Variational Autoencoder Geostatistical Approach (VEGA) with Subsurface Applications

Jonghyun Harry Lee[1] and Hongkyu Yoon[2]

[1]Civil and Environmental Engineering, University of Hawaii at Manoa, Honolulu, HI, USA; [2]Geomechanics Department, Sandia National Laboratories, Albuquerque, NM, USA

Introduction

In this work, we present a fast Machine Learning (ML)-based inverse modeling for real-time subsurface history matching and forecasts. The work aims to support CO_2 related operations and forecasting CO_2 and pressure plume development for DOE Science-informed Machine Learning to Accelerate Real Time (SMART) initiative. The proposed method utilizes one of the current efficient data assimilation methods, i.e., *Hierarchical Bayesian with Gaussian Prior* (through power-transformation of unknowns). In Bayesian framework, the prior uncertainty of subsurface unknown properties, e.g., permeability field, is typically parameterized as log-normal with mean and covariance. However, there are two major challenges for high-dimensional applications:

- the number of (expensive) forward model simulations
- computational burdens for high-dimensional matrix-matrix computations.

These challenges were partly addressed in PCGA [Lee, Yoon et. al., 2016] that utilizes state-of-art linear dimension reduction techniques for the Bayesian solution.
Still, PCGA requires O(100) numerical model runs, which would not be feasible for extremely high dimension problems such as 10^9 unknown permeability characterization.
To address these challenges for next generation inversion, we use ML to construct a reduced order model (ROM) for the forward simulation and perform a nonlinear dimension reduction for inverse modeling. This approach will accelerate both forward modeling and inverse modeling tasks without losing much accuracy.

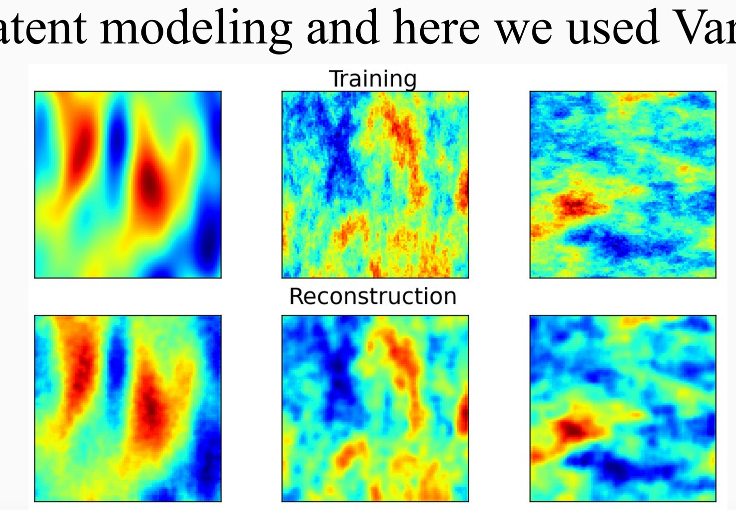
Model Reduction

Forward modeling:

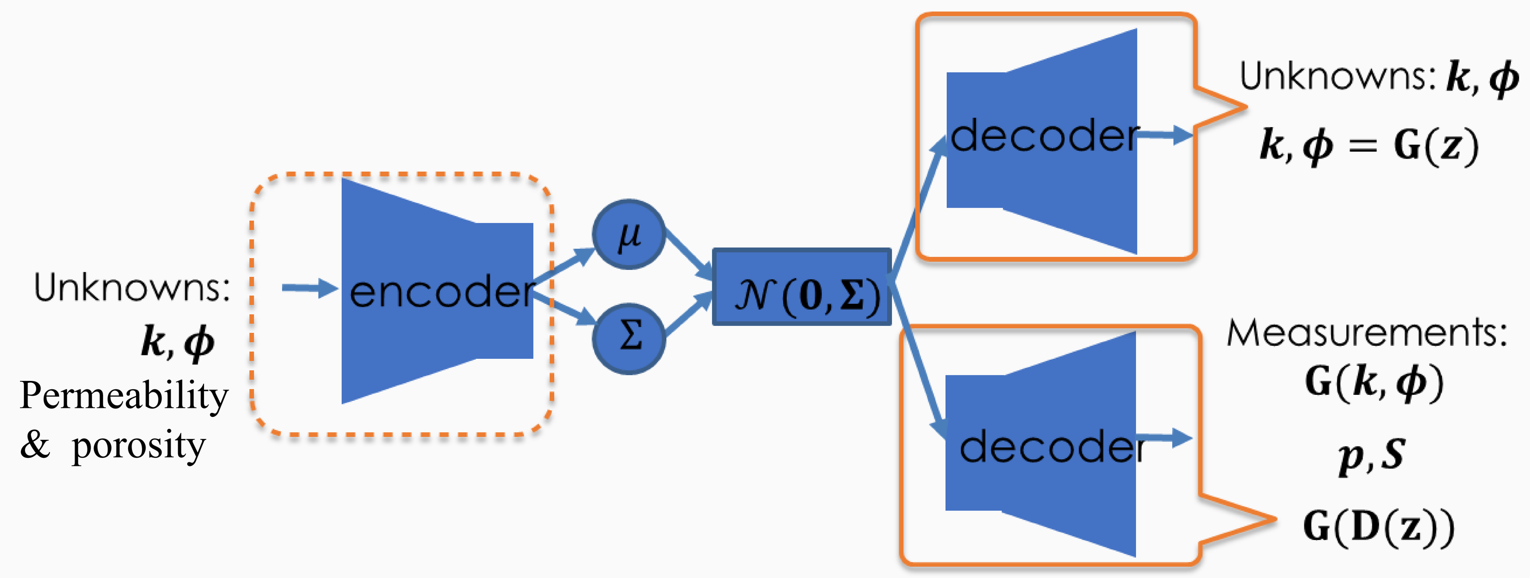
$$\mathbf{y} = \mathbf{G}(\mathbf{m}) \approx \mathbf{G}(\mathbf{D}(\mathbf{z}))$$

