



Delft University of Technology

Document Version

Final published version

Licence

CC BY

Citation (APA)

Seabra, G. S., Mücke, N. T., Santos Silva, V. L., Emerick, A. A., Voskov, D., & Vossepoel, F. C. (2026). Integrating score-based diffusion models with machine learning-enhanced localization for advanced data assimilation in geological carbon storage. *Computational Geosciences*, 30(4), Article 68. <https://doi.org/10.1007/s10596-026-10462-1>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

RESEARCH

Integrating score-based diffusion models with machine learning-enhanced localization for advanced data assimilation in geological carbon storage

Gabriel Serrão Seabra^{1,2} · Nikolaj T. Mücke^{1,3} · Vinicius Luiz Santos Silva^{2,4} · Alexandre A. Emerick² · Denis Voskov^{1,5} · Femke C. Vossepoel¹

Received: 3 December 2025 / Accepted: 24 June 2026

© The Author(s) 2026

Abstract

Accurate characterization of subsurface heterogeneity is important for the safe and effective implementation of geological carbon storage (GCS) projects. This paper explores how machine learning methods can enhance data assimilation for GCS with a framework that integrates score-based diffusion models with machine learning-enhanced localization in channelized reservoirs during CO₂ injection. We employ a machine learning-enhanced localization framework that uses large ensembles ($N_s = 5000$) with permeabilities generated by the diffusion model and states computed by simple ML algorithms to improve covariance estimation for the Ensemble Smoother with Multiple Data Assimilation (ESMDA). We apply ML algorithms to a prior ensemble of channelized permeability fields, generated with the geostatistical model FLUVSIM. Our approach is applied on a CO₂ injection scenario simulated using the Delft Advanced Research Terra Simulator (DARTS). Our ML-based localization maintains significantly more ensemble variance than when localization is not applied, while achieving comparable data-matching quality. This framework has practical implications for GCS projects, helping improve the reliability of uncertainty quantification for risk assessment.

Keywords Score-Based diffusion models · Data assimilation · Geological carbon storage · Uncertainty quantification · ESMDA · DARTS

1 Introduction

Geological carbon storage (GCS) is a promising technology for achieving large-scale carbon dioxide removal from the atmosphere, which is required to meet global climate targets [1]. The Intergovernmental Panel on Climate Change (IPCC) estimates that between 350 and 1,200 gigatons of CO₂ must be captured and stored by 2100 to limit global warming to 1.5°C above pre-industrial levels. This unprecedented scale demands advanced methodologies for characterizing and monitoring subsurface storage complexes to ensure safe and permanent sequestration while maximizing storage efficiency [2]. Data assimilation provides the mechanism to combine prior geological models with monitoring data to reduce uncertainty and improve forecasts, but practical deployments are constrained by computational resources as each assimilation cycle requires many forward simulations. This motivates adding machine learning components that assist DA without replacing the governing physics, for example by generating large prior ensembles and improving localization for small N_e . Long-term experience from Sleipner [3] and established best practices [4] inform the monitoring context in which we develop and test the approach.



Wells typically sample a small part of the reservoir volume, yet decisions affecting containment must be made based on this limited information. This data scarcity becomes particularly challenging in channelized formations, which can be attractive storage targets because of their high injectivity, where complex connectivity patterns defy simple statistical descriptions [5].

Existing geostatistical methods face inherent limitations when confronted with channelized systems. Two-point statistics cannot capture curvilinear channel geometries, whereas multiple-point statistics require training images that may poorly represent the true subsurface heterogeneity [6, 7]. When these prior models are updated through ensemble-based data assimilation, the limited number of ensemble members exacerbates these problems. Small ensemble sizes (typically $N_e \sim 50 - 200$ in operational settings) lead to poorly estimated covariance matrices with significant sampling errors, causing the assimilation update to introduce geologically implausible features, such as spurious high-permeability patches. In practice, N_e is capped by simulation cost per member, parallel scalability limits, and project turnaround requirements. The insufficient ensemble size results in artificial blurring of sharp geological boundaries, as the linear update mechanism cannot preserve the discrete nature of channel-background contrasts. This geological degradation becomes more severe with smaller ensembles, where sampling errors dominate the true covariance structure, ultimately producing posterior realizations that bear little resemblance to the complex channelized patterns that control the subsurface flow behavior. Although object-based geological models, such as FLUVSIM [8], can generate geologically realistic channelized systems that address these representational limitations, they are computationally expensive and time-consuming to run, making it impractical to generate the hundreds or thousands of realizations required on-the-fly during data assimilation workflows.

Data assimilation methods provide the mathematical framework for combining these prior geological models with observed data to reduce uncertainty in reservoir characterization [9, 10]. The fundamental approaches can be broadly categorized into variational methods, which solve optimization problems to find the most probable model realization [11], and ensemble methods, which use an ensemble to approximate probability distributions [12]. Ensemble methods quantify uncertainty through their sample-based representation, making them suitable for GCS applications where understanding the range of possible outcomes is as important as identifying the most likely scenario [13]. The evolution from the original Ensemble Kalman Filter [14] to smoothers and iterative variants reflects ongoing efforts to handle increasingly complex nonlinear problems while maintaining computational feasibility [15, 16]. Advances address the challenge of limited ensemble sizes through localization techniques [17] and iterative schemes that linearize nonlinear problems [18]. However, these methods still face challenges when applied to highly heterogeneous channelized systems, where traditional assumptions about spatial correlation and Gaussian distributions break down.

