Machine Learning Based Localization for Iterative Ensemble Smoothers

Vinicius Silva, Gabriel Serrão, **Alexandre Emerick**

20th EnKF Workshop, Norway 2025

INTERNA

Localization

- Distance-based localization is highly effective. However, key reservoir parameters lack a direct spatial relationship with observations.
- ES-MDA with Schur-product localization of the Kalman gain:

$$\mathbf{m}_j^{\ell+1} = \mathbf{m}_j^{\ell} + \mathbf{R} \circ \underbrace{\left[\widetilde{\mathbf{C}}_{\mathbf{md}}^{\ell}\left(\widetilde{\mathbf{C}}_{\mathbf{dd}}^{\ell} + \alpha_\ell \mathbf{C}_{\mathbf{e}}\right)^{-1}\right]}_{\widetilde{\mathbf{K}}^{\ell}} \left(\mathbf{d}_{\mathrm{obs},j}^{\ell} - \mathbf{g}\left(\mathbf{m}_j^{\ell}\right)\right)$$

- Several distance-free methods have been proposed for scalar parameters, although none of them is completely effective.

INTERNA

Pseudo-Optimal Localization*

- Minimization (term-by-term) of the Frobenius norm of the difference between the true covariance and its localized estimate

$$\mathbb{E}\left[\|\mathbf{C}-\mathbf{R}\circ\widetilde{\mathbf{C}}\|_F^2\right]=\mathrm{tr}\left(\mathbf{C}^2\right)-2\mathbb{E}\left[\mathrm{tr}\left(\mathbf{C}\left(\mathbf{R}\circ\widetilde{\mathbf{C}}\right)\right)\right]+\mathbb{E}\left[\left(\mathbf{R}\circ\widetilde{\mathbf{C}}\right)^2\right]$$

$$r_{ik}=\frac{c_{ik}^2}{c_{ik}^2+\frac{c_{ik}^2+c_{ii}c_{kk}}{N_e}} \quad\Rightarrow\quad r_{ik}=\frac{\rho_{ik}^2}{\rho_{ik}^2+\frac{1+\rho_{ik}^2}{N_e}}$$

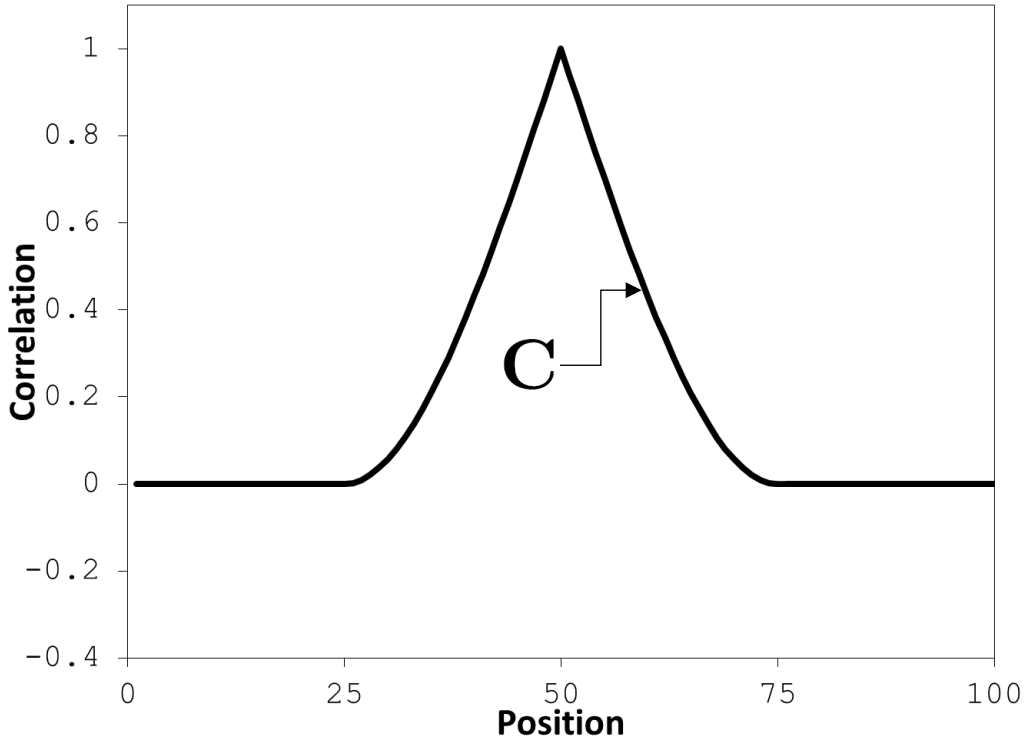
ρ_{ik} correlation coefficient
$i\mathrm{th}$ model parameter
$k\mathrm{th}$ datum

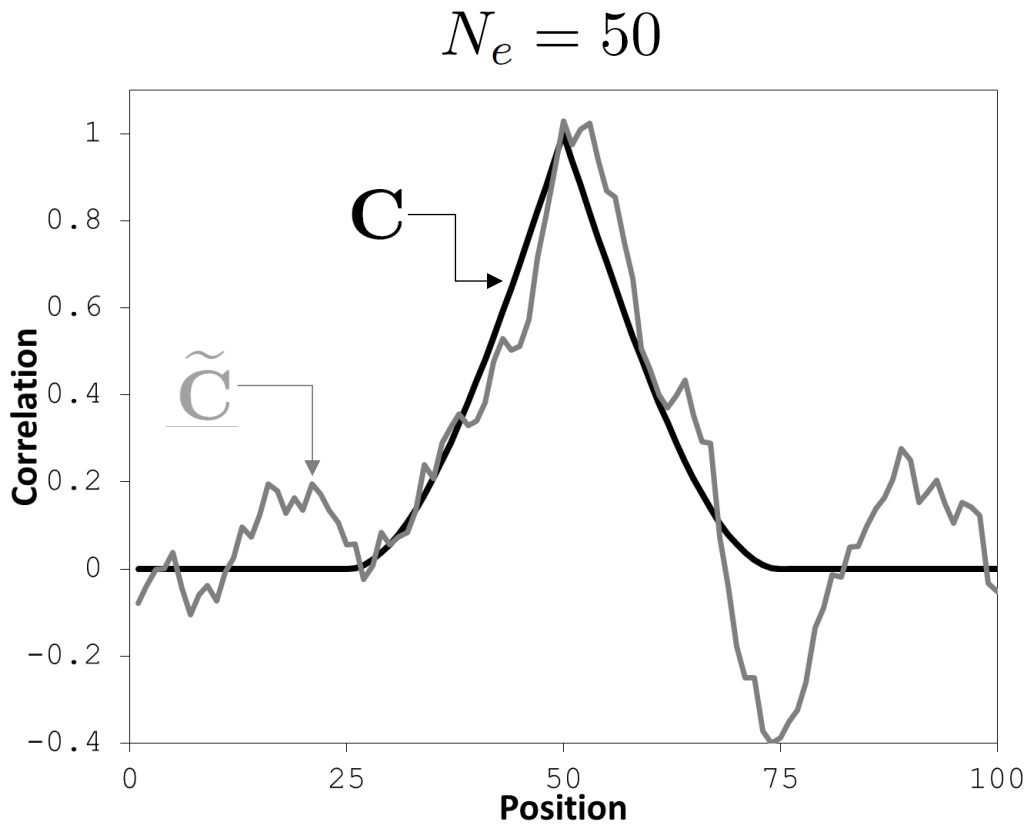
$$\widetilde{r}_{ik}=\frac{\widetilde{\rho}_{ik}^2}{\widetilde{\rho}_{ik}^2+\frac{1+\widetilde{\rho}_{ik}^2}{N_e}} \qquad \widetilde{r}_{ik}=0, \quad \text{if} \quad |\widetilde{\rho}_{ik}|<\tau$$

*Furrer, R., & Bengtsson, T.: **Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants**, *Journal of Multivariate Analysis*, *98*(2), 2007

