial Dropout:** The PReLU nonlinear activation function
 232 is used, and for each batch, an optimal leaky slope is learned. Pooling is used to reduce the
 233 feature space in half, and Spatial Dropout of 5% is used to regularize the learning process
 234 and increase robustness in prediction.

235 Step 5: **Final Encoding and Repeat:** From Steps 1-4, the geologic model realization m is en-
 236 coded into a latent representation z_m^k . We repeat Steps 1-4 three times to obtain three
 237 intermediate latent representations, namely z_m^1 , z_m^2 , and z_m^3 .

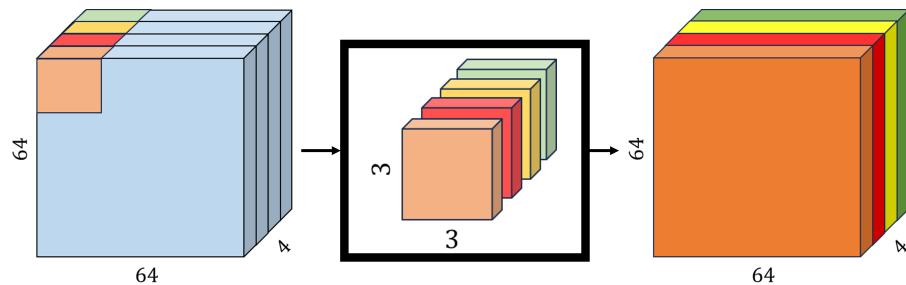


Figure 5: Schematic for a separable convolutional layer. Each channel is convolved with its own set of convolutional filters to obtain the best representation, as opposed to traditional convolutions where the same filter is applied to all channels in the data.

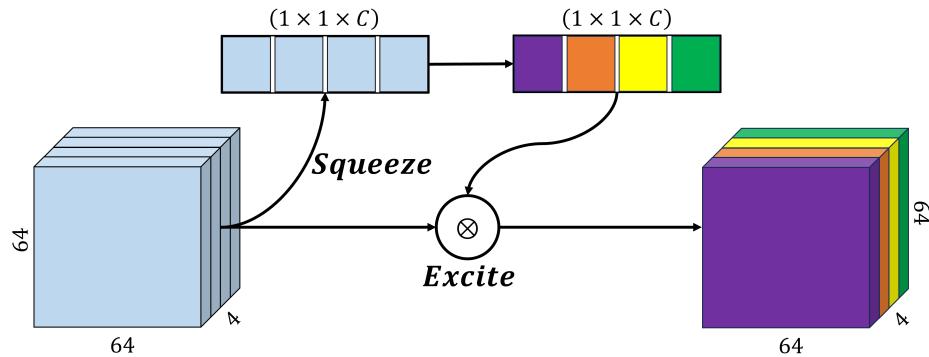


Figure 6: Schematic for a squeeze-and-excite layer. The "squeeze" layer takes the global average of the data for each channel, and the "excite" layer is a fully-connected layer with nonlinear activation to estimate the optimal weights for each channel in the data. The result is a weighted representation of the data based on their intrinsic global characteristics.

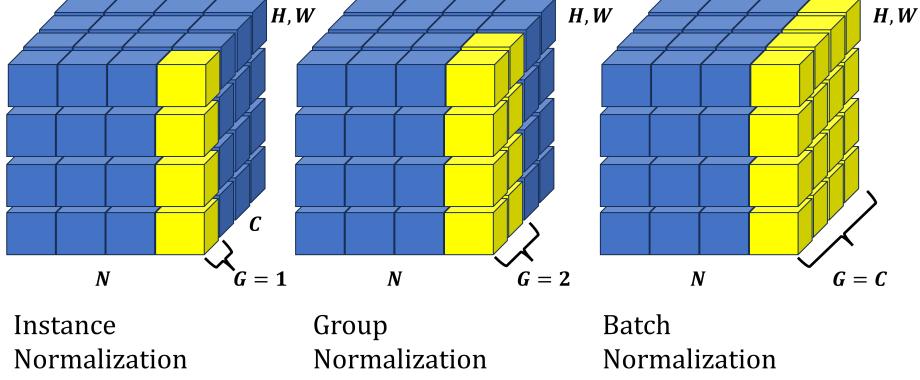


Figure 7: Schematic for instance normalization (left) compared to group normalization (center) and batch normalization (right). In an instance normalization layer, each channel will be normalized by themselves rather than normalizing the entire batch or a subset of channels (groups).

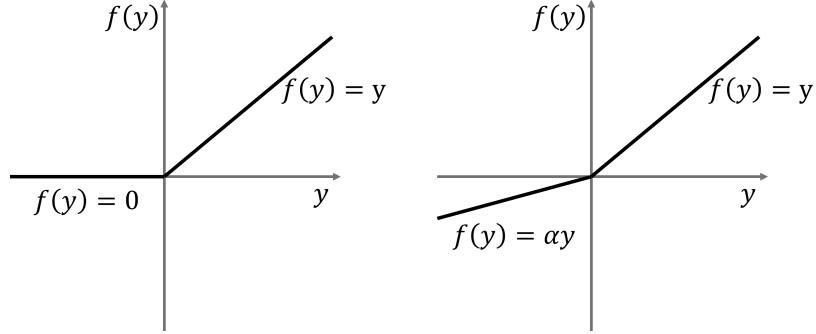


Figure 8: Schematic for the Parametric Rectified Linear Unit (PReLU) activation function (right) compared to the traditional ReLU activation function (left). The slope of the negative portion of the data, α , is learned for each batch.

238 The decoder, $Dec^t(\cdot)$, of the Stochastic pix2vid model extracts the spatiotemporal relationships
 239 from the latent representations of m to reconstruct the dynamic pressure and saturation distribu-
 240 tions over time, d . To accurately reconstruct the spatiotemporal structure from the static latent
 241 space, z_m , we employ a series of convolutional-recurrent layers, namely a convolutional long-short
 242 term memory layer (ConvLSTM). The general form of a 2D ConvLSTM layer is shown in Figure
 243 9. Through 3 convolutional-recurrent layers, we obtain the dynamic prediction of \hat{d} as follows:

244 Step 6: **Spatiotemporal decoding of z_m^3 :** The first ConvLSTM layer takes the smallest latent
 245 representation, z_m^3 , and reconstructs the first decoded timestep z_d^3 .

246 Step 7: **Residual concatenation of z_m^2 :** The first decoded timestep, z_d^3 , is concatenated with
 247 the intermediate static encoding z_m^2 to retain multi-scale features and improve prediction

Table 1: Encoder network architecture

Layer Number	Architecture	Shape in (h,w,c)	Shape out (h,w,c)
1	SeparableConv2D	$64 \times 64 \times 4 (m)$	
	Squeeze-and-Excite		
	Instance Norm		
	PReLU + Pooling		
2	Spatial Dropout		$32 \times 32 \times 64 (z_m^1)$
	SeparableConv2D	$32 \times 32 \times 64$	
	Squeeze-and-Excite		
	Instance Norm		
3	PReLU + Pooling		
	Spatial Dropout		$16 \times 16 \times 128 (z_m^2)$
	SeparableConv2D	$16 \times 16 \times 128$	
	Squeeze-and-Excite		
	Instance Norm		
	PReLU + Pooling		
	Spatial Dropout		$8 \times 8 \times 256 (z_m^3)$

248 performance and resolution.

249 Step 8: **Intermediate spatiotemporal decoding:** The second ConvLSTM layer takes the resi-
 250 dential concatenation of the intermediate latent representations, $[z_m^2, z_d^3]$, to predict the next
 251 spatiotemporal representation z_d^2 .

252 Step 9: **Residual concatenation of z_m^1 :** The intermediate decoded timestep, z_d^2 , is concatenated
 253 with the largest static encoding z_m^1 .

254 Step 10: **Final spatiotemporal decoding:** The third ConvLSTM layer takes the residual con-
 255 catenation of the larger latent representations, $[z_m^1, z_d^2]$, to predict the full-scale dynamic
 256 output, d .

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

262 This process yields the first video frame prediction, d_1 , from the latent representation of the

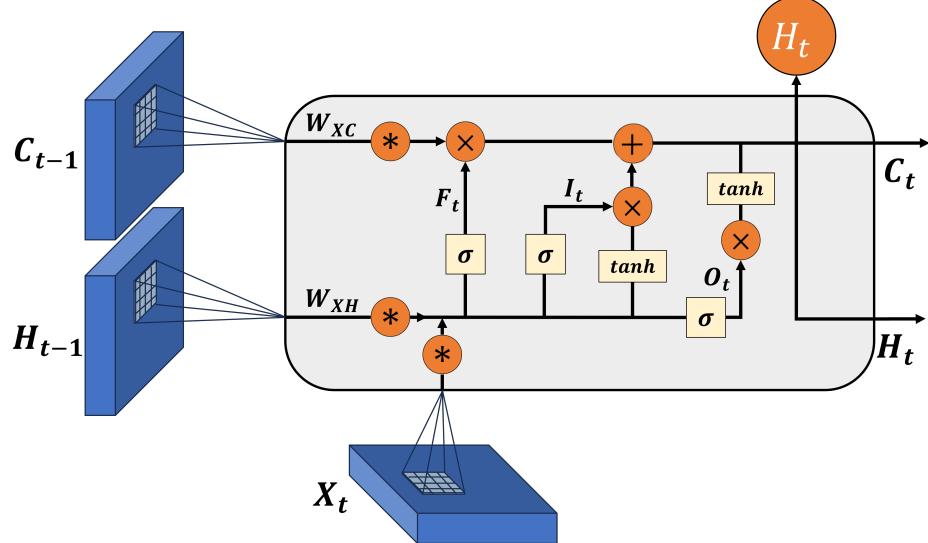


Figure 9: Schematic of a convolutional-LSTM (ConvLSTM) layer. The layer applies convolutional operations to the input data using a set of learnable filters to capture the spatial patterns. The recurrent part is a long short-term memory layer with memory and forget gates to capture the temporal patterns. LSTM units are applied to each spatial location separately allowing to capture both spatial and temporal dependencies in the data.

263 geologic realizations z_m . Each subsequent video frame prediction is obtained by another set of
 264 residual concatenation of the previous timestep dynamic decoded representation. The static latent
 265 representation z_m is concatenated at each timestep with the previous dynamic decoded represen-
 266 tation for each layer such that we have $[z_m, z_{d_t}^i]$, where i is the decoding layer number and t is the
 267 timestep. By recursively implementing spatiotemporal decoding to the latent representation z_m ,
 268 we obtain the prediction of the dynamic response d_t at times for each timestep $t = 1, \dots, n$.

269 The complete Stochastic pix2vid architecture is shown in Figure 10. Here we observe the
 270 spatial compression of the geologic models, m , through the encoding portion of the network, and
 271 the spatiotemporal decoding and residual multi-scale concatenations through the decoder portion of
 272 the network. The resulting architecture provides proxy model from a subsurface static uncertainty
 273 model (images) to subsurface dynamic response (videos).