- **y** is a (nobs x 1) simulated observation vector,
- **G** is a multiphase flow model that produces outputs at obs. locations
- **m** is a (m x 1) permeability vector,
- **D** is a decoder or deterministic map from z to s
- **z** is a (k x 1) latent space vector.

G can be obtained from any latent modeling and here we used Variational Auto Encoder (VAE).

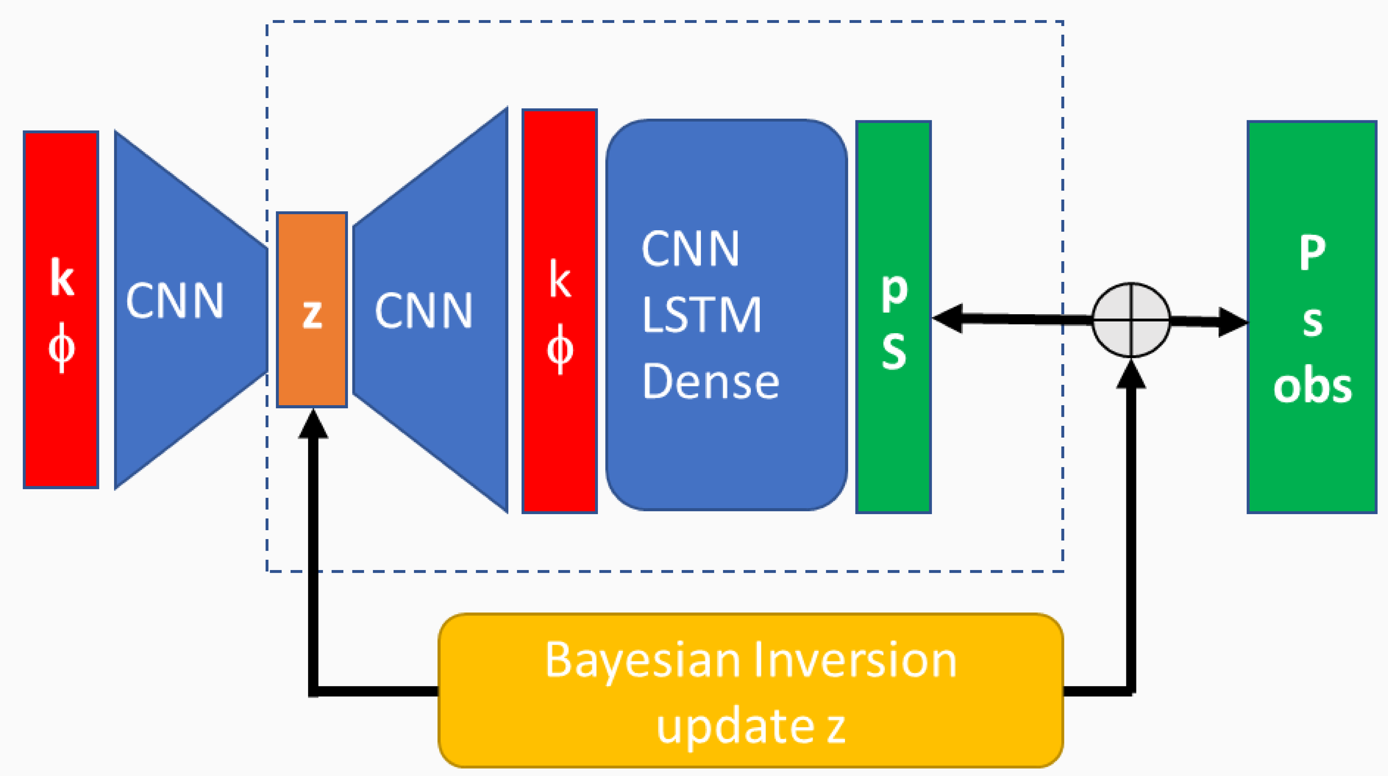


Machine Learning Architecture



Encoder-Decoder Architecture

Variational Autoencoder Geostatistical Approach (VEGA)