INTERNA

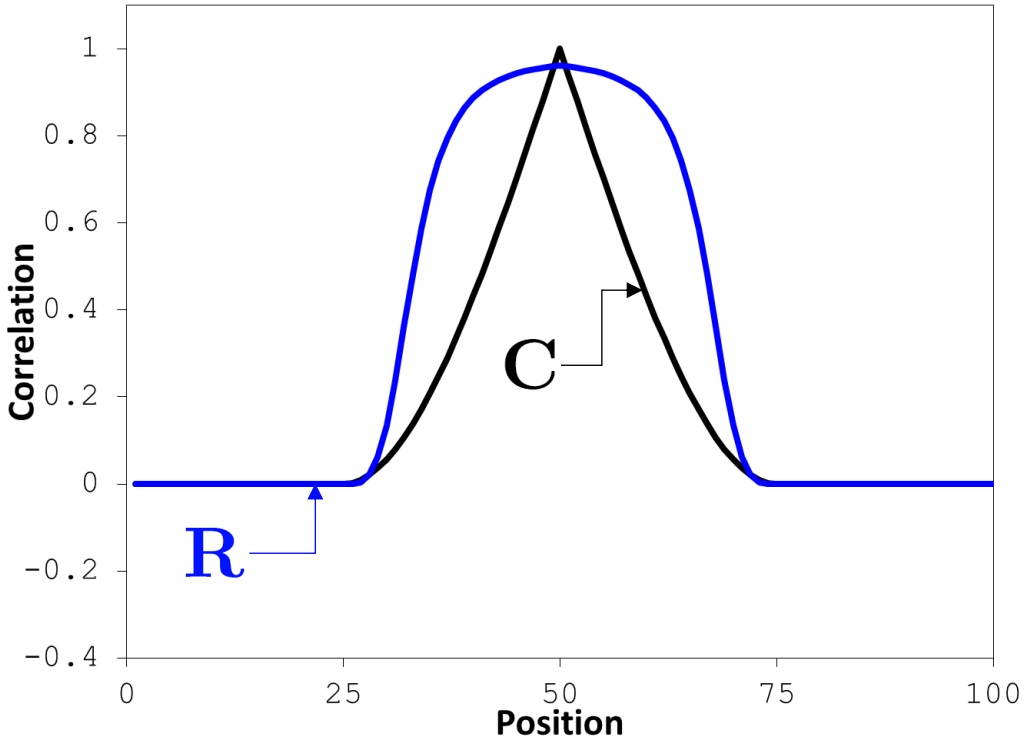
Pseudo-Optimal Localization



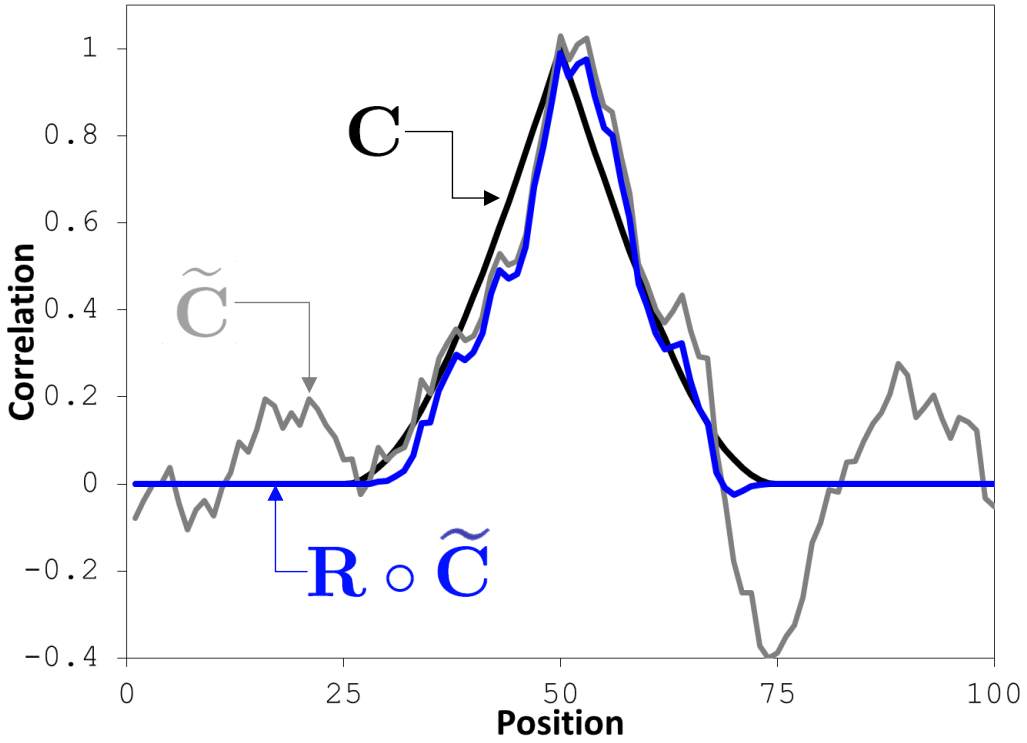


INTERNA

Pseudo-Optimal Localization

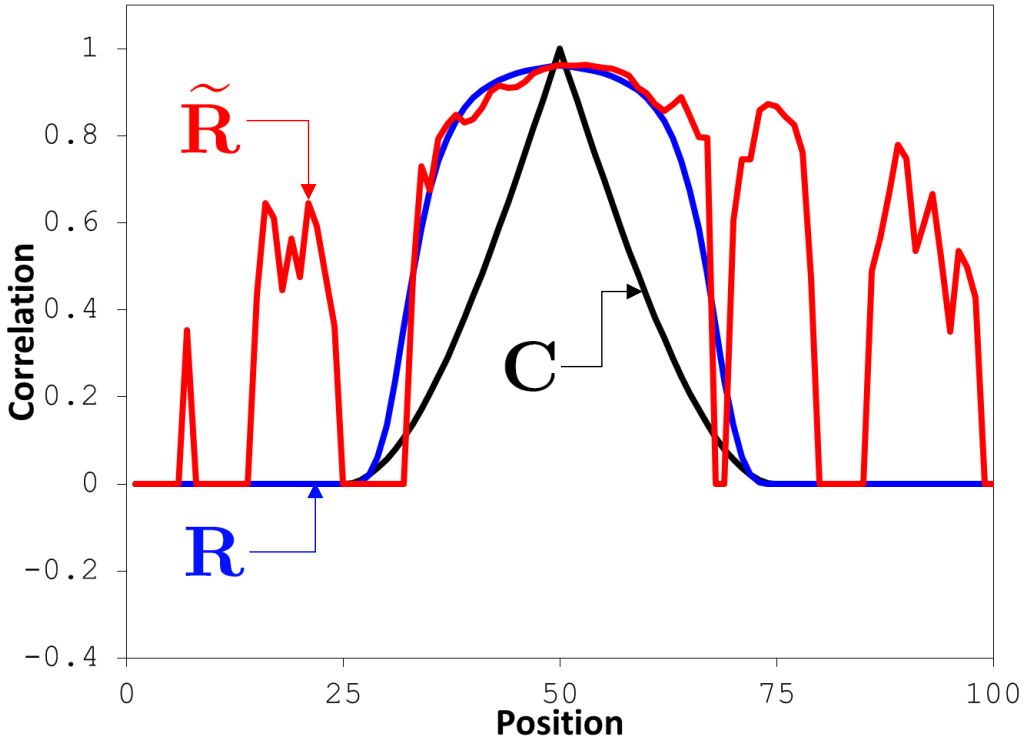


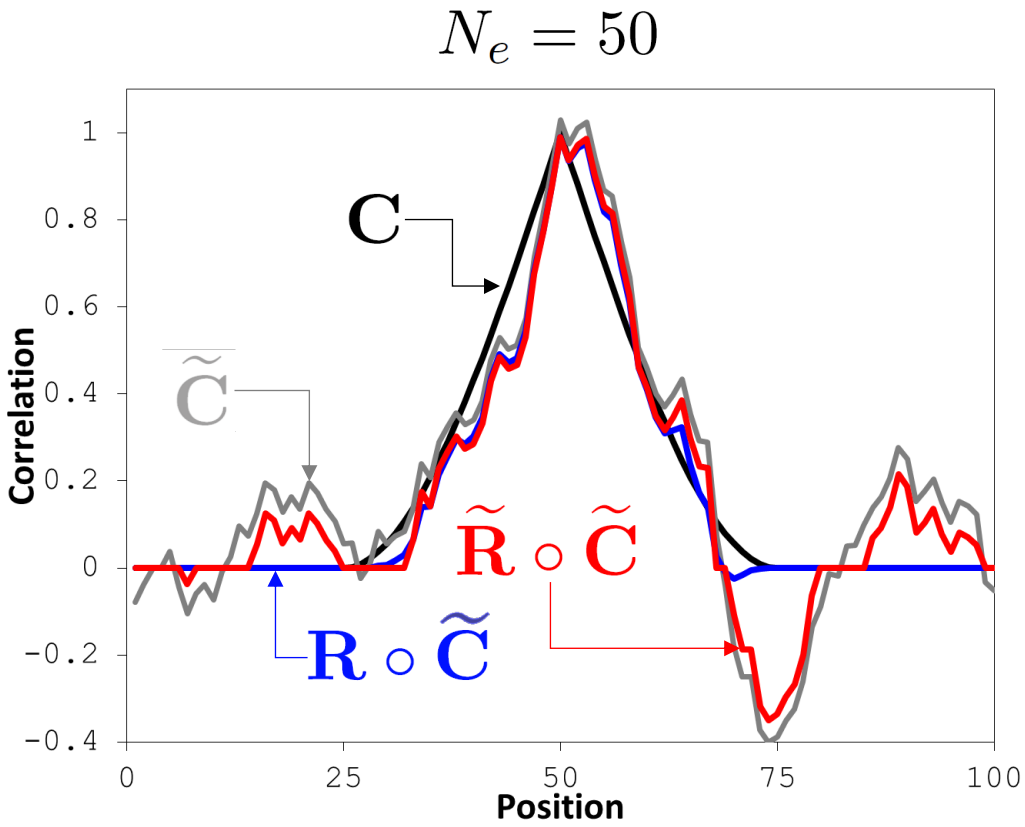
$N_e = 50$



INTERNA

Pseudo-Optimal Localization





INTERNA

Machine Learning Localization

1. Simulate the prior ensemble: $\left\{\mathbf{m}_j\right\}_{j=1}^{N_e} \xrightarrow{\text{Forward simulator}} \left\{\mathbf{d}_j = \mathbf{g}(\mathbf{m}_j)\right\}_{j=1}^{N_e}$

2. Train the machine learning proxy: $\widehat{\mathbf{d}} = \mathcal{G}(\mathbf{m})$

3. Use the ML proxy in a large ensemble: $\left\{\mathbf{m}_j\right\}_{j=1}^{N_E} \xrightarrow{\text{ML proxy}} \left\{\widehat{\mathbf{d}}_j = \mathcal{G}(\mathbf{m}_j)\right\}_{j=1}^{N_E} \quad N_E \gg N_e$

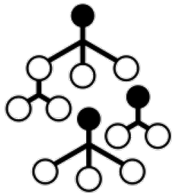
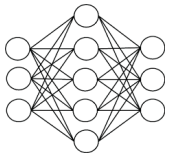
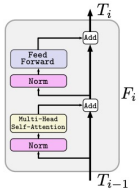
4. Compute the cross-covariance: $\widetilde{\mathbf{C}}_{\mathbf{md}} = \frac{1}{N_E - 1}\sum_{j=1}^{N_E}\left(\mathbf{m}_j - \overline{\mathbf{m}}\right)\left(\widehat{\mathbf{d}}_j - \overline{\widehat{\mathbf{d}}}\right)^{\top}$