274 2.3 Training Strategy

275 The inputs to the Stochastic pix2vid are the geologic realizations, comprised of the distributions of
 276 porosity, permeability, facies, and injection well(s) location, represented as a matrix m of dimensions
 277 $64 \times 64 \times 4$. The outputs are the results from the numerical reservoir simulation, namely pressure

Table 2: Decoder network architecture

Layer Number	Architecture	Shape in (t,h,w,c)	Shape out (t,h,w,c)
1	ConvLSTM2D	$1 \times 8 \times 8 \times 256$	
	BatchNorm + LeakyReLU		
	Conv2DTranspose		
	Spatial Dropout		
	Concatenate (z_m^3)		
2	Conv2D + Sigmoid		$t \times 16 \times 16 \times 128 (z_{d_t}^3)$
	ConvLSTM2D	$t \times 16 \times 16 \times 128$	
	BatchNorm + LeakyReLU		
	Conv2DTranspose		
	Spatial Dropout		
3	Concatenate (z_m^2)		
	Conv2D + Sigmoid		$t \times 32 \times 32 \times 64 (z_{d_t}^2)$
	ConvLSTM2D	$t \times 32 \times 32 \times 64$	
	BatchNorm + LeakyReLU		
	Conv2DTranspose		
278	Spatial Dropout		
	Concatenate (z_m^1)		
	Conv2D + Sigmoid		$t \times 64 \times 64 \times 2 (z_{d_t}^1)$

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 are difficult to resolve [78, 79]. 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; therefore the output matrices, (d, \hat{d}) , have a final dimension of $64 \times 64 \times 11 \times 2$. This is done to make the problem more tractable and speed up the training and prediction process, while retaining majority of the temporal information.

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° image rotation, making the network agnostic to orientation and encourage 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 training set minibatch is further split into a training and validation subset using an 80/20 split. The validation set is only used to adjust

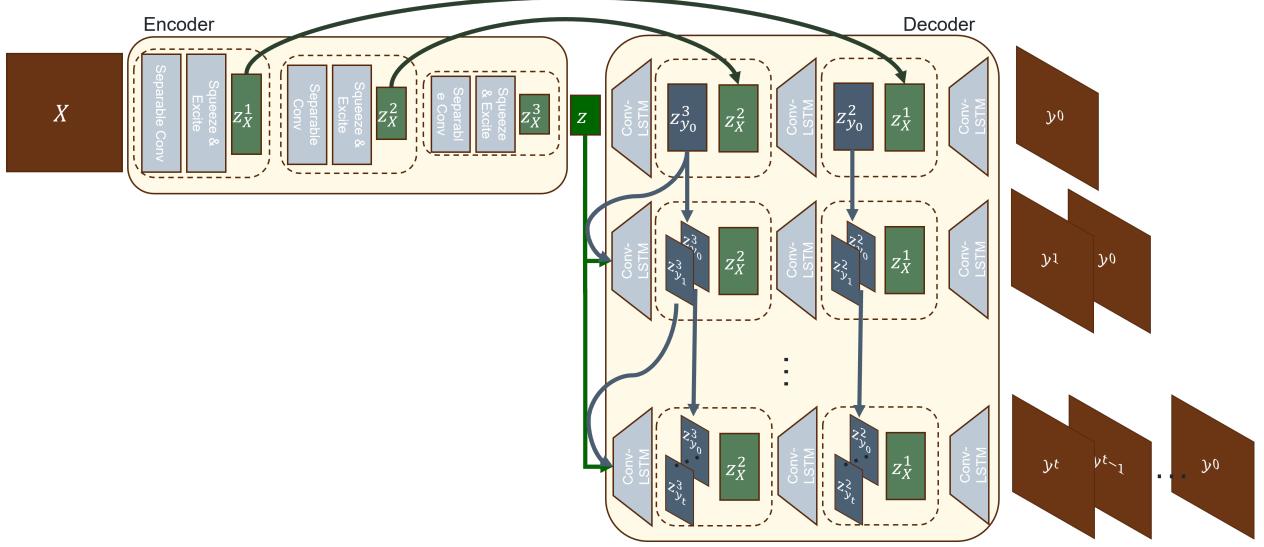


Figure 10: Architecture of our proposed Stochastic pix2vid method. The input data, $X \equiv m$, is encoded through a series of convolutional layers to capture the spatial dependencies in the geologic models. The latent representation, z_m , is recursively passed through a spatiotemporal decoder with convolutional-recurrent layers, and concatenated with the residuals of the encoder to reconstruct iteratively the frames of the output (video) data, $y \equiv d$.

293 the trainable model parameters for each batch at each epoch and is randomly partitioned from
 294 the training batch at every epoch, while the testing data remains unseen to quantify the model
 295 performance after training.

296 A custom three-part loss function is used to accurately predict pixel-wise and perceptual in-
 297 formation in the predictions. The mean squared error (MSE) is used to reconstruct the pixel-
 298 wise intensity values, while the mean absolute error (MAE) is used to optimize for the pressure
 299 and saturation plume edges. The third part is the structural similarity index metric (SSIM),
 300 which provides a perceptual image-to-image comparison of luminance, contrast, and structure
 301 [80]. For optimal training, the aim is to minimize the MSE and MAE while maximizing the
 302 SSIM for the true versus predicted outputs, d and \hat{d} , such that the total loss is given by $\mathcal{L} =$
 303 $\alpha(1 - SSIM) + (1 - \alpha)[\beta MSE + (1 - \beta)MAE]$, where α and β are weighting coefficients obtained
 304 empirically as 0.33 and 0.66, respectively.

305 The model is trained using the AdamW optimizer [81]. This variant of the well-known adaptive
 306 momentum (Adam) optimizer [82] includes an added method to decay weights for the adaptive
 307 estimation of first-order and second-order moments. We implement a learning rate of 1×10^{-3} with
 308 a weight decay term of 1×10^{-5} .

309 **3 Results**

310 This section describes the geologic model generation, training performance and discusses the ap-
 311 plication of the Stochastic pix2vid proxy to rapidly forecast CO₂ plume migration for a large-scale
 312 GCS operation.

313 **3.1 Reservoir Model and Simulation**

314 We use SGeMS [83] to construct the subsurface uncertainty model, an ensemble of static feature
 315 realiations that is representative of various potential geologic scenarios for CO₂ storage. Using
 316 sequential Gaussian co-simulation [84], we generate a set of 1,000 random porosity (ϕ) and perme-
 317 ability (k) distributions with a wide range of values, as shown in Figure 11. Facies distributions are
 318 obtained from a library of deepwater fluvial training images [85, 86]. These encompass a wide range
 319 of possible geologic scenarios including marked point (lobe, ellipse, and bar), FluvSim (channel,
 320 channel-levee, and channel-levee-splay), surface based (compensational cycles of lobes), and bank
 321 retreat (channel complex). To generate consistent porosity and permeability distributions with the
 322 facies-based geologic scenarios, we condition the original porosity and permeability distributions to
 323 the facies distributions. The resulting fluvial distributions are shown in Figure 12.

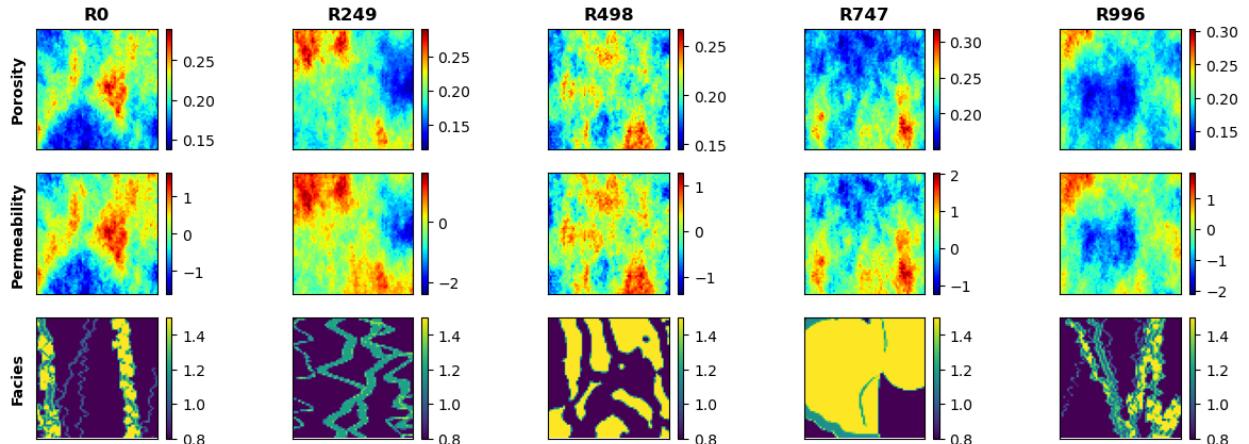


Figure 11: Spatial distribution of porosity (top), permeability (middle), and facies (bottom) for 5 random realizations.

324 The model has dimensions of 1km-1km-100m in the x-, y-, and z-directions, respectively. We use
 325 64 uniform grid cells in the x- and y-directions. The grid design is sufficiently refined to resolve the
 326 pressure and saturation plumes in highly heterogeneous reservoirs while remaining computationally

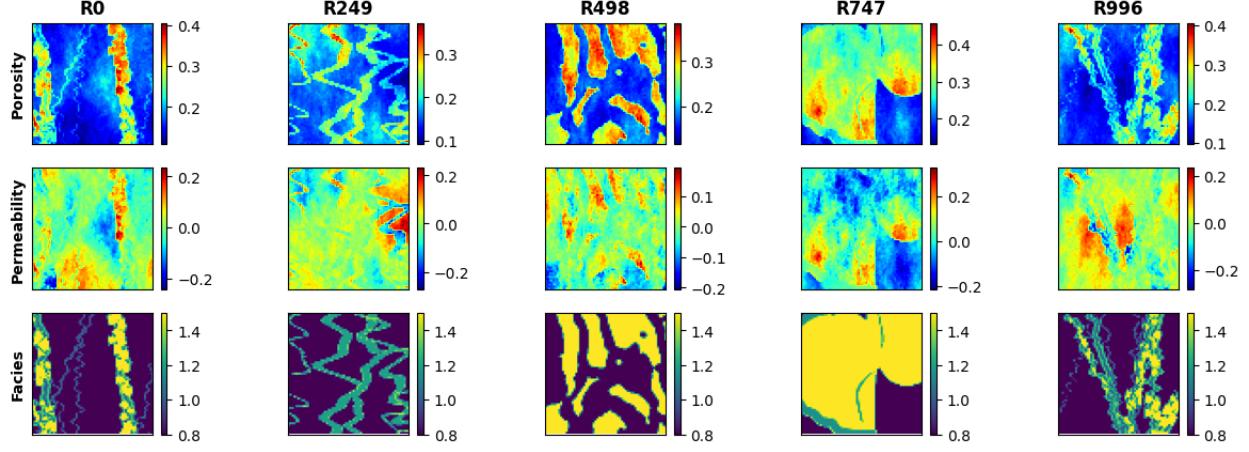


Figure 12: Spatial distribution conditioned to facies (top) for porosity (middle) and permeability (bottom) for 5 random realizations.

