Enhancing data assimilation and uncertainty quantification: machine learning for better covariance estimation

Vinicius L. S. Silva*
Petrobras, Rio de Janeiro, Brazil
v.santos-silva19@alumni.imperial.ac.uk

Gabriel S. Seabra
Petrobras, Rio de Janeiro, Brazil
g.serraoseabra@tudelft.nl

Alexandre A. Emerick
Petrobras, Rio de Janeiro, Brazil
emerick@petrobras.com.br

Abstract

Ensemble-based data assimilation is a powerful tools for updating geological reservoir models using dynamic data, with applications in hydrocarbon production, groundwater management, carbon storage, and geothermal energy. However, small ensemble sizes —due to the high computational cost of simulations— can introduce sampling errors and spurious correlations, leading to poor covariance estimations and degraded uncertainty quantification. Localization techniques help mitigate these effects by tapering updates based on the distance between observations and model parameters. Nonetheless, for cases lacking spatial relationships, distance-based localization is ineffective. To overcome these limitations, we propose a novel distance-free localization strategy using machine learning models tailored for tabular data. Additionally, we introduce a simple correction to the prior cross-covariance to improve localization in low-dimensional problems. Integrated into the Ensemble Smoother with Multiple Data Assimilation (ES-MDA), our methods were tested on scalar and grid-based parameters. Results show improved cross-covariance estimates and enhanced data assimilation performance for all cases (with/without spatial relationships), preserving ensemble variance while maintaining data match quality.

1 Introduction

Ensemble-based data assimilation methods, particularly iterative ensemble smoothers, are widely used in subsurface reservoir modeling to integrate dynamic data and reduce uncertainty in predictions [Chen and Oliver, 2013, Emerick and Reynolds, 2013, Luo et al., 2015, Silva et al., 2017, Raanes et al., 2019]. These methods, such as the Ensemble Smoother with Multiple Data Assimilation (ES-MDA) [Emerick and Reynolds, 2013], rely on ensembles of model realizations to estimate covariances between uncertain parameters and observed data. However, practical constraints on computational resources limit ensemble sizes, leading to sampling errors and spurious correlations. These errors can cause excessive variance reduction or ensemble collapse, compromising the reliability of the data assimilation process.

To mitigate these issues, localization techniques are employed [Houtekamer and Mitchell, 2001, Furrer and Bengtsson, 2007, Anderson, 2007, Bishop and Hodyss, 2009, Zhang and Oliver, 2010,

*The code and data for the machine learning-based localization (ML-localization) can be found at https://github.com/viluiz/ml_localization, and the full paper at https://arxiv.org/pdf/2506.13362

1st Workshop on Differentiable Systems and Scientific Machine Learning @ EurIPS 2025.

Luo et al., 2018, Vishny et al., 2024]. Traditional localization relies on spatial proximity between model parameters and observations, tapering updates based on distance between parameters and observed data [Emerick, 2025, Chap. 7]. Nonetheless, selecting an appropriate localization region is a problem-specific task that depends on the reservoir's dynamic behavior and the spatial data distribution. Furthermore, many key reservoir parameters lack these spatial relationship.

This paper introduces a novel distance-free localization strategy using machine learning (ML) tailored for tabular data. The approach leverages ML proxies to better estimate cross-covariances, enabling more accurate localization coefficients. Additionally, a simple analytical correction to the prior cross-covariance is proposed to enhance performance in low-dimensional settings.