5. Compute the localization coefficients: $\widetilde{r}_{md} = \dfrac{\widetilde{\rho}_{md}^2}{\widetilde{\rho}_{md}^2 + \frac{1+\widetilde{\rho}_{md}^2}{N_e}} \qquad \widetilde{\rho}_{md} = \dfrac{\widetilde{c}_{md}}{\sqrt{\widetilde{c}_{mm}\widetilde{c}_{dd}}}$

$$\widetilde{r}_{md} = 0, \quad \text{if} \quad |\widetilde{\rho}_{md}| < \tau = 10^{-3}$$

INTERNA

Machine Learning Methods

- Support Vector Regression 
- Extra Tree
- Decision Tree 
- Random Forest
- XGBoost
- LightGBM
- Neural Network (Multilayer Perceptron) 
- Tabnet
- FT-Transformer 

Goals:

- No additional reservoir simulations (beyond those already used in the data assimilation)
- No additional tuning parameter (no hyperparameter tuning)

INTERNA

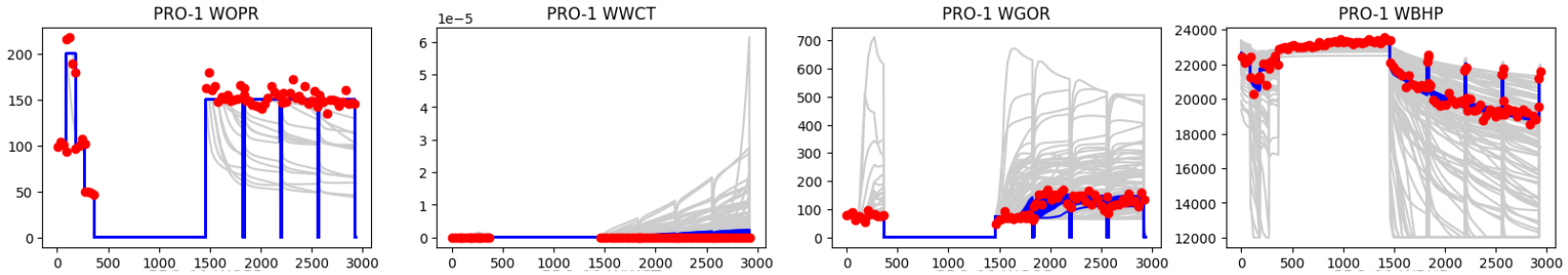
Part 1 – Localization for Scalar Parameters

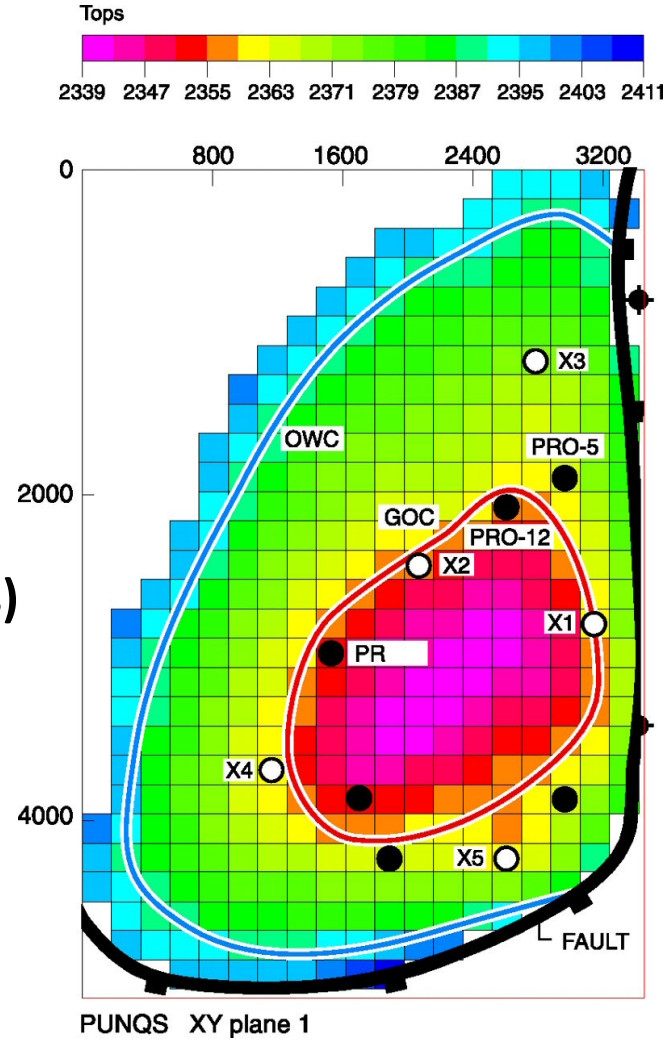
INTERNA

Test Case (PUNQ-S3* modified model)

- **Model parameters (20 scalar variables)**
  - 5 Porosity multipliers
  - 5 Permeability multipliers
  - Compressibility
  - Water-oil contact depth
  - Gas-oil contact depth
  - 2 Aquifers radius
  - **5 Dummy variables**

- **Observed data from 6 producers -- 8 years history (1530 observed points)**





*Floris, F.J., Bush, M.D., Cuypers, M., Roggero, F., & Syversveen, A.R.: **Methods for Quantifying the Uncertainty of Production Forecasts: a Comparative Study**, *Petroleum Geoscience*, 7(S), 2001