327 tractable for the purpose of training deep learning models. A random number of injection wells,
 328 $w \in [1, 3]$, are placed randomly along the reservoir for each of the 1,000 realizations, no closer than
 329 250m from the boundaries, as shown in Figure 13. The injection well(s) are randomly placed and
 330 not conditioned to zones of preferential porosity, permeability, nor facies. Each injection well has
 331 a constant radius of 0.1m and a single and continuous perforation that injects pure supercritical
 332 CO₂ at a constant rate such that the total injection rate of the w well(s) is 0.5 megatons per year.

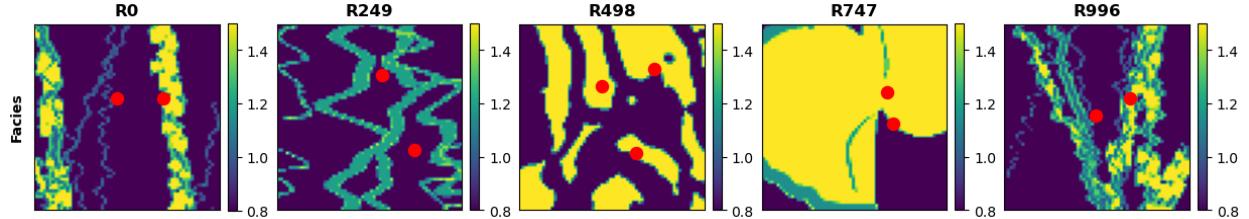


Figure 13: CO₂ injection well(s) location (red) overlaid over facies distributions for 5 random realizations.

333 The conditional fluvial porosity and permeability distributions are used as input models for the
 334 numerical simulation of geologic CO₂ storage using MRST [87] to calculate the response models
 335 for training our proposed model. The reservoir is initialized as a fully water saturated zone (i.e.,
 336 aquifer) with an initial pressure of 4,000 psi. The reservoir has constant isothermal conditions and
 337 constant pressure boundary conditions to represents a large-scale geologic CO₂ storage project with
 338 negligible dip, such as found in the Illinois Basin and parts of the North Sea and Gulf of Mexico.

339 The numerical simulation is run for 5 years, monitored monthly, for a total of 60 timesteps. At
 340 each grid cell and for each time step, we resolve the implicit pressure, explicit saturation (IMPES)
 341 formulation of Eq. (1) to obtain the corresponding dynamic pressure and saturation distributions
 342 over time (videos) from the static geologic realizations of porosity and permeability conditioned to
 343 the fluvial facies (images) with random well(s) configuration. The pressure and saturation responses
 344 corresponding to the geologic model realizations are shown in Figures 14 and 15, respectively.

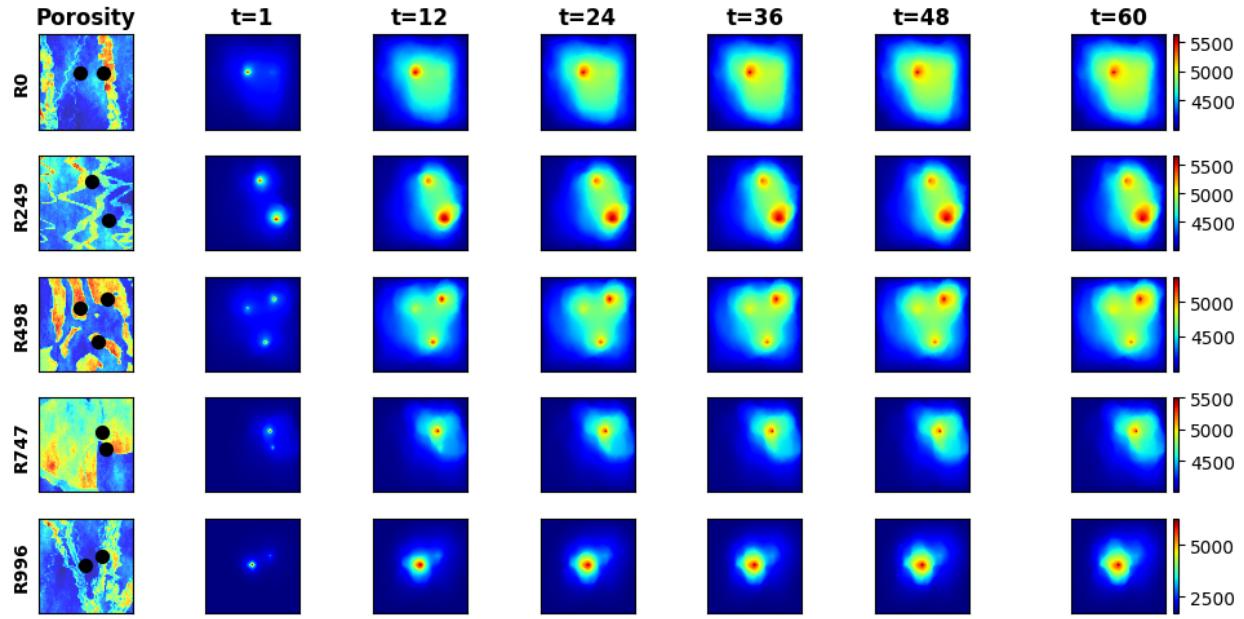


Figure 14: Pressure response distributions over time (in psia) obtained by HFS for the 5 random realizations from Fig. 12.

345 3.2 Training Performance

346 Using an NVIDIA Quadro M6000 GPU, we train for 100 epochs with batch size of 50. The model
 347 has in total 97,523,370 parameters, and the training time required is approximately 88 minutes
 348 for all 1,500 training realizations. The training and validation performance per epoch is shown
 349 in Figure 16. We observe minimal overfit in the validation set, corresponding to good model
 350 generalizability and prediction accuracy within the training data. Using physics-based numerical
 351 simulation, each realization requires approximately 30 seconds to obtain the dynamic pressure and
 352 saturation predictions from the static geologic models. Our Stochastic pix2vid model obtains the
 353 same results in approximately 4.59 milliseconds, corresponding to a 6,500 \times speedup. The average

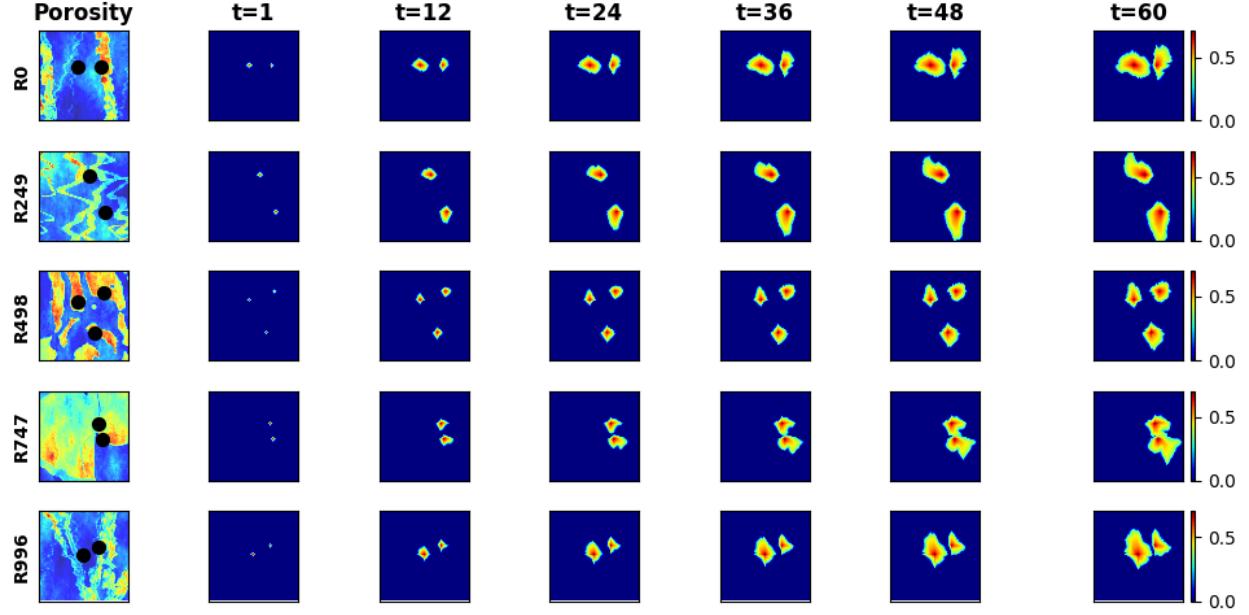


Figure 15: Saturation response distributions over time obtained by HFS for the 5 random realizations obtained from Fig. 12.

354 MSE for the ensemble is 9.21×10^{-4} and 9.70×10^{-4} for training and testing, respectively. Similarly,
 355 the average SSIM for the ensemble is 98.97% and 97.91% for training and testing, respectively.

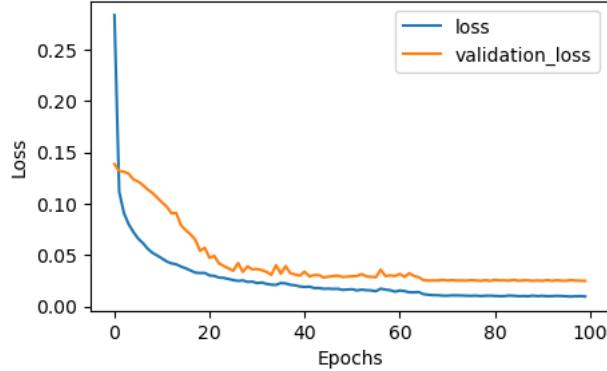


Figure 16: The total training and validation losses, \mathcal{L} , as a function of epoch number.

356 3.3 Prediction Results

357 After training the Stochastic pix2vid model with 1,500 realizations of static geologic models, $m =$
 358 $\{\phi, k, facies, w\}$, to predict the dynamic reservoir response, $d = \{P(t), S(t)\}$, we can compare the
 359 performance of the predictions for the training and unseen testing data.

360 Figures 17 and 18 show the predicted dynamic pressure and saturation distributions, respec-
361 tively, along with the absolute difference to HFS for 3 training realizations. We observe reasonable
362 agreement between the true and predicted CO₂ pressure and saturation plumes over time, pixel-wise
363 with an average MSE of 3.25×10^{-4} and perceptually with SSIM of 98.59% for pressure predictions
364 and MSE of 1.50×10^{-4} and SSIM of 97.31% for saturation predictions.

365 Similarly, Figures 19 and 20 show the pressure and saturation distributions predictions along
366 with the absolute difference to HFS for 3 testing realizations. We observe a similar performance,
367 with an average MSE of 3.71×10^{-4} and SSIM of 97.55% for pressure predictions and MSE of
368 1.61×10^{-3} and SSIM of 96.19% for saturation predictions. This indicates that the Stochastic
369 pix2vid model is generalizable and achieves on par performance with HFS at a fraction of the
370 computational cost.