2 Background

Ensemble data assimilation methods, particularly ES-MDA, are commonly used in subsurface reservoir engineering. ES-MDA performs multiple global updates, avoiding inconsistencies common in sequential approaches like the ensemble Kalman filter [Emerick and Reynolds, 2013]. Its iterative corrections improve robustness in nonlinear problems, making it particularly effective for data assimilation and uncertainty quantification of subsurface reservoirs. The ES-MDA update equation for iteration k is:

$$\mathbf{m}_j^{k+1} = \mathbf{m}_j^k + \tilde{\mathbf{C}}_{md}^k \left(\tilde{\mathbf{C}}_{dd}^k + \alpha_k \mathbf{C}_e \right)^{-1} \left(\mathbf{d}_{\text{obs}} + \mathbf{e}_j^k - \mathbf{d}_j^k \right). \tag{1}$$

For $j = 1, \ldots, N_e$, where N_e is the ensemble size. The update uses: $\mathbf{m}_j^k \in \mathbb{R}^{N_m}$ model parameters; $\mathbf{d}_{\text{obs}} \in \mathbb{R}^{N_d}$ observed data; $\mathbf{e}_j^k \sim \mathcal{N}(0, \alpha_k \mathbf{C}_{\text{e}})$ random perturbations; $\mathbf{C}_{\text{e}} \in \mathbb{R}^{N_d \times N_d}$ data-error covariance; $\mathbf{d}_j^k \in \mathbb{R}^{N_d}$ predicted data; $\mathbf{C}_{\text{md}}^k \in \mathbb{R}^{N_m \times N_d}$ cross-covariance; $\mathbf{C}_{\text{dd}}^k \in \mathbb{R}^{N_d \times N_d}$ auto-covariance. All covariances are estimated from the current ensemble (size N_e). The inflation coefficients α_k must be selected prior to the data assimilation such that: $\sum_{k=1}^{N_a} \frac{1}{\alpha_k} = 1$. Where N_a is the number of iterations.

Localization modifies this update by applying a Schur product (as a tapering) with a localization matrix $\mathbf{R} \in \mathbb{R}^{N_m \times N_d}$:

$$\mathbf{m}_j^{k+1} = \mathbf{m}_j^k + \mathbf{R} \circ \left[\tilde{\mathbf{C}}_{md}^k \left(\tilde{\mathbf{C}}_{dd}^k + \alpha_k \mathbf{C}_e \right)^{-1} \left(\mathbf{d}_{\text{obs}} + \mathbf{e}_j^k - \mathbf{d}_j^k \right) \right]. \tag{2}$$

The most common localization used is distance-based, where r_{ik} is computed via a compact correlation function based on the distance between the ith parameter and kth data point, typically using the Gaspari-Cohn function [Gaspari and Cohn, 1999]. For parameters that do not have a distance relationship, one of the best-performing methods is he pseudo-optimal (PO) localization Furrer and Bengtsson [2007], Lacerda et al. [2019], that estimates localization coefficients without spatial information: $r_{ik} = \frac{c_{ik}^2}{c_{ik}^2 + \frac{c_{ii} c_{kk}}{N_e}}$.

where $c_{i,k}$ represents the covariance between the i-th model parameter and the k-th predicted data point. However, PO-localization struggles with small ensembles sizes.

3 Proposed Methods

3.1 ML-Based Localization

The core idea is to estimate the cross-covariance from a large ensemble generated by a ML surrogate instead of the numerical reservoir simulations. The goal is to improve localization accuracy requiring no additional reservoir simulations beyond those already performed during the data assimilation process. The method is as follows:

Algorithm 1: ML-localization

1. Use prior ensemble $\{\mathbf{m}_j, \mathbf{d}_j\}_{j=1}^{N_e}$ as training data.
2. Train ML proxy $G(\cdot)$ to predict data from parameters.

3. Generate large ensemble $\{\mathbf{m}_j\}_{j=1}^{N_E}$, where $N_E \gg N_e$.
4. Predict data: $\hat{\mathbf{d}}_j = G(\mathbf{m}_j)$.
5. Estimate cross-covariance $\hat{\mathbf{C}}_{md}$ and compute localization coefficients using PO-localization: $r_{ik} = \frac{\hat{c}_{ik}^2}{\hat{c}_{ik}^2 + \frac{\hat{c}_{ii}\hat{c}_{kk}}{N_e}}$, set $r_{ik} = 0$ if $|\hat{c}_{ik}| < \eta\sqrt{\hat{c}_{ii}\hat{c}_{kk}}$

An analytical correction of the cross-covariance based on the auto-covariance of the model parameters is also proposed. This improves cross-covariance estimation without requiring ML proxies and is particularly effective for low-dimensional problems. More details please see Silva et al. [2025].