INTERNA

Results

- Ensemble size (training set): $N_e = 100$
- Reference: $N_E = 5000$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_m N_d} \sum_{i=1}^{N_m} \sum_{k=1}^{N_d} \left(\widetilde{\rho}_{ik}^{5000} - \widetilde{r}_{ik} \cdot \widetilde{\rho}_{ik}^{100} \right)^2}$$

	Training time (s)	RMSE (Correlation)
No localization	-	0.053
Pseudo-optimal localization	-	0.047
Decision Tree	0.2	0.052
Support Vector Regression	2.7	0.041
Random Forest	10.3	0.046
Extra Tree	6.3	0.044
XGBoost	58.1	0.043
LightGBM	31.7	0.040
Multilayer Perceptron	111	0.039
FT-Transform	269.8	0.039
Tabnet	50.9	0.051

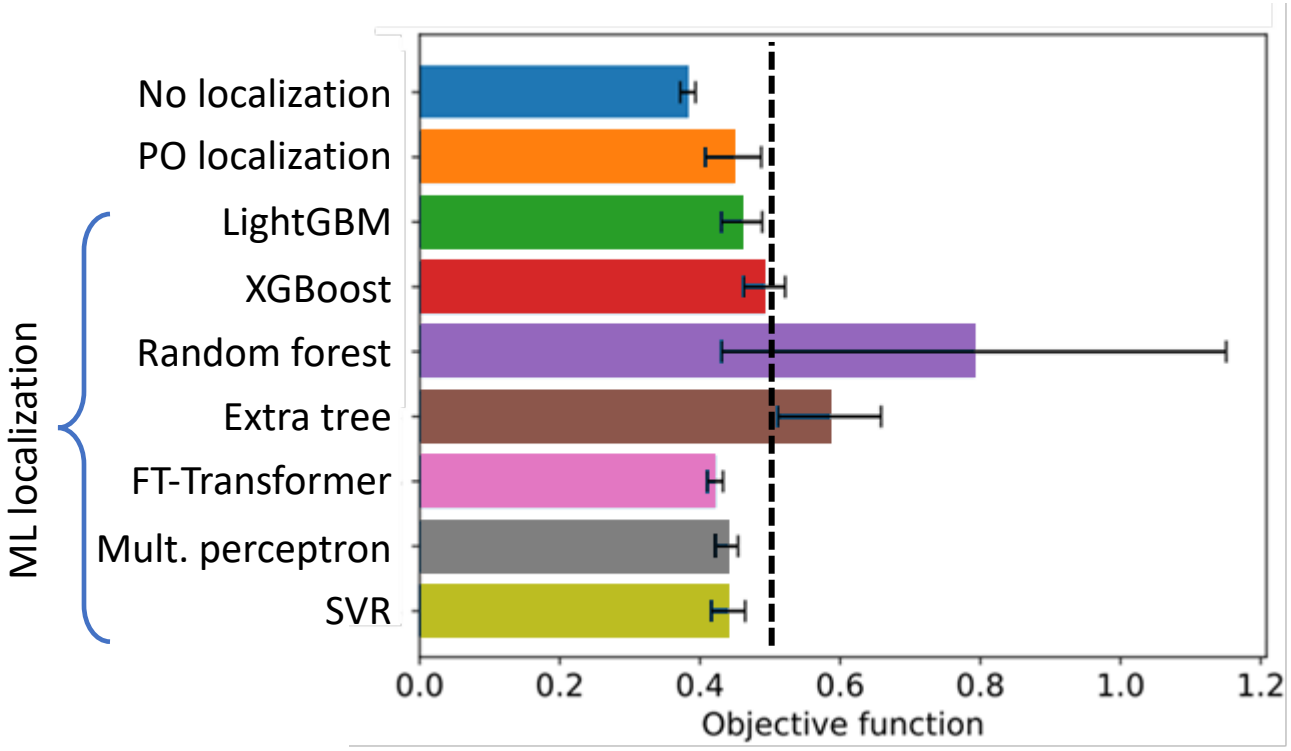
INTERNA

Data Assimilation Results

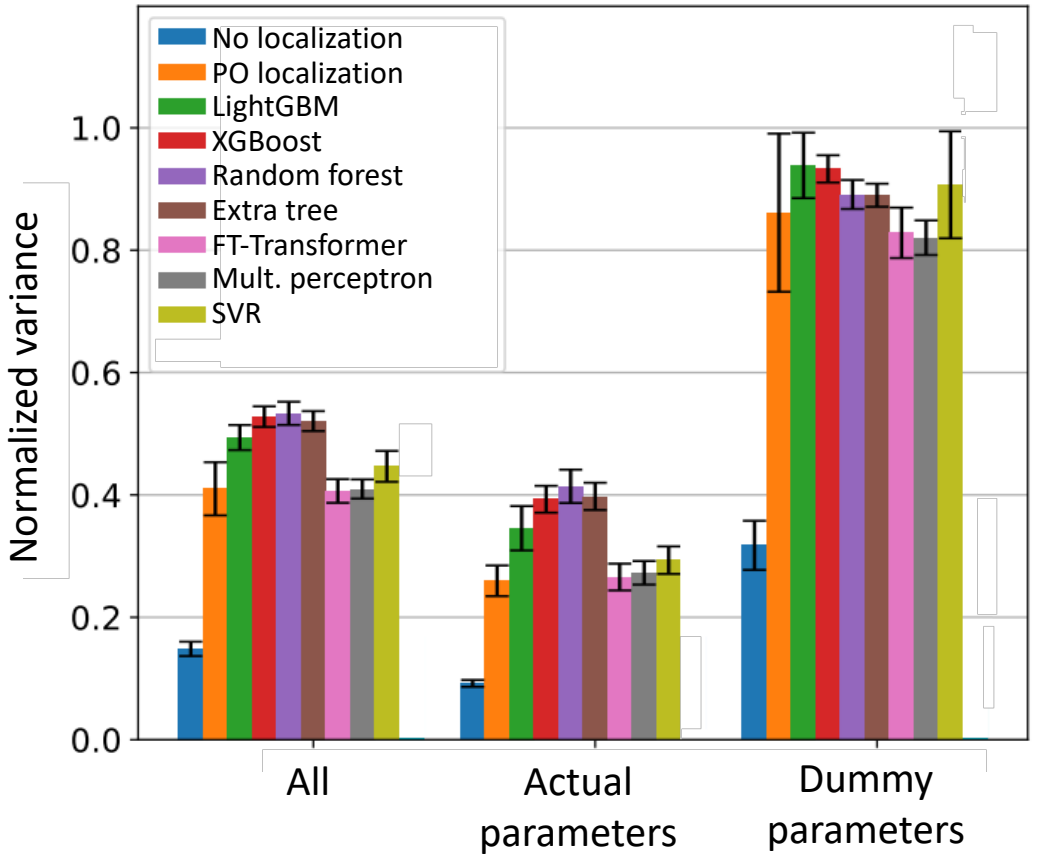
- 10 data assimilation experiments (N_e = *100*)

$$\mathrm{NV} = \frac{\mathrm{var}\left[\mathbf{m}_{\mathrm{post}}\right]}{\mathrm{var}\left[\mathbf{m}_{\mathrm{prior}}\right]}$$