Main conceptual idea

Forward problem: $\mathbf{y} = \mathbf{G}(\mathbf{m})$
with l Gauss Newton iterations from $\mathbf{m}^0 = \mathbf{m}_{prior}$

$$\boldsymbol{m}^{l+1} = \boldsymbol{m}^0 + \mathbf{C}_{\text{prior}}\mathbf{J}\left(\mathbf{J}\mathbf{C}_{\text{prior}}\mathbf{J}^{\mathrm{T}} + \mathbf{C}_{\mathbf{obs}}\right)^{-1}\left(\mathbf{y} - \mathbf{G}(\mathbf{m}^l) + \mathbf{J}(\boldsymbol{m}^l - \boldsymbol{m}^0)\right)$$

With any (nonlinear) dimension reduction $\mathbf{G}$ of $\mathbf{m}$

$$\boldsymbol{y} = \mathbf{G}(\mathbf{D}(\mathbf{z})),\ \dim(z) \ll \dim(m)$$

$$\mathbf{z}^{l+1} = \mathbf{z}^l + \alpha\left(\mathbf{J}_{\mathbf{z}}^{\mathbf{T}}\mathbf{C}_{\text{obs}}^{-1}\mathbf{J}_{\mathbf{z}} + \mathbf{C}_{\text{prior}_{\mathrm{z}}}^{-1}\right)^{-1}\left(\mathbf{y} - \mathbf{G}\left(\mathbf{D}(\mathbf{z}^l)\right) - \mathbf{C}_{\text{prior}_{\mathrm{z}}}^{-1}\mathbf{z}^l\right)$$

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

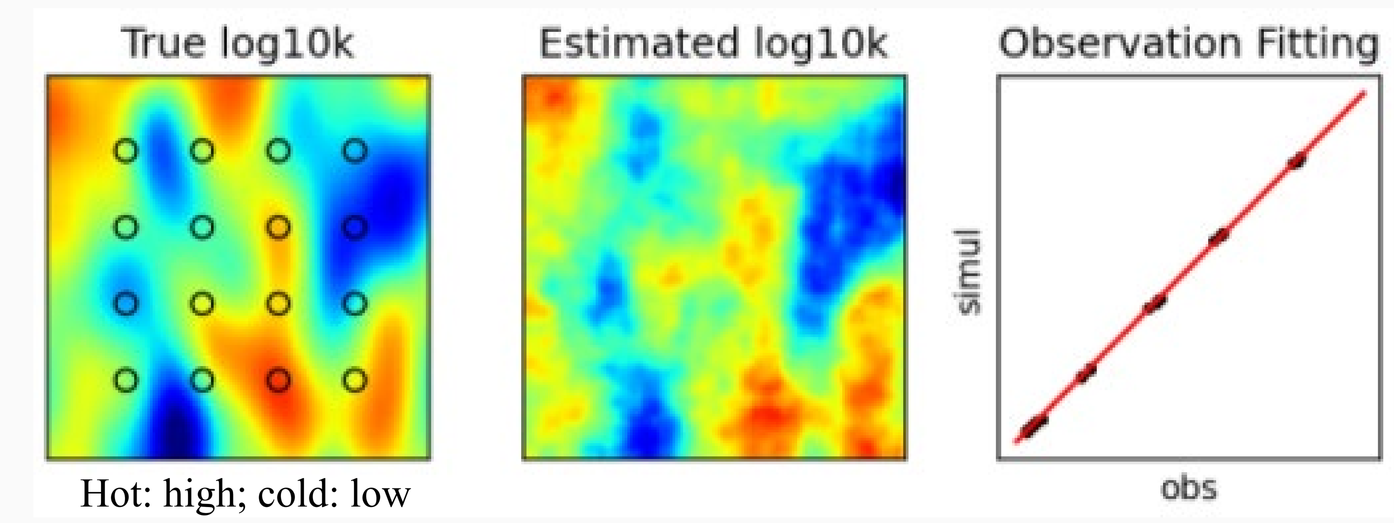
$$\mathbf{C}_{\text{posterior}_{\mathrm{z}}} = \left(\mathbf{J}_{\mathbf{z}}^{\mathbf{T}}\mathbf{C}_{\text{obs}}^{-1}\mathbf{J}_{\mathbf{z}} + \mathbf{C}_{\text{prior}_{\mathrm{z}}}^{-1}\right)^{-1}$$

VAE is our choice since it constructs the prior $\boldsymbol{z} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I})$!