4 Results

4.1 Scalar Parameters – PUNQ-S3 Case

The PUNQ-S3 3D reservoir model is a benchmark case developed to evaluate the performance of data assimilation methods, particularly in terms of uncertainty quantification [Floris et al., 2001, Lacerda et al., 2019]. The observed data include measurements of well water cut (WWCT), well gas-oil ratio (WGOR), and well bottom-hole pressure (WBHP) over 8 years period. The total number of data points is 1,530. Using 15 scalar parameters and 5 dummy parameters, the methods were evaluated on cross-correlation accuracy and data assimilation performance. We use an ensemble of $N_e = 100$ simulations and $N_E = 5000$ for the ML-localization.

In terms of cross-correlation accuracy, we compare the localized correlations presented in Table 1 with a gold standard correlation generated running 5000 simulations (with is usually unfeasible in practice). We present the norms of the difference correlation matrix in Table 1.

Table 1: Localization performance: cross-correlation difference

Method	Frobenius Norm	Spectral Norm
No Localization	0.053	6.105
PO-localization	0.047	5.904
ML-localization (LightGBM)	0.040	5.303

We also run the data assimilation 10 times and evaluate the localization performance for different ML methods. Figure 1 shows the data assimilation results. The plot on the left shows the values of the data mismatch, while the plot on the right displays the normalized variance (NV: posterior variance divided by prior variance), for dummy parameters (NV_d), non-dummy parameters (NV_nd), and all parameters combined (NV_t). The results indicate that ML algorithms based on ensemble of decision trees (Random Forest, Extra Trees, XGBoost, and LightGBM), yielded higher NV values for the non-dummy parameters and NV values close to one for the dummy parameters. However, Random Forest and Extra Trees also produced higher data mismatch objective values, both in terms of mean and standard deviation, suggesting reduced robustness compared to the other two methods. Among all ML methods tested in this study, those based on gradient boosting decision trees (XGBoost and LightGBM) showed the best overall performance.

The superior performance of ensemble of decision tree models is consistent with the findings of Grinsztajn et al. [2022], typically in most reservoir applications, uninformative features are present, and the dataset is not rotationally invariant. Figure 3 in the Appendix shows the data match for one well in one data assimilation (same random seed).

4.2 Grid Parameters – CCS Model

A 2D CO_2 injection model with 1,024 grid parameters is also used [Seabra et al., 2024]. The observations consist of pressure measurements at four monitoring well locations over a two-year period, resulting in a total of 96 data points. We perform the data assimilation using the ES-MDA with four iterations. We use an $N_E = 5000$ for the ML-localization. Fig. 2 shows the values of NV obtained after data assimilation with different ensemble sizes. Here, we also test a distance-based localization (DB-localization) since the grid parameters have a distance relationship with the observed data from the wells. For ML-localization we used the XGBoost method. The results in this figure