371 It is interesting to note that the Stochastic pix2vid model is trained on a triple-loss function
372 with MSE, MAE and SSIM. For both training and testing cases, we see that the average MSE for
373 pressure is higher than that of saturation, while the opposite is true for the average SSIM. This
374 can be attributed to the fact that there are more pixel-wise variations in pressure predictions, thus
375 the loss focuses on matching those individual pixel-wise values. On the other hand, for saturation
376 predictions, the contrast, luminance, and structure play a bigger role in the prediction than the
377 pixel-wise intensity values. Therefore, it is important to take into account both metrics for training
378 and validating spatiotemporal subsurface prediction models.

379 These results imply that our Stochastic pix2vid is capable of learning the spatiotemporal re-
380 lationship between the static geologic models and the dynamic reservoir response. Thus, our
381 image-to-video architecture can outperform current image-to-image and encoder-recurrent-decoder
382 architectures for improved reservoir behavior prediction. A comparison of true versus predicted
383 results for pressure and saturation responses for the testing data is shown in Figure 22. For the
384 pressure and saturation predictions, the average R^2 over time is approximately 99% with narrow
385 95% prediction bands that recursively narrow over time. From Figure 22 we observe the Stochastic
386 pix2vid model’s performance at recursively refining the predictions over time due to the residual
387 connections in the spatiotemporal decoder network.

388 From Section 2.2, the first step of the Stochastic pix2vid model is to take the static geologic
389 realizations, m , and compresses them into a latent space representation, z_m , using the spatial

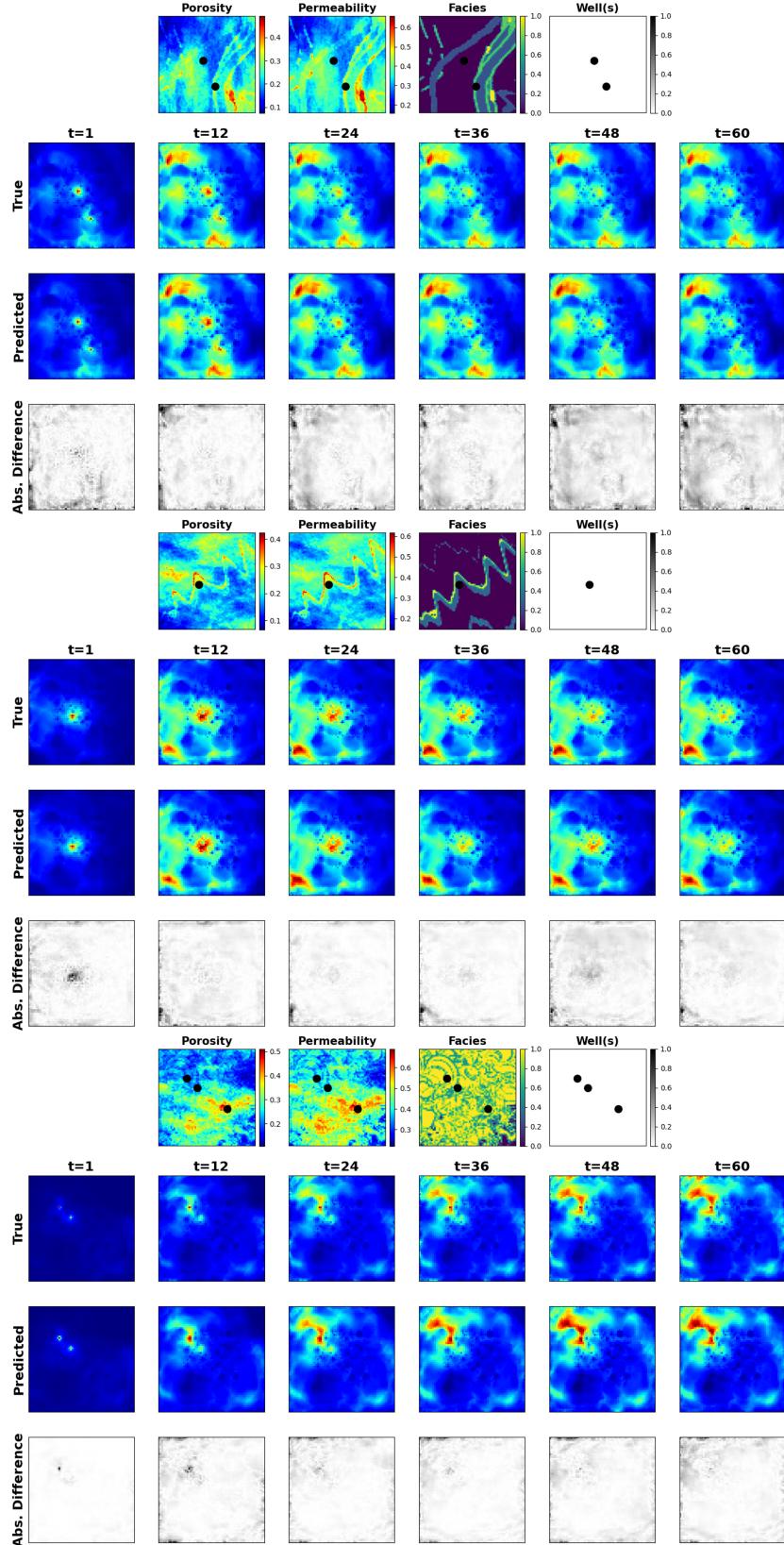


Figure 17: Normalized pressure distribution over time for 3 random training realization. For each panel, the top row is the ground truth from the HFS, the middle row is the Stochastic pix2vid prediction, and the bottom row is the absolute difference to HFS.

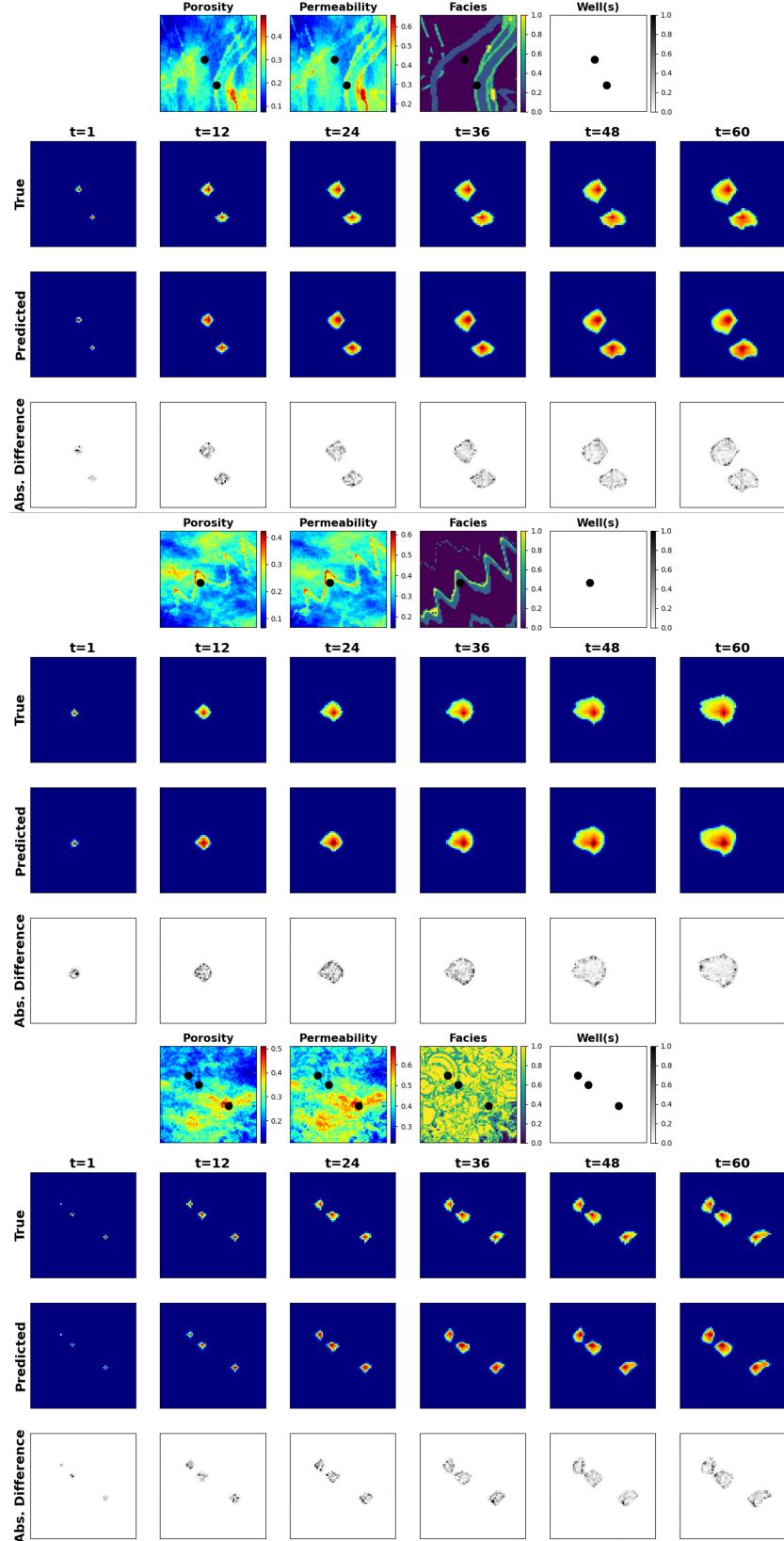


Figure 18: Saturation distribution over time for 3 random training realization. For each panel, the top row is the ground truth from the HFS, the middle row is the Stochastic pix2vid prediction, and the bottom row is the absolute difference to HFS.