Data Mismatch

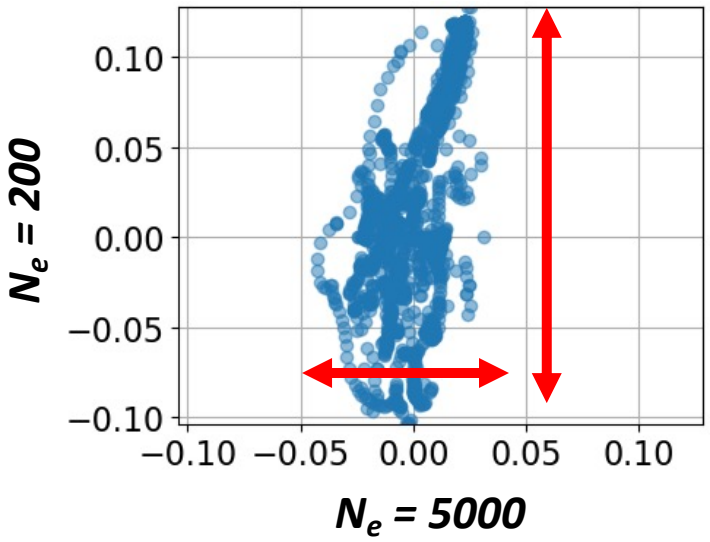


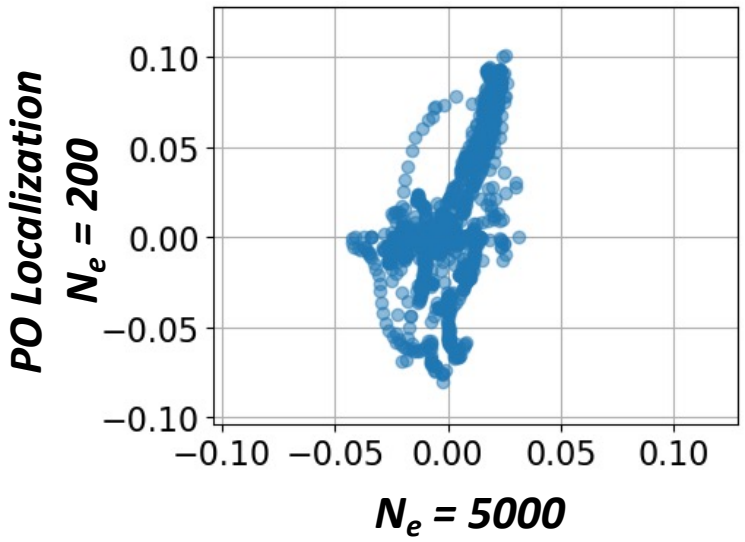
Normalized Variance

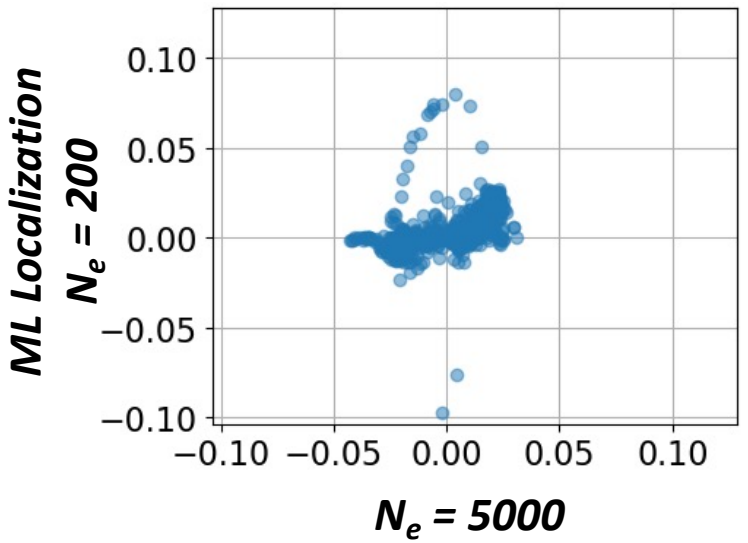


INTERNA

Correlations (Dummy vs Predicted Data)







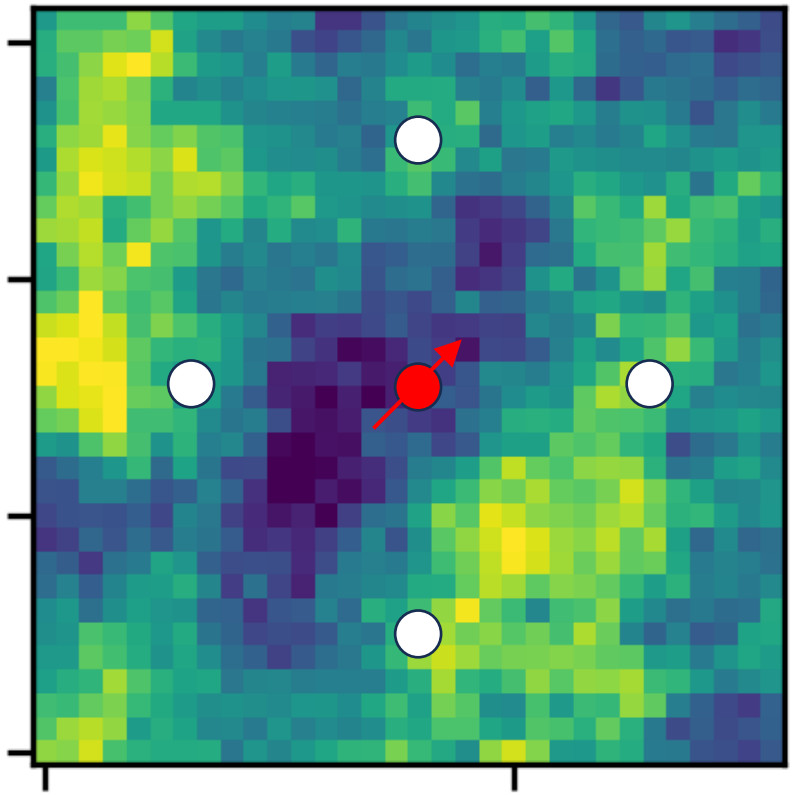
INTERNA

Part 2 – Localization for Grid Parameters

INTERNA

Test Case 1

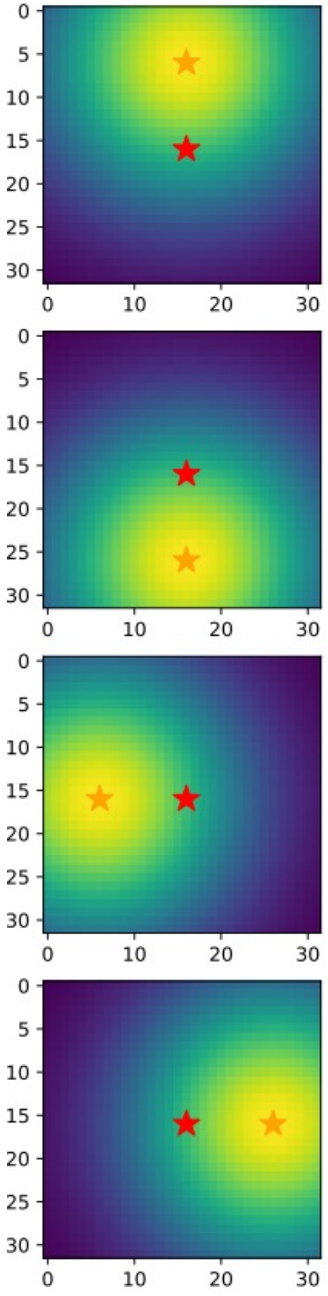
- CO_2 injection
- Pressure measurements at 4 monitor wells
- Estimate log-permeability distribution
- Ensemble with $N_e = 100$
- ML method: XGBoost