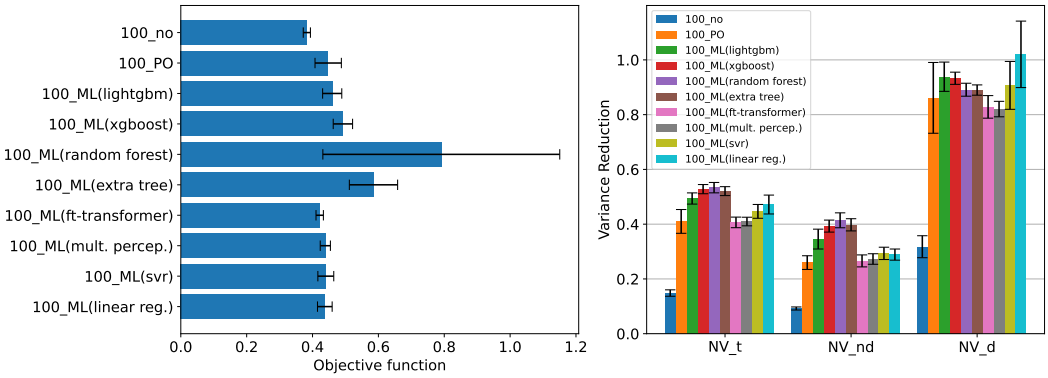

Figure 1: Data assimilation results with $N_e = 100$, comparing ML-localization using different machine learning methods. The labels NV_t, NV_nd, and NV_d denote the normalized variance of all parameters, non-dummy parameters, and dummy parameters, respectively.

indicate that ML-localization resulted in larger NV for all ensemble sizes. In terms of data match quality, all methods resulted in similar values of data mismatch. This is illustrated in Fig. 4 in the Appendix, which shows the predicted well bottom-hole pressure (WBHP) for all localization cases with an ensemble of 100 realizations.

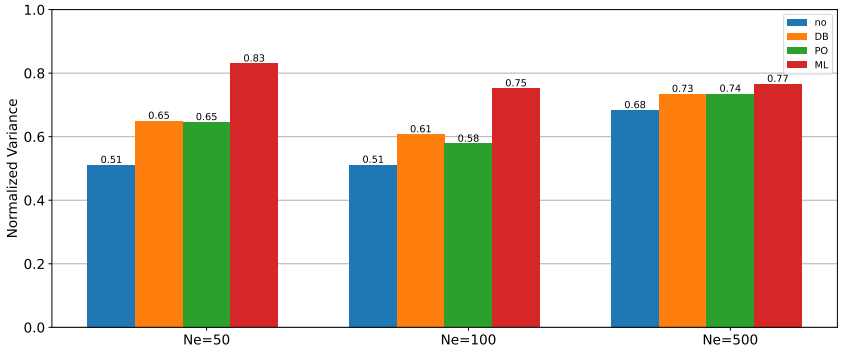

Figure 2: Normalized variance results with varying ensemble sizes for the CCS Model. This figure includes the results of data assimilation without localization (no), DB-localization (DB), PO-localization (PO), and ML-localization (ML).

Figure 5 in the Appendix shows the values of the localization in the simulation grid. Each image includes the location of the CO_2 injection well (red star) and the corresponding monitoring well (orange star). Distance-based localization produces a smooth spatial distribution, with values decaying from one at the data location to zero in regions far from the well. PO-localization produces patterns where the regions between injection and monitoring wells are clearly emphasized. However, these maps also show high variability in areas far from the wells, likely due to spurious correlations. In contrast, ML-localization effectively suppresses distant correlations, yielding clean localization maps concentrated along the region connecting each pair of wells.

5 Conclusion

This work presents a machine learning-based localization method for ensemble data assimilation that avoids spatial assumptions and improves covariance estimation. Combined with a simple prior correction, the approach enhances performance in both scalar and grid-based reservoir models. ML-localization retains ensemble variance, reduces spurious updates, and matches data effectively, even with small ensembles. The method is easy to implement, requires no additional simulations, and scales well with problem size, making it suitable for practical reservoir applications.

References

J. L. Anderson. Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D: Nonlinear Phenomena*, 230(1–2):99–111, 2007. doi: 10.1016/j.physd.2006.02.011.

C. H. Bishop and D. Hodyss. Ensemble covariances adaptively localized with ECO-RAP. part 1: tests on simple error models. *Tellus*, 61:84–96, 2009. doi: 10.1111/j.1600-0870.2008.00371.x.

Y. Chen and D. S. Oliver. Levenberg-Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Computational Geosciences*, 17:689–703, 2013. doi: 10.1007/s10596-013-9351-5.

A. A. Emerick. *Ensemble Data Assimilation Applied to Geological Reservoir Models*. Petrobras, first edition, 2025. ISBN 978-6588763292. URL `https://publicacoesup.petrobras.com.br/peld/catalog/book/54`.