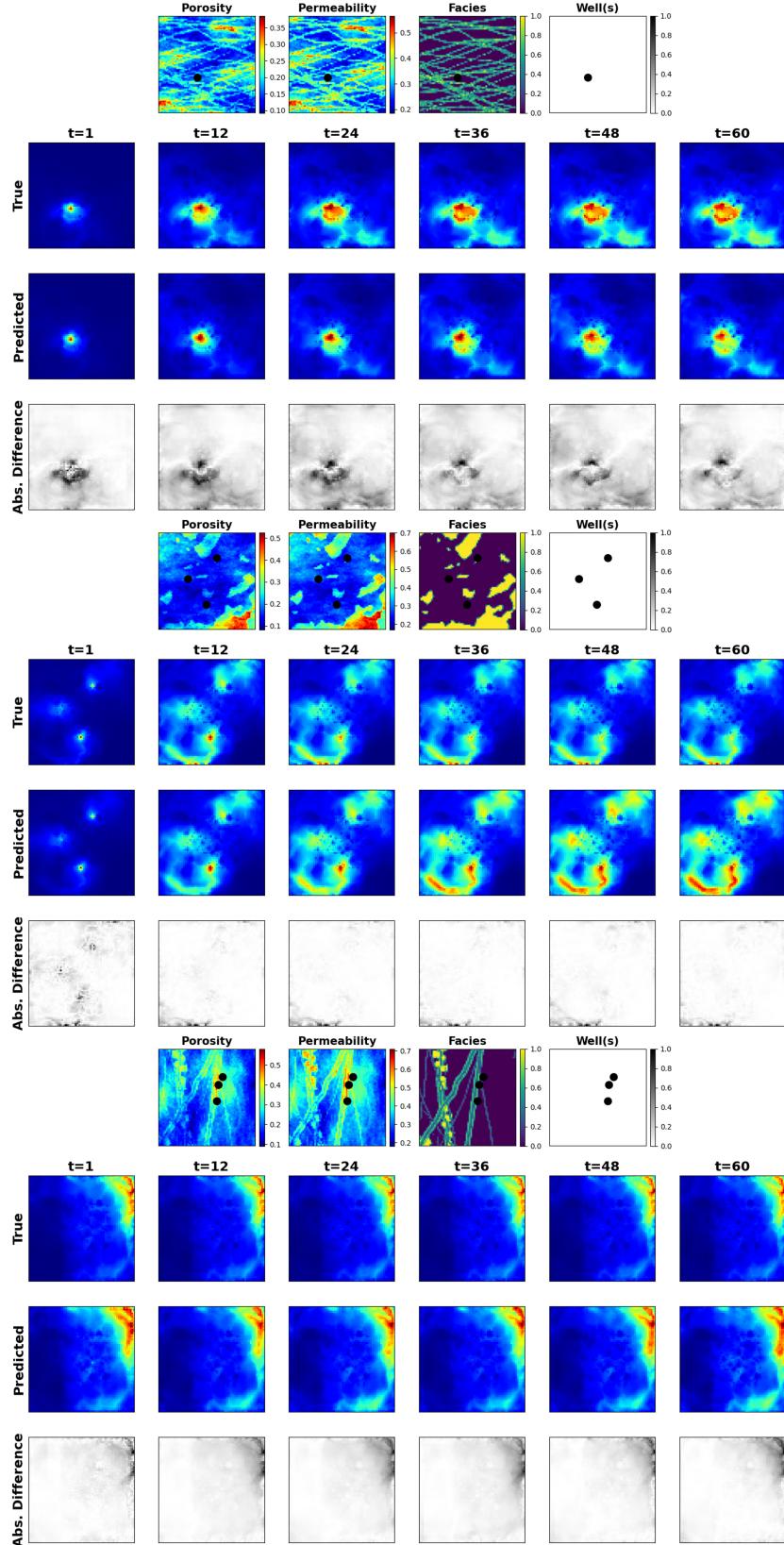


Figure 19: Normalized pressure distribution over time for 3 random testing realization. For each panel, the top row is the ground truth from the HFS, the middle row is the Stochastic pix2vid prediction, and the bottom row is the absolute difference to HFS.

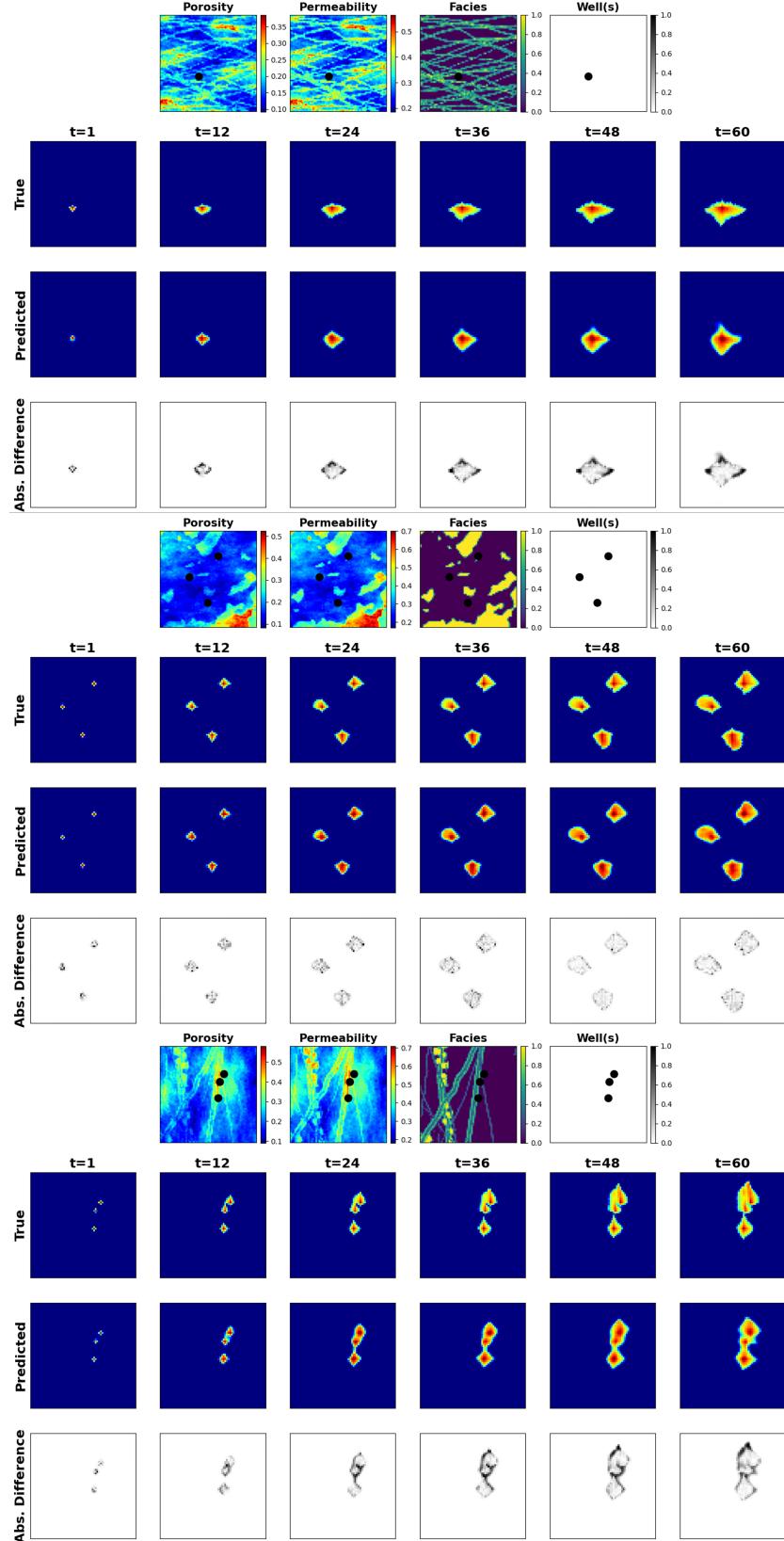


Figure 20: Saturation distribution over time for 3 random testing realization. For each panel, the top row is the ground truth from the HFS, the middle row is the Stochastic pix2vid prediction, and the bottom row is the absolute difference to HFS.

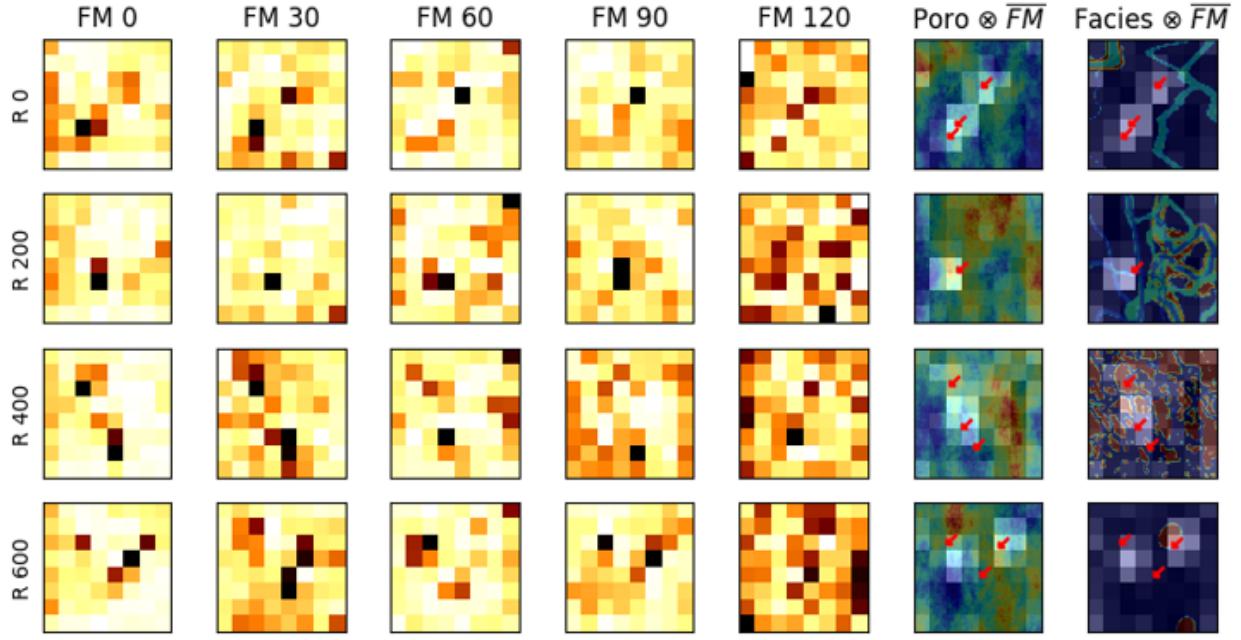


Figure 21: Five random feature maps (FM) of z_m^3 for 4 random realizations. Their average is superimposed on top of the porosity and facies distributions to show the attention mechanism of the encoder. Bright colors represent higher attention and dark colors represent lower attention.

encoder structure. Figure 21 show a random selection of latent feature maps, along with their superposition on the porosity and facies distribution. This can be interpreted as an analog to the attention head mechanisms recently developed in transformer-based architectures [88]. 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.

These results imply 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 to provide improved reservoir behavior prediction closer to that of traditional numerical simulation. To quantify the uncertainty in predictions, a comparison of true (d) versus predicted (\hat{d}) response for pressure and saturation distributions for the testing data is shown in Figure 22. The average R^2 over time is approximately 99% with narrow 95% prediction bands that recursively narrow over time. From Figure 22 we observe the advantage in implementing recursive refining of predictions over time with recurrent residual connections in the spatiotemporal decoder

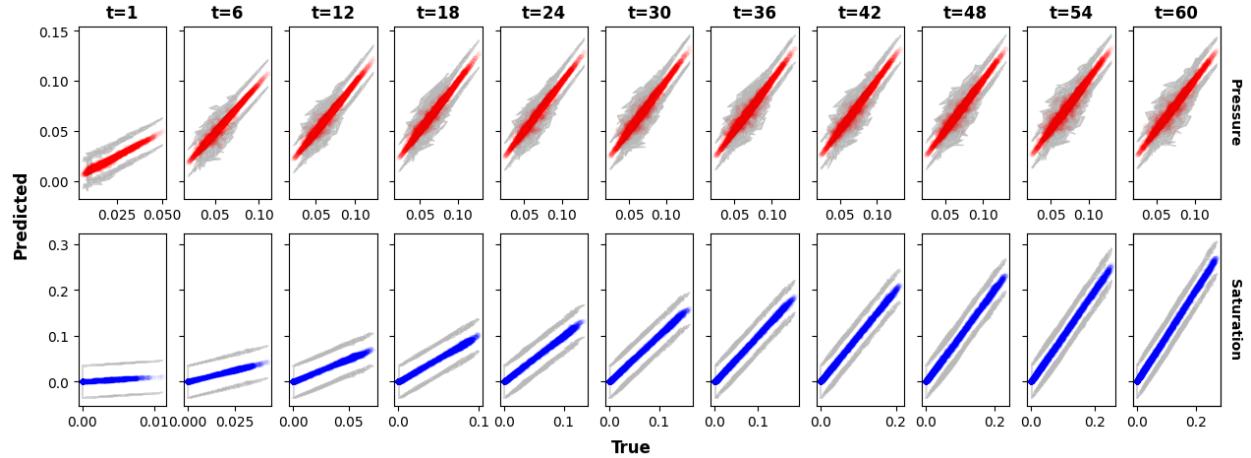


Figure 22: True versus predicted average normalized pressure (top) and saturation (bottom) over time for the testing data. The gray portion represents the 95% confidence bands, which narrow over time.