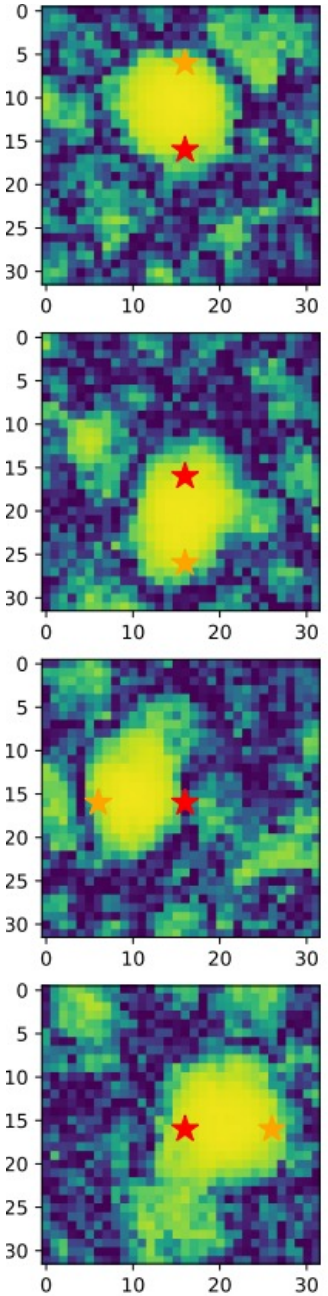
INTERNA

Localization Results

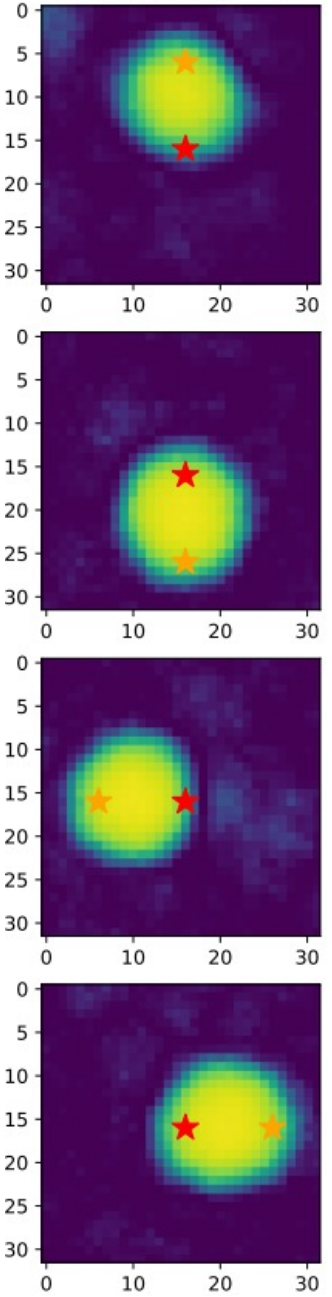
Distance-based localization



Pseudo-optimal localization



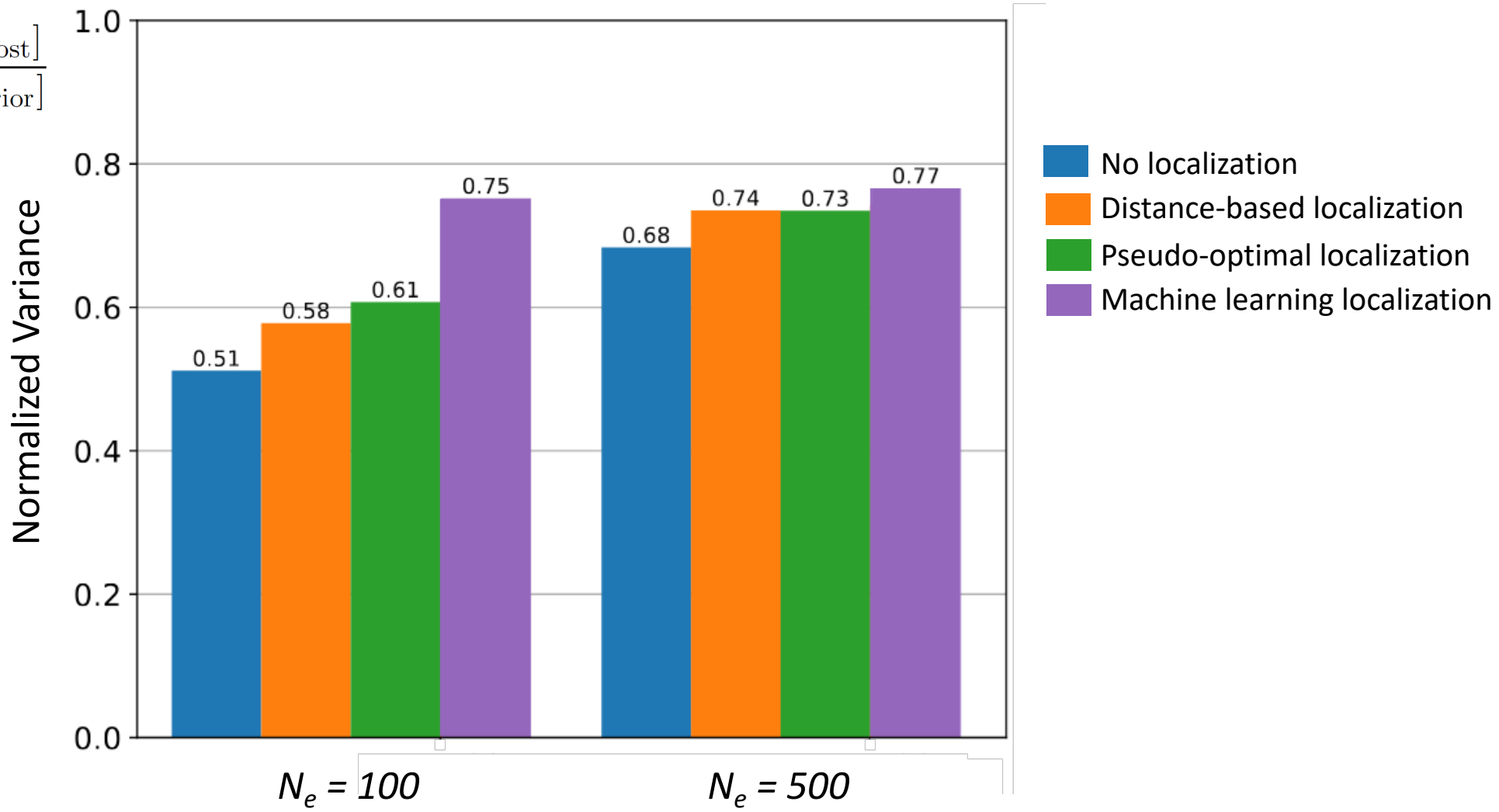
Machine learning localization



INTERNA

Data Assimilation Results

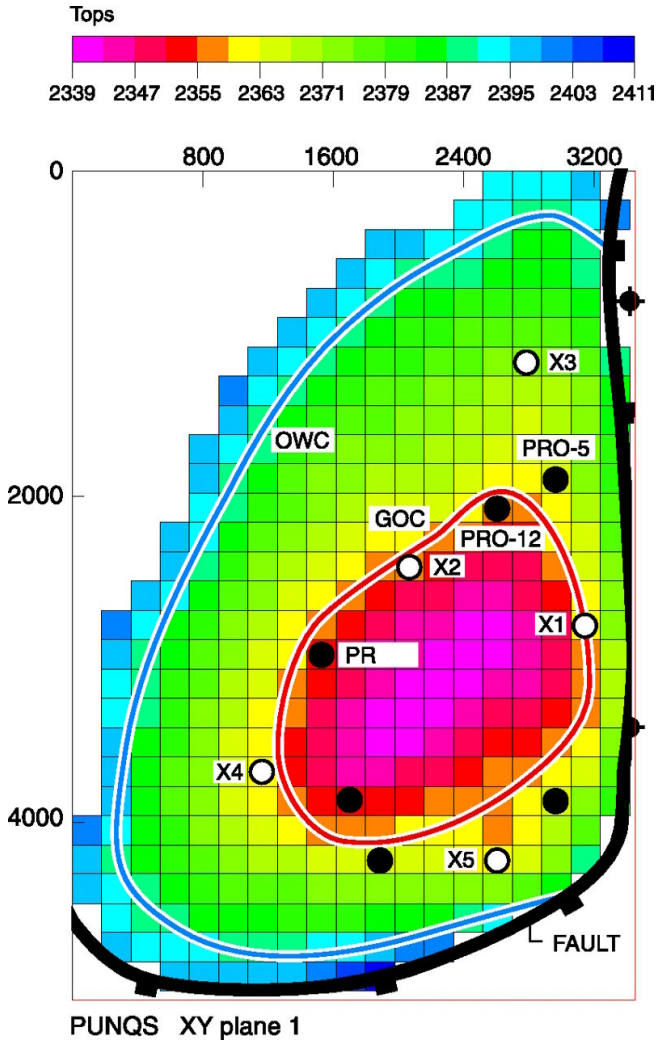
$$\mathrm{NV} = \frac{\mathrm{var}\left[\mathbf{m}_{\mathrm{post}}\right]}{\mathrm{var}\left[\mathbf{m}_{\mathrm{prior}}\right]}$$