A. A. Emerick and A. C. Reynolds. Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 55:3–15, 2013. doi: 10.1016/j.cageo.2012.03.011.

F. J. T. Floris, M. D. Bush, M. Cuypers, F. Roggero, and A. R. Syversveen. Methods for quantifying the uncertainty of production forecasts: a comparative study. *Petroleum Geoscience*, 7(SUPP): 87–96, 2001. doi: 10.1144/petgeo.7.S.S87.

R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007. doi: 10.1016/j.jmva.2006.08.003.

G. Gaspari and S. E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757, 1999. doi: 10.1002/qj.49712555417.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 2022.

P. L. Houtekamer and H. L. Mitchell. A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129(1):123–137, 2001. doi: 10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

J. M. Lacerda, A. A. Emerick, and A. P. Pires. Methods to mitigate loss of variance due to sampling errors in ensemble data assimilation with non-local model parameters. *Journal of Petroleum Science and Engineering*, 172:690–706, 2019. doi: 10.1016/j.petrol.2018.08.056.

X. Luo, A. S. Stordal, R. J. Lorentzen, and G. Nævdal. Iterative ensemble smoother as an approximate solution to a regularized minimum-average-cost problem: Theory and applications. *SPE Journal*, 20(5), 2015. doi: 10.2118/176023-PA.

X. Luo, T. Bhakta, and G. Nævdal. Correlation-based adaptive localization with applications to ensemble-based 4D-seismic history matching. *SPE Journal*, 23(2):396–427, 2018. doi: 10.2118/185936-PA.

P. N. Raanes, A. S. Stordal, and G. Evensen. Revising the stochastic iterative ensemble smoother. *Nonlinear Processes in Geophysics*, 26(3), 2019. doi: 10.5194/npg-2019-10.

G. S. Seabra, N. T. Mücke, V. L. S. Silva, D. Voskov, and F. C. Vossepoel. AI enhanced data assimilation and uncertainty quantification applied to geological carbon storage. *International Journal of Greenhouse Gas Control*, 136, 2024. doi: 10.1016/j.ijggc.2024.104190.

V. L. Silva, G. S. Seabra, and A. A. Emerick. Mitigating loss of variance in ensemble data assimilation: machine learning-based and distance-free localizations for better covariance estimation. *arXiv preprint arXiv:2506.13362*, 2025.

V. L. S. Silva, A. A. Emerick, P. Couto, and J. L. D. Alves. History matching and production optimization under uncertainties – application of closed-loop reservoir management. *Journal of Petroleum Science and Engineering*, 157:860–874, 2017. doi: 10.1016/j.petrol.2017.07.037.

D. Vishny, M. Morzfeld, K. Gwirtz, E. Bach, O. R. A. Dunbar, and D. Hodyss. High-dimensional covariance estimation from a small number of samples. *Journal of Advances in Modeling Earth Systems*, 16(9), 2024. doi: 10.1029/2024MS004417.

Y. Zhang and D. S. Oliver. Improving the ensemble estimate of the Kalman gain by bootstrap sampling. *Mathematical Geosciences*, 42:327–345, 2010. doi: 10.1007/s11004-010-9267-8.