405 network, thus reducing the spatiotemporal uncertainty in the predictions.

406 CO₂ saturation and pressure buildup fronts are important quantities for geologic CO₂ storage
 407 projects and are often used for regulatory oversight [89, 90], monitoring metrics or history matching
 408 purposes [91, 92]. The distance between the injection well(s) and the saturation fronts represents
 409 the maximum extent of the CO₂ plume; however, these are often very difficult to capture accurately
 410 with data-driven proxy models. Our Stochastic pix2vid method shows greater absolute error on and
 411 around the plume fronts compared to within the plumes. However, the overall shape and intensity of
 412 the pressure and saturation distributions over time is very well captured for all realizations despite
 413 being highly heterogeneous. Therefore, the Stochastic pix2vid model can be used as a reliable
 414 replacement for expensive numerical reservoir simulations, especially in cases where large number
 415 of runs are required to obtain dynamic estimates (e.g., well placement and control optimization,
 416 history matching, uncertainty quantification).

417 3.4 Discussion

418 In our Stochastic pix2vid model, the encoder block is composed of separable convolutions, squeeze
 419 and excite layers, and instance normalization. These three particular implementations allow for
 420 precise parameterization of the geologic realization into a latent representation, without mixing
 421 the effects of Gaussian-distributed properties against binary or binomial-distributed properties.

422 Using recursive residual ConvLSTM layers, the decoder block iteratively predicts each dynamic
 423 state, or video frame, from the concatenation of the previous dynamic latent representation and
 424 the intermediate encoding parameterizations. Thus, our architecture makes the proxy model an
 425 image-to-video prediction formulation for dynamic reservoir states from a static geologic realization.

426 To further demonstrate the effectiveness of our Stochastic pix2vid model for geologic CO₂ stor-
 427 age operations, we plot the cumulative pixel-wise CO₂ saturation as a surrogate for the cumulative
 428 CO₂ volume injected. For all training and testing realizations, Figure 23 shows the sum of pixel-wise
 429 CO₂ saturation and the probability density function (PDF) of the true versus predicted saturations.
 430 We observe an R^2 of 98% for training and 96% for testing in the cumulative CO₂ saturation of true
 431 versus predicted results, and a conformable PDFs for both training and testing.

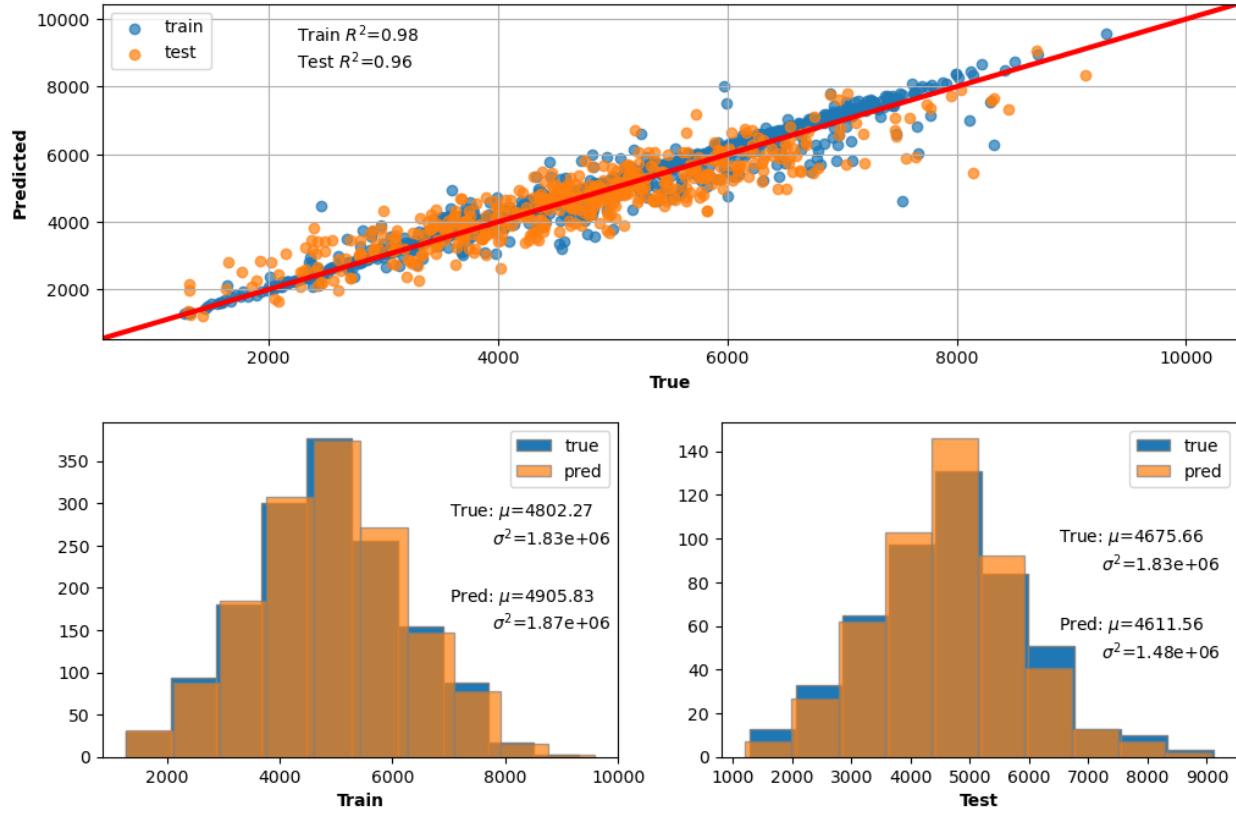


Figure 23: (Top) True vs. predicted cumulative CO₂ volume injected via pixel-wise saturation. (Bottom) True vs. predicted distributions of cumulative CO₂ saturation for training (left) and testing (right).

432 Our Stochastic pix2vid method has several limitations. In order to learn the spatiotemporal
 433 relationships between input images and output videos, the model requires substantial amounts of

434 training data, which in turn require expensive physics-based numerical simulation runs. Moreover,
435 the method would require retraining in order to apply to a different subsurface flow and transport
436 problem, increasing the time required for generating the training data and the time required to
437 retrain the model. One major limitation is the inability to predict for timesteps beyond those
438 present in the training data. The architecture of the Stochastic pix2vid is designed to reconstruct
439 only the 11 timesteps present in d , therefore it is capable of interpolation for steps in between the
440 training timesteps, but incapable to forecast beyond $t = 5$ years (60 months). Lastly, the method
441 is designed for images at the resolution of 64×64 pixels, and preprocessing is required to reshape
442 training data of other dimensions to this size.

443 4 Conclusions

444 We develop a deep learning-based spatiotemporal proxy model to provide efficient flow predictions
445 for a large-scale GCS operations to support optimum decision making. Our proposed method,
446 Stochastic pix2vid, introduces the use of a spatiotemporal convolutional-recurrent architecture for
447 dynamic predictions of CO₂ pressure and saturation distributions over time from a static geo-
448 logic realization representing the subsurface uncertainty model. The framework is developed as
449 an image-to-video prediction, which is an under-determined estimation problem. Specifically, the
450 implementation expands upon the architectures of current encoder-recurrent-decoder models and
451 provides a fast and accurate proxy as a replacement for physics-based numerical reservoir simula-
452 tion.

453 The spatiotemporal proxy is applied to a synthetic 2D GCS project with multiple uncertain
454 geologic scenarios and random number and location of injection well(s). A total of 1,000 geologic
455 models are obtained from a variety of possible geologic scenarios including fluvial, turbidite, and
456 deepwater lobe systems. The spatial distribution of porosity, permeability and facies, and the spatial
457 location of the injector well(s) are used as the input data. The proxy model is used to predict the
458 dynamic reservoir response over time, namely the video frames, corresponding to the dynamic CO₂
459 pressure and saturation distributions, which are obtained offline for training using HFS. The total
460 training time is 88 minutes on a single NVIDIA Quadro M6000 GPU, and predictions are obtained
461 with 98-99% accuracy within approximately 4.6 milliseconds, compared to the approximate 30

462 seconds required for HFS, a $6,500\times$ speedup.

463 There are several opportunities for future work. First, an extension to 3D geologic models and
464 their corresponding dynamic predictions is key to scaling up this method for real-world applications.
465 Similarly, although the Stochastic pix2vid proxy model is only trained for GCS prediction, it is
466 applicable for a range of processes such as ground-water, compositional, geothermal, or conventional
467 oil and gas systems. Moreover, it is possible to extend the Stochastic pix2vid model from a data-
468 driven mapping to a PINN by including the discretized form of the governing PDE in the loss
469 function and minimizing the residuals. Another future opportunity is to test the performance of
470 the Stochastic pix2vid model on unseen timesteps, either interpolating the training timesteps or
471 extrapolating beyond the training timesteps. Furthermore, the Stochastic pix2vid model can be
472 used as a proxy in workflows for history matching and closed-loop reservoir management.

473 **Reproducibility**

474 The code will be made publicly available on the author's repository (github.com/misaelmmorales).

475 **Funding**

476 This research did not receive any specific grant from funding agencies in the public, or not-for-profit
477 sectors.

478 **Declarations**

479 The authors declare no conflict of interests.

480 **Acknowledgements**

481 The authors thank the Formation Evaluation (FE) and Digital Reservoir Characterization Tech-
482 nology (DIRECT) Industry Affiliate Programs at the University of Texas at Austin for supporting
483 this work.