INTERNA

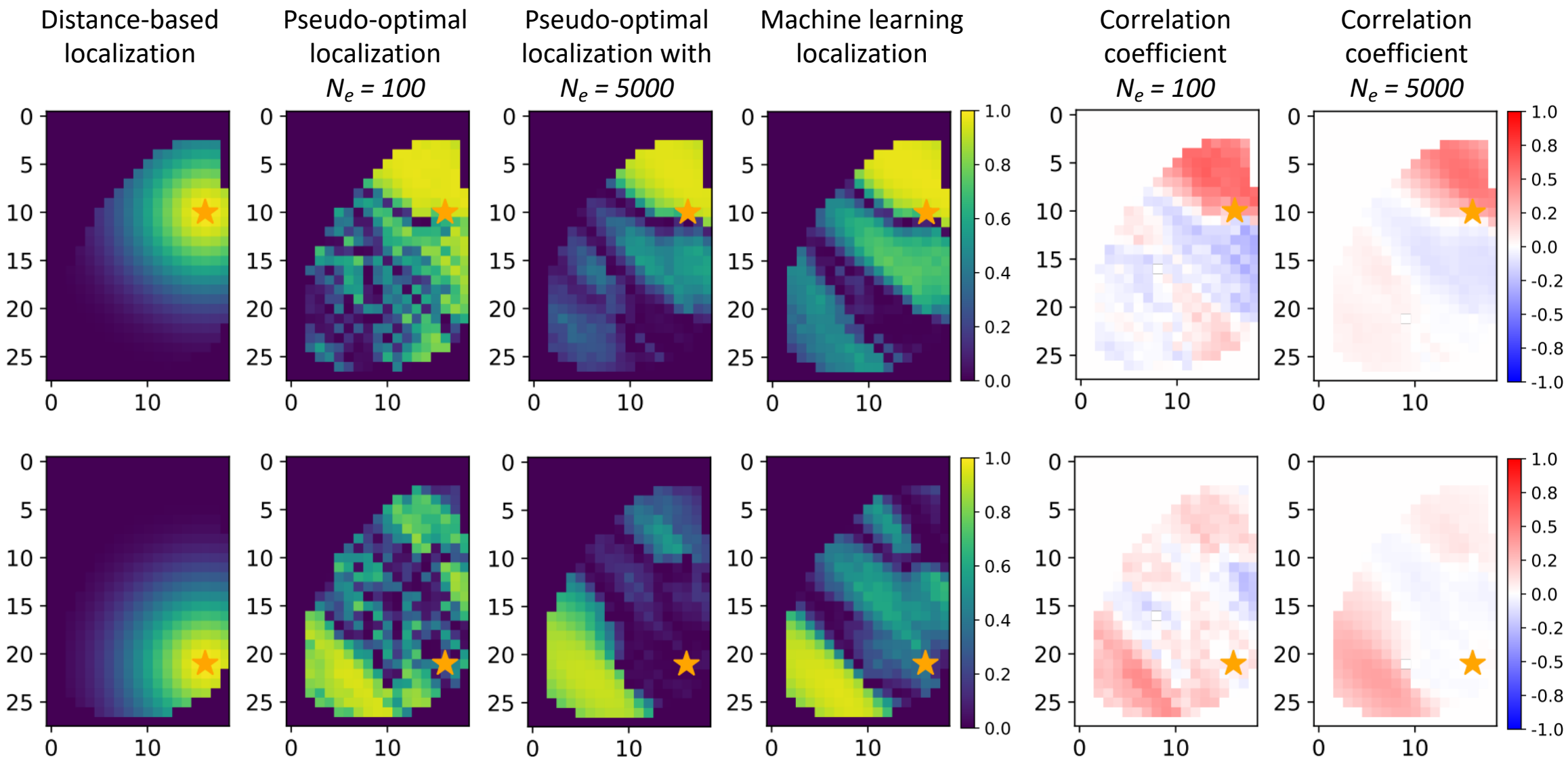
Test Case 2 (PUNQ-S3 modified model)

- Estimate porosity, horizontal and vertical permeability distribution
- Measurements of water cut, gas-oil ratio, and pressure
- Ensemble with $N_e = 100$
- ML method: LightGBM