A Data Assimilation Results

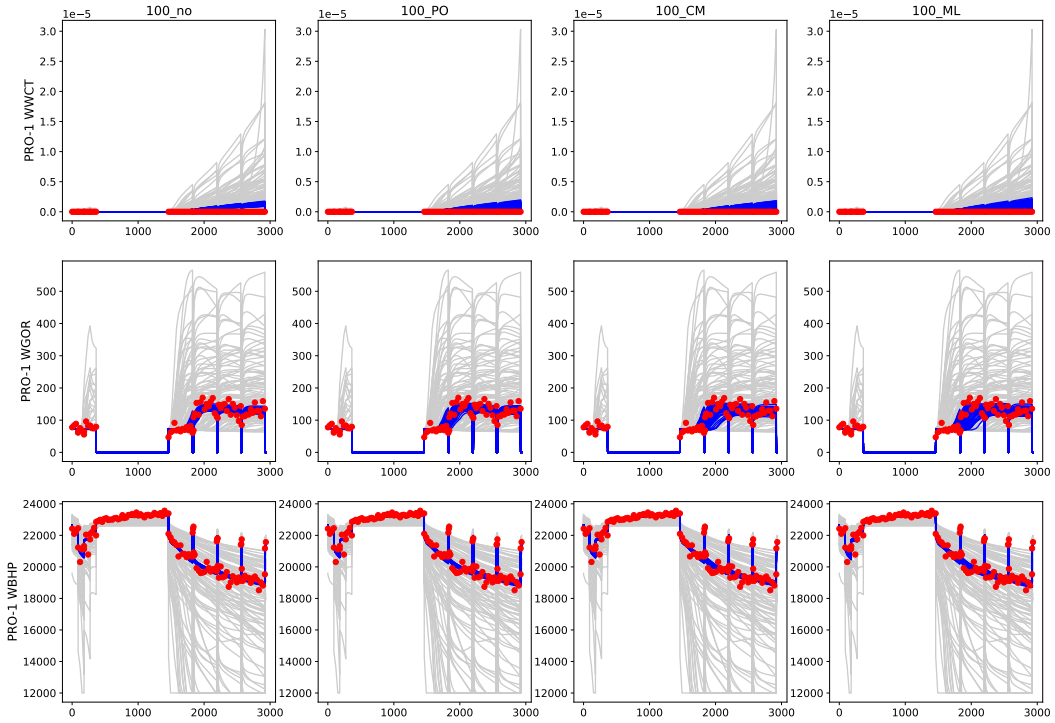

Figure 3: Predicted data for well PRO-1 obtained using the prior ensemble (gray curves) and posterior ensemble (blue curves) for different localization strategies. All results are based on an ensemble size of 100. The first column (100_no) shows the results without localization, the second column (100_PO) shows results with PO-localization, the third column (100_CM) presents results with CM-localization, and the fourth column (100_ML) shows results with ML-localization. The red dots represent the observed data.