484 **References**

- 485 [1] K. Michael, A. Golab, V. Shulakova, J. Ennis-King, G. Allinson, S. Sharma, and T. Aiken.
486 Geological storage of co₂ in saline aquifers—a review of the experience from existing storage
487 operations. *International Journal of Greenhouse Gas Control*, 4(4):659–667, 2010. ISSN 1750-
488 5836. doi: <https://doi.org/10.1016/j.ijggc.2009.12.011>.
- 489 [2] A. Goodman, G. Bromhal, B. Strazisar, T. Rodosta, W.F. Guthrie, D. Allen, and G. Guthrie.
490 Comparison of methods for geologic storage of carbon dioxide in saline formations. *Interna-*
491 *tional Journal of Greenhouse Gas Control*, 18:329–342, 2013. doi: 10.1016/j.ijggc.2013.07.016.
492 cited By 48.
- 493 [3] J.S. Levine, I. Fukai, D.J. Soeder, G. Bromhal, R.M. Dilmore, G.D. Guthrie, T. Rodosta,
494 S. Sanguinito, S. Frailey, C. Gorecki, W. Peck, and A.L. Goodman. U.s. doe netl methodology
495 for estimating the prospective co₂ storage resource of shales at the national and regional scale.
496 *International Journal of Greenhouse Gas Control*, 51:81–94, 2016. doi: 10.1016/j.ijggc.2016.
497 04.028. cited By 81.
- 498 [4] Bert Metz, Ogunlade Davidson, HC De Coninck, Manuela Loos, and Leo Meyer. *IPCC special*
499 *report on carbon dioxide capture and storage*. Cambridge: Cambridge University Press, 2005.
- 500 [5] Energy 2020. European commission. In *A strategy for competitive, sustainable and secure*
501 *energy*, 2010.
- 502 [6] United nations. Agreement, p. *United Nations Treaty Collect*, pages 1–27, 2015.
- 503 [7] S. Bachu. Review of co₂ storage efficiency in deep saline aquifers. *International Journal of*
504 *Greenhouse Gas Control*, 40:188–202, 2015. doi: 10.1016/j.ijggc.2015.01.007. cited By 277.
- 505 [8] J.F.D. Tapia, J.-Y. Lee, R.E.H. Ooi, D.C.Y. Foo, and R.R. Tan. Optimal co₂ allocation and
506 scheduling in enhanced oil recovery (eor) operations. *Applied Energy*, 184:337–345, 2016. doi:
507 10.1016/j.apenergy.2016.09.093.
- 508 [9] N. Castelletto, P. Teatini, G. Gambolati, D. Bossie-Codreanu, O. Vincké, J.-M. Daniel, A. Bat-
509 tistelli, M. Marcolini, F. Donda, and V. Volpi. Multiphysics modeling of co₂ sequestration in

- 510 a faulted saline formation in italy. *Advances in Water Resources*, 62:570–587, 2013. doi:
511 10.1016/j.advwatres.2013.04.006. cited By 25.
- 512 [10] Elnara Rustamzade, Wen Pan, John T. Foster, and Michael Pyrcz. Comparison of commin-
513 gled and sequential production schemes by sensitivity analysis for gulf of mexico paleogene
514 deepwater turbidite oil fields: A simulation study. *Energy Exploration & Exploitation*, 0(0):
515 01445987231195679, 2023. doi: 10.1177/01445987231195679. URL <https://doi.org/10.1177/01445987231195679>.
- 516
- 517 [11] K. Rashid, W. Bailey, B. Couët, and D. Wilkinson. An efficient procedure for expensive
518 reservoir-simulation optimization under uncertainty. *SPE Economics and Management*, 5(4):
519 21–33, 2013. doi: 10.2118/167261-PA. cited By 16.
- 520 [12] C. Luo, S.-L. Zhang, C. Wang, and Z. Jiang. A metamodel-assisted evolutionary algorithm for
521 expensive optimization. *Journal of Computational and Applied Mathematics*, 236(5):759–764,
522 2011. doi: 10.1016/j.cam.2011.05.047. cited By 29.
- 523 [13] Javier E. Santos, Bernard Chang, Alex Gigliotti, Eric Guiltinan, Mohamed Mehana, Arvind
524 Mohan, James McClure, Qinjun Kang, Hari Viswanathan, Nicholas Lubbers, Masa Pro-
525 danovic, and Michael Pyrcz. Learning from a big dataset of digital rock simulations. In
526 *AGU Fall Meeting Abstracts*, volume 2021, pages H25O–1207, December 2021.
- 527 [14] Bailian Chen, Dylan R. Harp, Youzuo Lin, Elizabeth H. Keating, and Rajesh J. Pawar.
528 Geologic co2 sequestration monitoring design: A machine learning and uncertainty quan-
529 tification based approach. *Applied Energy*, 225:332–345, 9 2018. ISSN 03062619. doi:
530 10.1016/j.apenergy.2018.05.044.
- 531 [15] Wenyue Sun and Louis J. Durlofsky. Data-space approaches for uncertainty quantification of
532 co2 plume location in geological carbon storage. *Advances in Water Resources*, 123:234–255,
533 1 2019. ISSN 03091708. doi: 10.1016/j.advwatres.2018.10.028. cited By 23.
- 534 [16] Bailian Chen, Dylan R. Harp, Zhiming Lu, and Rajesh J. Pawar. Reducing uncertainty in
535 geologic co2 sequestration risk assessment by assimilating monitoring data. *International*

- 536 *Journal of Greenhouse Gas Control*, 94, 3 2020. ISSN 17505836. doi: 10.1016/j.ijggc.2019.
537 102926.
- 538 [17] B. Li and S.M. Benson. Influence of small-scale heterogeneity on upward co₂ plume migration
539 in storage aquifers. *Advances in Water Resources*, 83:389–404, 2015. doi: 10.1016/j.advwatres.
540 2015.07.010. cited By 84.
- 541 [18] Su Jiang and Louis J. Durlofsky. Use of multifidelity training data and transfer learning for
542 efficient construction of subsurface flow surrogate models. *Journal of Computational Physics*,
543 474, 2 2023. ISSN 10902716. doi: 10.1016/J.JCP.2022.111800.
- 544 [19] *Best Practices in Automatic Permeability Estimation: Machine-Learning Methods vs. Con-*
545 *ventional Petrophysical Models*, volume Day 4 Tue, June 13, 2023 of *SPWLA Annual Logging*
546 *Symposium*, 06 2023. doi: 10.30632/SPWLA-2023-0084.
- 547 [20] H. Wu, N. Lubbers, H.S. Viswanathan, and R.M. Polleyea. A multi-dimensional parametric
548 study of variability in multi-phase flow dynamics during geologic co₂ sequestration accelerated
549 with machine learning. *Applied Energy*, 287, 2021. doi: 10.1016/j.apenergy.2021.116580. cited
550 By 14.
- 551 [21] Siddharth Misra, Yusuf Falola, Polina Churilova, Rui Liu, Chung-Kan Huang, and Jose F.
552 Delgado. Deep learning assisted extremely low-dimensional representation of subsurface earth.
553 *SSRN Electronic Journal*, 8 2022. doi: 10.2139/SSRN.4196705.
- 554 [22] Ademide O. Mabadeje and Michael J. Pyrcz. Rigid transformations for stabilized lower di-
555 mensional space to support subsurface uncertainty quantification and interpretation, 2023.
- 556 [23] Mingliang Liu, Dario Grana, and Tapan Mukerji. Randomized tensor decomposition for large-
557 scale data assimilation problems for carbon dioxide sequestration. *Mathematical Geosciences*,
558 54:1139–1163, 5 2022. ISSN 18748953. doi: 10.1007/S11004-022-10005-1/FIGURES/17.
- 559 [24] S.W.A. Canchumuni, A.A. Emerick, and M.A.C. Pacheco. Towards a robust parameterization
560 for conditioning facies models using deep variational autoencoders and ensemble smoother.
561 *Computers and Geosciences*, 128:87–102, 2019. doi: 10.1016/j.cageo.2019.04.006. cited By 80.

- 562 [25] Y. Zhang, P. Vouzis, and N.V. Sahinidis. Gpu simulations for risk assessment in co2 geologic
563 sequestration. *Computers and Chemical Engineering*, 35(8):1631–1644, 2011. doi: 10.1016/j.
564 compchemeng.2011.03.023. cited By 20.
- 565 [26] Bicheng Yan, Dylan Robert Harp, Bailian Chen, and Rajesh J. Pawar. Improving deep
566 learning performance for predicting large-scale geological co2 sequestration modeling through
567 feature coarsening. *Scientific Reports*, 12:1–12, 11 2022. ISSN 2045-2322. doi: 10.1038/
568 s41598-022-24774-6.
- 569 [27] Zeeshan Tariq, Murtada Saleh Aljawad, Amjad Hasan, Mobeen Murtaza, Emad Mohammed,
570 Ammar El-Husseiny, Sulaiman A Alarifi, Mohamed Mahmoud, and Abdulazeez Abdulraheem.
571 A systematic review of data science and machine learning applications to the oil and gas
572 industry. *Journal of Petroleum Exploration and Production Technology*, pages 1–36, 2021.
- 573 [28] Mohammad Ali Mirza, Mahtab Ghoroori, and Zhangxin Chen. Intelligent petroleum engineer-
574 ing. *Engineering*, 18:27–32, 2022. ISSN 2095-8099. doi: <https://doi.org/10.1016/j.eng.2022.06.009>.
- 576 [29] Jean-Paul Chiles and Pierre Delfiner. *Geostatistics: modeling spatial uncertainty*, volume 713.
577 John Wiley & Sons, 2012.
- 578 [30] Michael J Pyrcz and Clayton V Deutsch. *Geostatistical reservoir modeling*. Oxford University
579 Press, USA, 2014.
- 580 [31] Proctor Joshua Brunton, Steve and Nathan Kutz. Discovering governing equations from data
581 by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy
582 of Sciences of the United States of America*, 2016. doi: 10.1073/pnas.1517384113.
- 583 [32] He Xiaolong Fries, William and Youngsoo Choi. Lasdi: Parametric latent space dynamics
584 identification. *Computer Methods in Applied Mechanics and Engineering*, 2022. doi: 10.1016/
585 j.cma.2022.115436.
- 586 [33] Choi Youngsoo Fries William Belof Jonathan He, Xiaolong and Jiun-Shyan Chen. glasdi: Para-
587 metric physics-informed greedy latent space dynamics identification. *Journal of Computational
588 Physics*, 2023.

- 589 [34] M. Liu and D. Grana. Time-lapse seismic history matching with an iterative ensemble smoother
590 and deep convolutional autoencoder. *Geophysics*, 85(1):M15–M31, 2020. cited By 2.
- 591 [35] Syamil Mohd Razak, Anyue Jiang, and Behnam Jafarpour. Latent-space inversion (lsi): a
592 deep learning framework for inverse mapping of subsurface flow data. *Computational Geo-*
593 *science*, 26:71–99, 11 2022. doi: 10.1007/s10596-021-10104-8.
- 594 [36] S. Oladyshkin, H. Class, and W. Nowak. Bayesian updating via bootstrap filtering com-
595 bined with data-driven polynomial chaos expansions: Methodology and application to history
596 matching for carbon dioxide storage in geological formations. *Computational Geosciences*, 17
597 (4):671–687, 2013. doi: 10.1007/s10596-013-9350-6. cited By 36.
- 598 [37] Anqi Bao, Eduardo Gildin, Abhinav Narasingam, and Joseph S. Kwon. Data-driven model
599 reduction for coupled flow and geomechanics based on dmd methods. *Fluids*, 4:138, 7 2019.
600 ISSN 2311-5521. doi: 10.3390/FLUIDS4030138.
- 601 [38] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu
602 Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.
- 603 [39] Liu Yang, Dongkun Zhang, and George Em Karniadakis. Physics-informed generative adver-
604 sarial networks for stochastic differential equations, 2018.
- 605 [40] N. Wang, H. Chang, and D. Zhang. Efficient uncertainty quantification for dynamic subsurface
606 flow with surrogate by theory-guided neural network. *Computer Methods in Applied Mechanics
607 and Engineering*, 373, 2021. doi: 10.1016/j.cma.2020.113492. cited By 33.
- 608 [41] Emilio Jose Rocha Coutinho, Marcelo Dall’Aqua, and Eduardo Gildin. Physics-aware deep-
609 learning-based proxy reservoir simulation model equipped with state and well output predic-
610 tion. *Frontiers in Applied Mathematics and Statistics*, 7:49, 9 2021. ISSN 22974687. doi:
611 10.3389/FAMS.2021.651178/BIBTEX.
- 612 [42] Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. Physics-
613 constrained deep learning for high-dimensional surrogate modeling and uncertainty quantifi-
614 cation without labeled data. *Journal of Computational Physics*, 394:56–81, oct 2019. doi:
615 10.1016/j.jcp.2019.05.024. URL <https://doi.org/10.1016%2Fj.jcp.2019.05.024>.