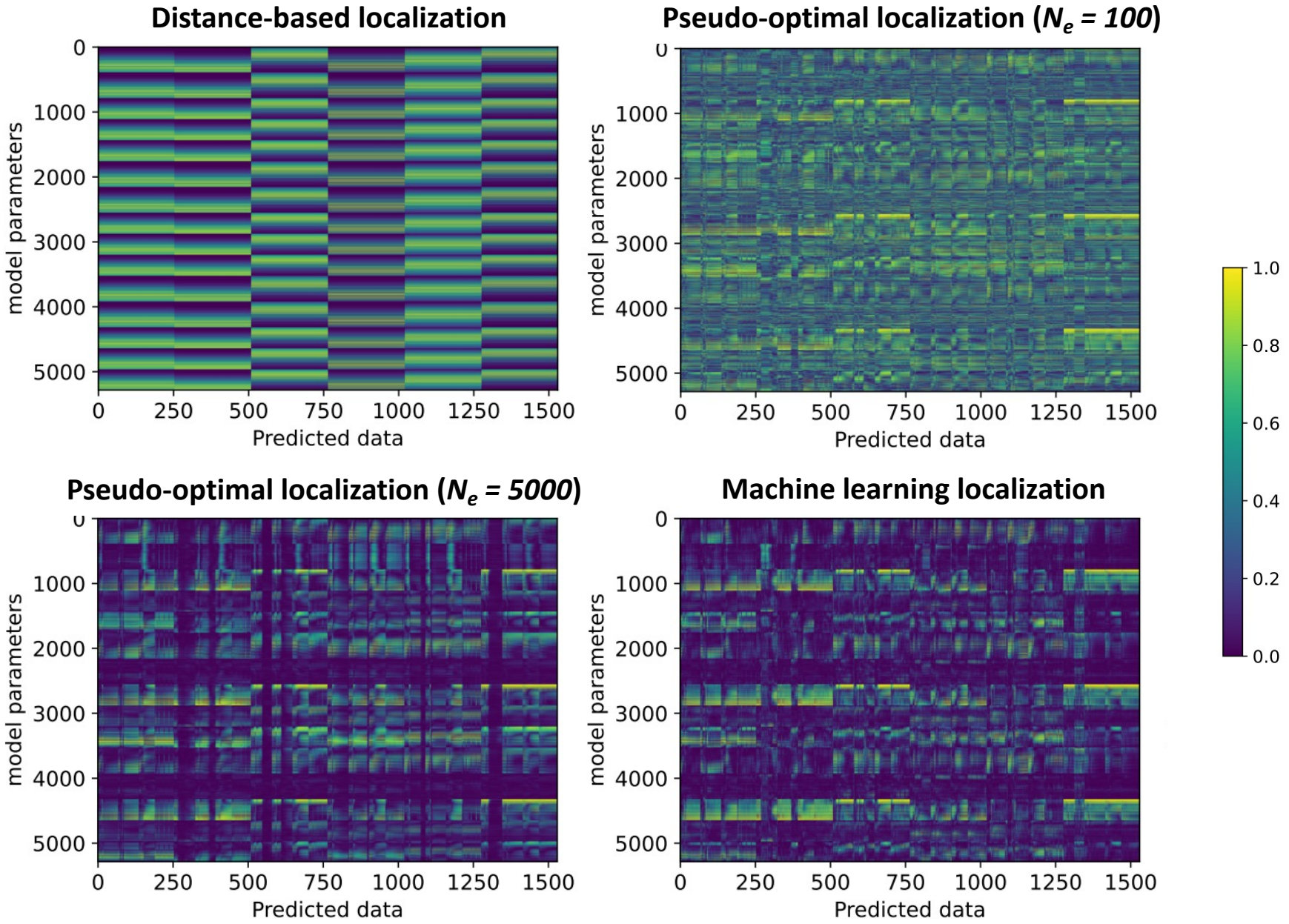
INTERNA

Localization Results – Pressure



INTERNA

Localization Matrix

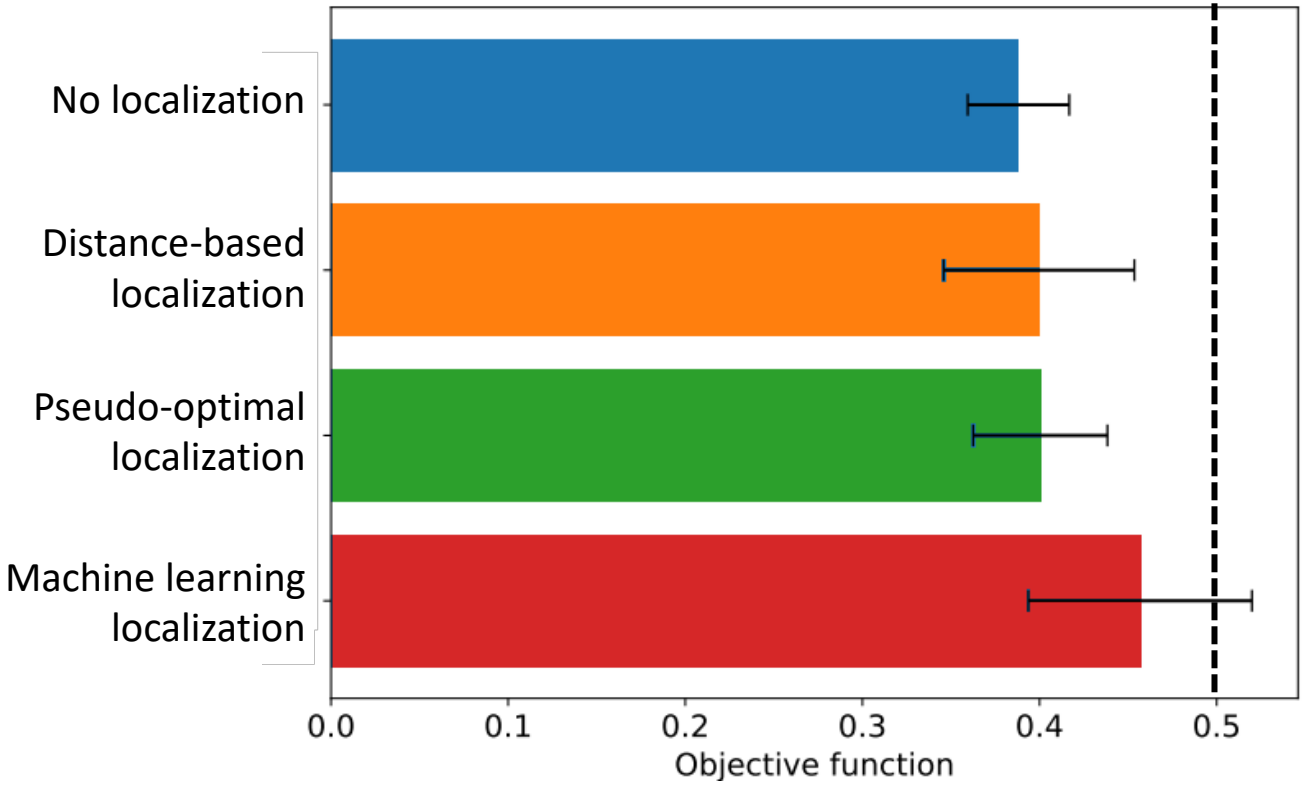


INTERNA

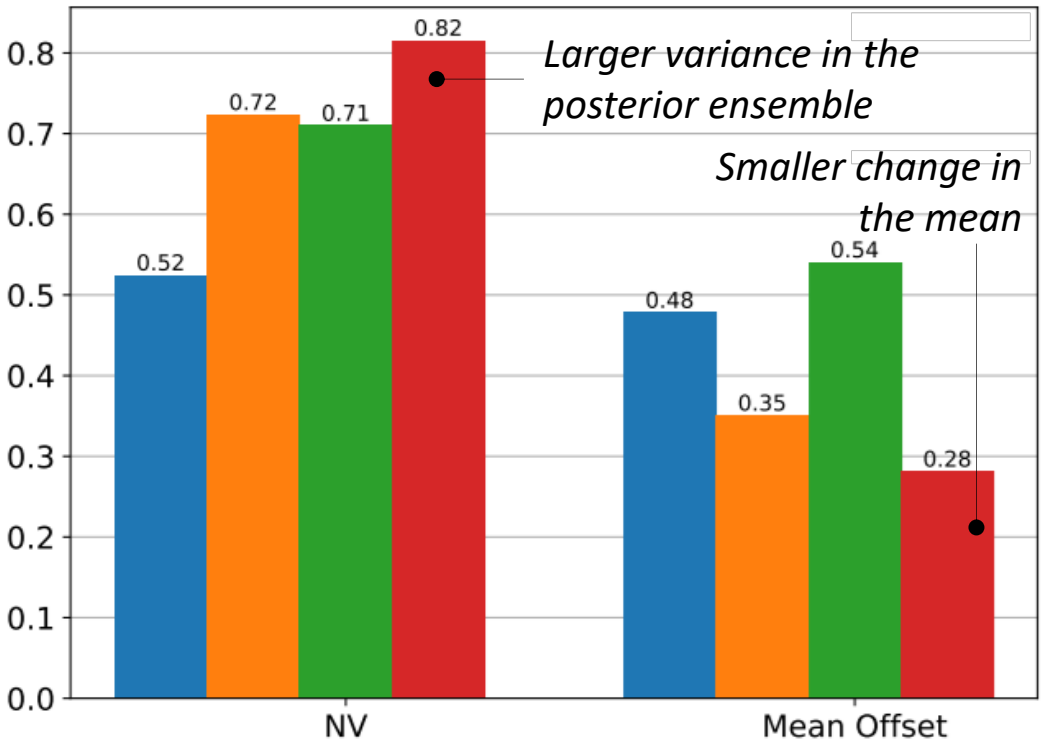
Data Assimilation Results

- 10 data assimilation experiments ($N_e = 100$)

Data Mismatch



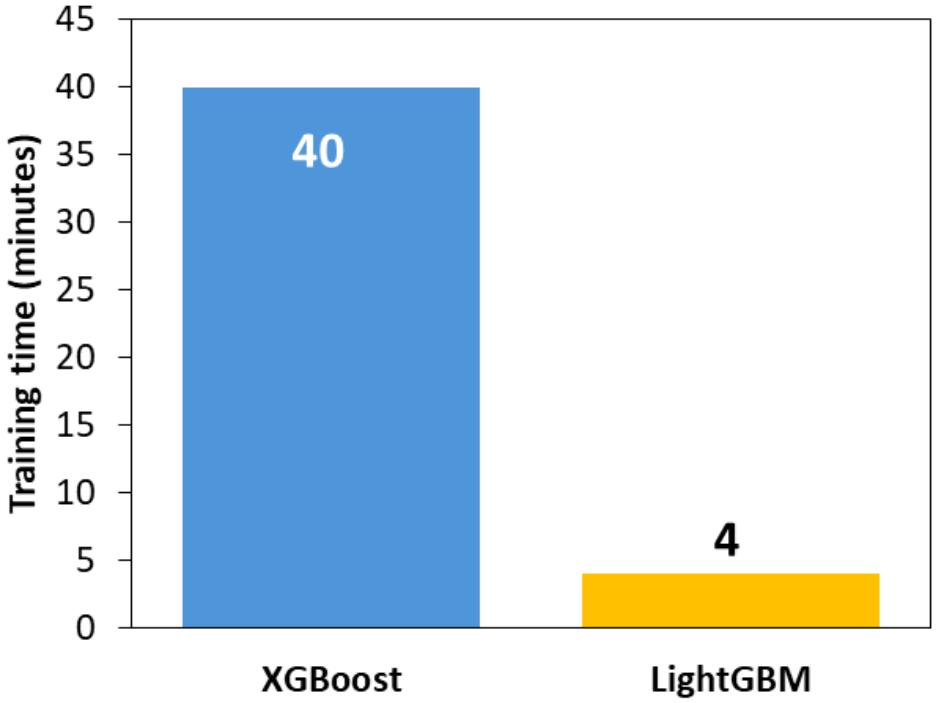
Normalized Variance and Mean Offset



INTERNA

Training Time*

- Number of model parameters: 5238
- Number of data points: 1530
- CPU Intel(R) Xeon(R) with 16 cores



*Preliminary result

INTERNA

Final Comments

- Distance-free localization scheme
- No additional forward simulations
- Out-of-box machine learning implementations (XGBoost or LightGBM)
- No additional tuning parameters
- Scalability for large scale problems needs further investigation

INTERNA

References

- Furrer, R., & Bengtsson, T.: **Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants**, *Journal of Multivariate Analysis*, *98*(2), 2007

- Floris, F.J., Bush, M.D., Cuypers, M., Roggero, F., & Syversveen, A.R**.: Methods for Quantifying the Uncertainty of Production Forecasts: a Comparative Study**, *Petroleum Geoscience*, 7(S), 2001

- Silva, V.L.S., Seabra, G.S. & Emerick, A.A.: **Machine Learning to Enhance the Covariance Estimations of Non-Local Model Parameters in Ensemble Based-Data Assimilation**, *Proceedings of the SPE Reservoir Simulation Conference*, SPE-223908-MS, 2025

- Silva, V.L.S., Seabra, G.S. & Emerick, A.A.: **Mitigating Loss of Variance in Ensemble Data Assimilation: Machine Learning-Based and Distance-Free Localizations for Better Covariance Estimation**, *In preparation*, 2020

INTERNA