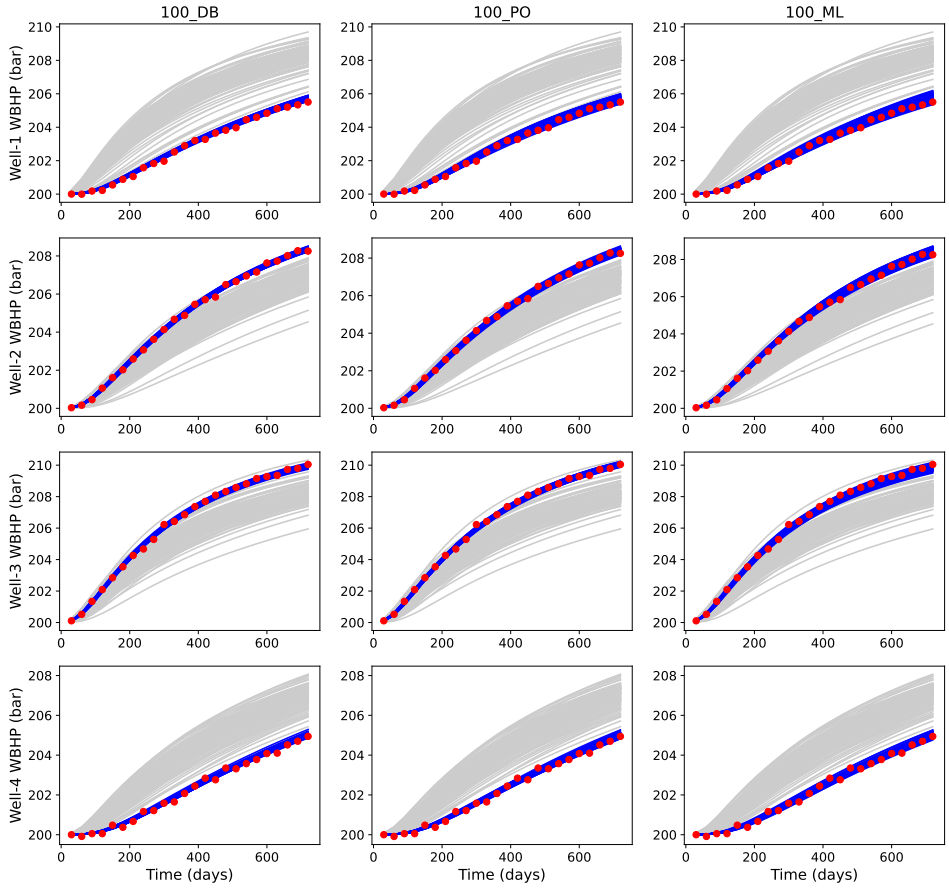

Figure 4: Predicted WBHP data before and after data assimilation for an ensemble size of $N_e = 100$. Each row represents a monitoring well and each column a different localization scheme. The red dots represent the observed data, the gray curves the prior ensemble, and the blue curves the posterior ensemble.

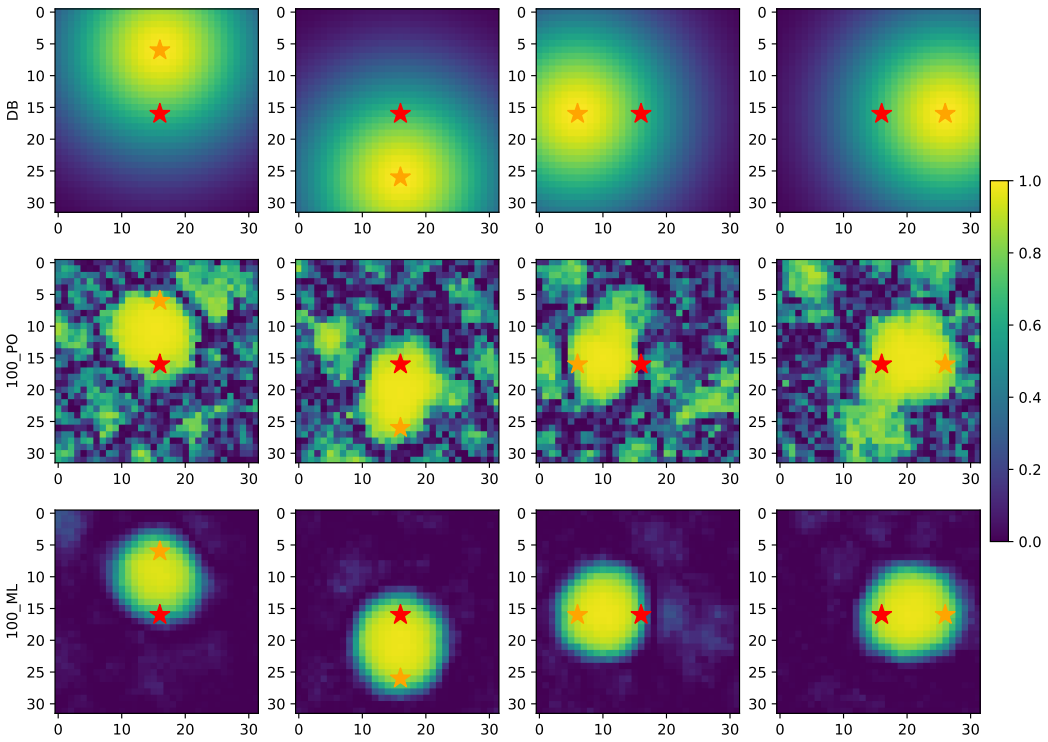

Figure 5: Localization values ($N_e = 100$) for the CCS model. Each column corresponds to a different monitoring well, and each row shows results from a different localization scheme. The red star indicates the CO_2 injection well, and the orange stars denote the positions of the monitoring wells.