Experimental Results

- Validation with Single Phase Flow Application

Here we used USGS MODFLOW as "full" physics single phase flow model. 10,000 (100x100) unknown k field has been compressed to 32 latent dimension through VAE. With noisy 16 observation wells with head data and 33 forward model runs/iteration, the best estimate is converged in 3 iterations.

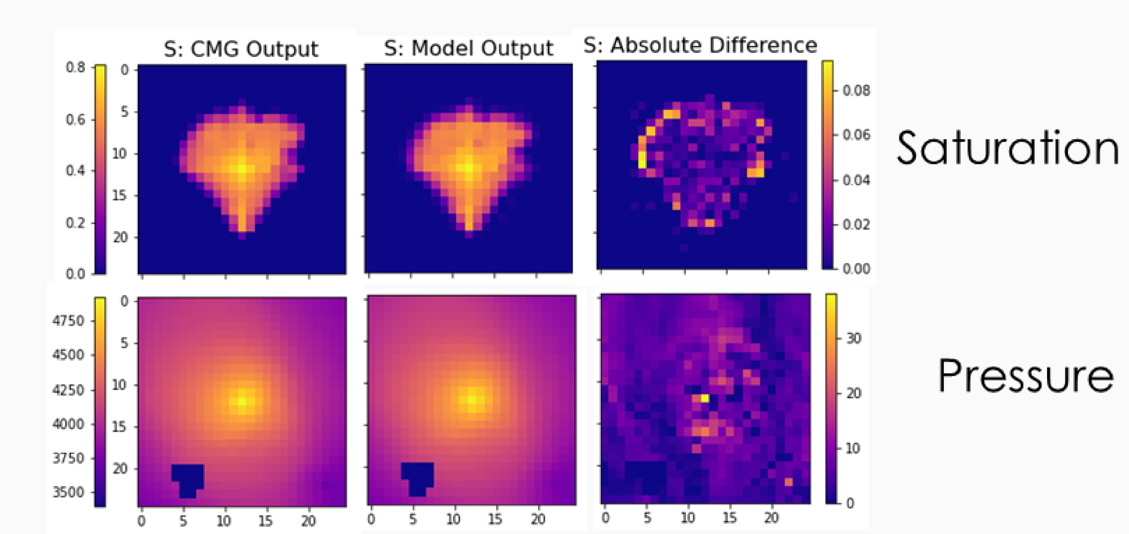


Hot: high; cold: low

- Multiphase Flow Application

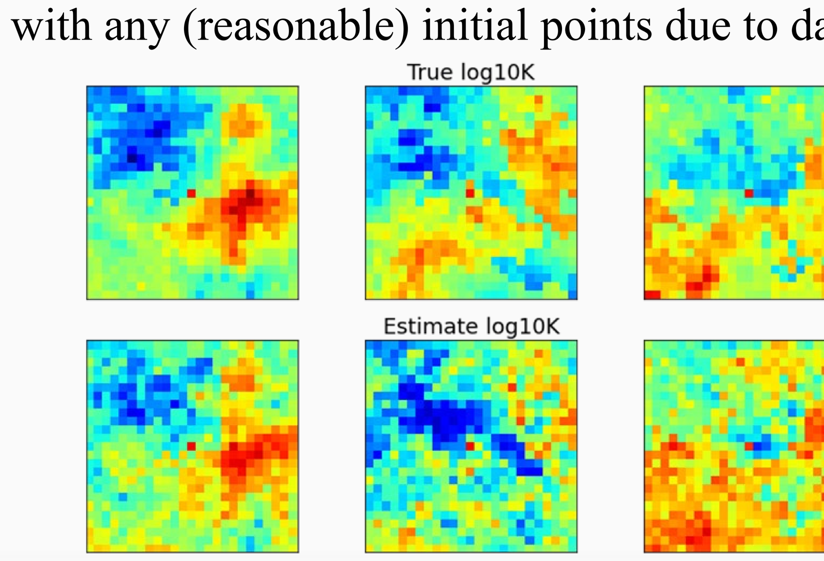
ROM: ML-based model with physics-based loss functions

- Multiphase flow with CNN-LSTM-DenseLayer for in a deep reservoir.
- MSE values of 5E-05 for both pressure and CO_2 saturation prediction
- a forward run time of trained model is less than 1 second



ROM: ML-based model with physics-based loss functions

- 25x25x3 unknown k => z with 32 latent dimension
- 720,000 noisy transient pressure observations
- Only ~3 min inversion time on a single core laptop
- Convergence with any (reasonable) initial points due to data-driven prior



Acknowledgement

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories and also DOE Office of Fossil Energy project -Science-informed Machine Learning to Accelerate Real Time (SMART) Decisions in Subsurface Applications.

SAND2021-6598 C