- 616 [43] B Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.
- 617 [44] Jeff Heaton. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit
618 press, 2016, 800 pp, isbn: 0262035618. *Genetic programming and evolvable machines*, 19(1-2):
619 305–307, 2018.
- 620 [45] Yimin Liu and Louis J Durlofsky. 3d cnn-pca: A deep-learning-based parameterization for
621 complex geomodels. *Computers & Geosciences*, 148:104676, 2021.
- 622 [46] Zixiao Yang, Qiyu Chen, Zhesi Cui, Gang Liu, Shaoqun Dong, and Yiping Tian. Automatic re-
623 construction method of 3d geological models based on deep convolutional generative adversarial
624 networks. *Computational Geosciences*, 26:1135–1150, 2022. doi: 10.1007/s10596-022-10152-8.
- 625 [47] Su Jiang and Louis J Durlofsky. Data-space inversion using a recurrent autoencoder for time-
626 series parameterization. *Computational Geosciences*, 25:411–432, 2021.
- 627 [48] Yanrui Ning, Hossein Kazemi, and Pejman Tahmasebi. A comparative machine learning study
628 for time series oil production forecasting: Arima, lstm, and prophet. *Computers and Geo-
629 sciences*, 164:105126, 7 2022. ISSN 00983004. doi: 10.1016/j.cageo.2022.105126.
- 630 [49] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining
631 Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings
632 of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- 633 [50] Liuqing Yang, Sergey Fomel, Shoudong Wang, Xiaohong Chen, Wei Chen, Omar M. Saad,
634 and Yangkang Chen. Porosity and permeability prediction using a transformer and periodic
635 long short-term network. *Geophysics*, 88(1):WA293–WA308, 01 2023. ISSN 0016-8033. doi:
636 10.1190/geo2022-0150.1.
- 637 [51] Eduardo Maldonado Cruz and Michael J Pyrcz. Multi-horizon well performance forecasting
638 with temporal fusion transformers. *Available at SSRN 4403939*.
- 639 [52] Wen Pan, Carlos Torres-Verdín, and Michael J. Pyrcz. Stochastic pix2pix: A new ma-
640 chine learning method for geophysical and well conditioning of rule-based channel reser-
641 voir models. *Natural Resources Research*, 30:1319–1345, 4 2021. ISSN 15738981. doi:
642 10.1007/S11053-020-09778-1/FIGURES/24.

- 643 [53] Bogdan Sebacher and Stefan Adrian Toma. Bridging deep convolutional autoencoders and en-
644 semble smoothers for improved estimation of channelized reservoirs. *Mathematical Geosciences*,
645 54:903–939, 7 2022. ISSN 18748953. doi: 10.1007/S11004-022-09997-7/TABLES/3.
- 646 [54] Jichao Bao, Liangping Li, and Arden Davis. Variational autoencoder or generative ad-
647 versarial networks? a comparison of two deep learning methods for flow and transport
648 data assimilation. *Mathematical Geosciences*, 54:1017–1042, 8 2022. ISSN 18748953. doi:
649 10.1007/S11004-022-10003-3/FIGURES/17.
- 650 [55] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image
651 segmentation. *CoRR*, 2015. cited By 358.
- 652 [56] Eduardo Maldonado-Cruz and Michael J. Pyrcz. Fast evaluation of pressure and saturation
653 predictions with a deep learning surrogate flow model. *Journal of Petroleum Science and*
654 *Engineering*, 212:110244, 5 2022. ISSN 0920-4105. doi: 10.1016/J.PETROL.2022.110244.
- 655 [57] Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M. Benson.
656 U-fno—an enhanced fourier neural operator-based deep-learning model for multiphase flow.
657 *Advances in Water Resources*, 163:104180, 2022. ISSN 0309-1708. doi: <https://doi.org/10.1016/j.advwatres.2022.104180>.
- 659 [58] Gege Wen, Zongyi Li, Qirui Long, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M.
660 Benson. Real-time high-resolution co 2 geological storage prediction using nested fourier neural
661 operators. *Energy & Environmental Science*, 2023. ISSN 1754-5692. doi: 10.1039/d2ee04204e.
- 662 [59] Honggeun Jo, Wen Pan, Javier E Santos, Hyungsik Jung, and Michael J Pyrcz. Machine
663 learning assisted history matching for a deepwater lobe system. *Journal of Petroleum Science*
664 *and Engineering*, 207:109086, 2021.
- 665 [60] Feng Zhang, Long Nghiem, and Zhangxin Chen. Evaluating reservoir performance using a
666 transformer based proxy model. *Geoenergy Science and Engineering*, 226:211644, 2023.
- 667 [61] Daowei Zhang and Heng Li. Efficient surrogate modeling based on improved vision transformer
668 neural network for history matching. *SPE Journal*, pages 1–17, 2023.

- 669 [62] Yong Do Kim and Louis J. Durlofsky. Convolutional – recurrent neural network proxy for
670 robust optimization and closed-loop reservoir management. *Computational Geosciences*, pages
671 1–24, 1 2023. ISSN 1420-0597. doi: 10.1007/S10596-022-10189-9/TABLES/1.
- 672 [63] Meng Tang, Yimin Liu, and Louis J. Durlofsky. A deep-learning-based surrogate model for
673 data assimilation in dynamic subsurface flow problems. *Journal of Computational Physics*,
674 413, 7 2020. ISSN 10902716. doi: 10.1016/J.JCP.2020.109456.
- 675 [64] M. Tang, Y. Liu, and L.J. Durlofsky. Deep-learning-based surrogate flow modeling and ge-
676 ological parameterization for data assimilation in 3d subsurface flow. *Computer Methods in*
677 *Applied Mechanics and Engineering*, 376, 2021. doi: 10.1016/j.cma.2020.113636. cited By 39.
- 678 [65] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dy-
679 namics, 2016.
- 680 [66] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond
681 mean square error, 2016.
- 682 [67] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing
683 motion and content for natural video sequence prediction, 2018.
- 684 [68] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing
685 motion and content for video generation, 2017.
- 686 [69] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-
687 chun WOO. Convolutional lstm network: A machine learning approach for precipita-
688 tion nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett,
689 editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Asso-
690 ciates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.
- 692 [70] Michael Iliadis, Leonidas Spinoulas, and Aggelos K. Katsaggelos. Deep fully-connected net-
693 works for video compressive sensing, 2017.

- 694 [71] Kai Xu and Fengbo Ren. Csvideonet: A real-time end-to-end learning framework for high-
695 frame-rate video compressive sensing. In *2018 IEEE Winter Conference on Applications of*
696 *Computer Vision (WACV)*, pages 1680–1688. IEEE, 2018.
- 697 [72] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G.
698 Derpanis, and Björn Ommer. Stochastic image-to-video synthesis using cinns, 2021.
- 699 [73] Aleksander Holynski, Brian Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures
700 with eulerian motion fields, 2020.
- 701 [74] Karsten Pruess, Curtis M Oldenburg, and GJ Moridis. Tough2 user’s guide version 2. Technical
702 report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 1999.
- 703 [75] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Pro-*
704 , pages 1251–1258,
705 2017.
- 706 [76] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE*
707 *conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- 708 [77] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing
709 ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- 710 [78] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM,
711 2005.
- 712 [79] D.S. Oliver, A.C. Reynolds, and N. Liu. *Inverse theory for petroleum reservoir characterization*
713 *and history matching*, volume 9780521881517. 2008. doi: 10.1017/CBO9780511535642. cited
714 By 766.
- 715 [80] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assess-
716 ment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*,
717 13:600–612, 4 2004. ISSN 1941-0042. doi: doi.org/10.1109/TIP.2003.819861.
- 718 [81] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*
719 *arXiv:1711.05101*, 2017.

- [82] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [83] Nicolas Remy, Alexandre Boucher, and Jianbing Wu. *Applied Geostatistics with SGeMS: A User's Guide*. Cambridge University Press, 2009.
- [84] G. W. Verly. *Sequential Gaussian Cosimulation: A Simulation Method Integrating Several Types of Information*, pages 543–554. Springer Netherlands, Dordrecht, 1993. ISBN 978-94-011-1739-5. doi: 10.1007/978-94-011-1739-5_42.
- [85] M.J. Pyrcz, J.B. Boisvert, and C.V. Deutsch. A library of training images for fluvial and deepwater reservoirs and associated code. *Computers Geosciences*, 34(5):542–560, 2008. ISSN 0098-3004. doi: <https://doi.org/10.1016/j.cageo.2007.05.015>.
- [86] Misael M. Morales and Michael Pyrcz. GeostatsGuy/MLTrainingImages: MachineLearning-TrainingImages_v1.0.0, March 2023. URL <https://doi.org/10.5281/zenodo.7702128>.
- [87] Knut-Andreas Lie. *An introduction to reservoir simulation using MATLAB/GNU Octave: User guide for the MATLAB Reservoir Simulation Toolbox (MRST)*. Cambridge University Press, 2019.
- [88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [89] Q. Li and G. Liu. *Risk assessment of the geological storage of CO₂: A review*. 2016. doi: 10.1007/978-3-319-27019-7_13. cited By 39.
- [90] R.A. Chadwick, R. Arts, and O. Eiken. 4d seismic quantification of a growing co₂ plume at sleipner, north sea. *Petroleum Geology Conference Proceedings*, 6(0):1385–1399, 2005. doi: 10.1144/0061385. cited By 188.
- [91] R.A. Chadwick and D.J. Noy. History-matching flow simulations and timelapse seismic data from the sleipner co₂ plume. *7th Petroleum Geology Conference [FROM MATURE BASINS to NEW FRONTIERS] (London, 3/30/2009-4/2/2009) Proceedings*, 2:1171–1182, 2010. cited By 31.

747 [92] Ismael Dawuda and Sanjay Srinivasan. Geologic modeling and ensemble-based history match-
748 ing for evaluating co2 sequestration potential in point bar reservoirs. *Frontiers in Energy*
749 *Research*, 10:867083, 2022.