The Ensemble Smoother with Multiple Data Assimilation (ESMDA) is a common method of choice for practical reservoir characterization due to its computational efficiency and ability to handle nonlinear forward models [12, 15]. However, ESMDA has Gaussian assumptions that channelized permeability fields severely violate, and this can exacerbate the fact that small ensembles lose their variance, causing posterior models to be less uncertain than they should be, which affects the reliability of uncertainty quantification. The industry has widely adopted distance-based localization methods [17, 19, 20] as the primary solution to combat ensemble collapse. Covariance estimation in high-dimensional systems [21] and distance based methods, including the popular Gaspari-Cohn correlation function, are standard practice in several operational reservoir history matching systems. While those approaches suppress spurious correlations arising from sampling errors, they rely on an assumption that correlation strength decreases with physical distance — an assumption that fails in channelized systems. A circular or elliptical localization kernel cannot distinguish between updating along a high-permeability channel versus across a low-permeability barrier, leading to geologically inconsistent modifications that compromise long-term predictions. Despite these limitations, distance-based localization remains the industry standard due to its computational efficiency and ease of implementation. Recent work has shown the benefits of applying covariance localization to field cases [22] and comparing different adaptive localization methods for reservoir characterization [23].

Score-based diffusion models, generative AI models that learn to approximate the gradient of the log-density ($\nabla_x \log p_t(x)$) and can iteratively denoise random noise into realistic samples [24, 25], offer a promising approach for ensemble generation by training on established geological modeling outputs, such as those from FLUVSIM [8], to capture the complex spatial patterns and connectivity structures inherent in channelized systems. This capability addresses a need in ensemble-based data assimilation, where accurate covariance estimation benefits from much larger sample sizes than typically computationally feasible [21, 26]. Once trained, diffusion models can rapidly generate large numbers of geologically consistent realizations on demand, providing super-ensembles for enhanced covariance estimation without requiring additional, expensive reservoir simulations. The diffusion model alone is not sufficient to estimate covariances: it only provides parameter samples, and obtaining the parameter/data cross-covariance still requires a forward map. In this work that map is supplied by a fast ML surrogate trained on the working ensemble, so the enhanced covariance relies on both components. The generated realizations preserve essential geological features from the training data, including channel geometries and connectivity patterns that control fluid flow behavior in complex heterogeneous formations [5]. By leveraging these large synthetic ensembles within machine learning-enhanced localization frameworks [27–29], this approach can improve covariance estimation accuracy while maintaining the geological realism necessary for reliable uncertainty quantification in carbon storage applications.

Recent advances have demonstrated the potential of integrating machine learning with data assimilation across various fields. The evolution of DA-ML integration has progressed from early demonstrations on low-dimensional chaotic systems [30] to frameworks that merge DA and ML through unified Bayesian formulations [31]. The designers of these frameworks established that DA can provide complete state reconstructions from sparse observations, enabling ML models to learn dynamics that would otherwise be untrainable. The concept of “Data Learning”—combining ML’s pattern recognition capabilities with DA’s physics-based constraints—has emerged as a powerful paradigm [32], with theoretical foundations showing how many DA-ML approaches can be unified under expectation-maximization frameworks.

In the subsurface domain, researchers explored two primary integration strategies: using ML to enhance DA processes and leveraging DA/UQ methods to improve ML models through physics-informed constraints [33]. For enhancing DA, ML is applied to learn model error corrections [34, 35], infer unresolved scale parametrizations directly from observations [36], and enable training with partial observations through hybrid forecasting approaches [37]. Recent theoretical advances include Deep Data Assimilation [38], where neural networks learn to perform assimilation updates internally, embedding DA knowledge into the model itself. For geological carbon storage specifically, the integration of generative AI with DA represents a paradigm shift. Recent work demonstrates that latent diffusion models can parameterize complex facies-based geomodels while maintaining geological realism during history matching [39], working in latent space through VAE (Variational Auto Encoders) encoders. Score-based diffusion models have enabled real-time ensemble forecasts of 3D CO₂ plume evolution, generating hundreds of realizations in minutes rather than hours [40], although they require extensive training datasets. Beyond diffusion models, researchers employ conditional normalizing flows within Bayesian filtering frameworks to create uncertainty-aware “digital shadows” for continuous monitoring [41]. They have also developed hierarchical DA approaches using sequential Monte Carlo methods to handle uncertainty in both probability distributions and structural assumptions [42]. While these advances demonstrate that generative models can capture non-Gaussian uncertainties and maintain geological consistency, the challenge of adaptive localization that respects complex geological structures in ensemble-based methods remains an active area of research.

To explore how machine learning can enhance data assimilation for GCS, we combine score-based diffusion models with machine learning-enhanced localization to overcome the specific limitations identified above, namely (i) the inability of traditional geostatistical workflows to generate large ensembles at affordable cost in channelized settings, (ii) the sampling noise that dominates covariance estimates at operational ensemble sizes (N_e between 50 and 200), and (iii) the geological inconsistency introduced by distance-only localization in systems where

information propagates along channels rather than radially. Our investigation demonstrates that ML methods can improve data assimilation performance through an approach that operates on three levels:

1. **Efficient super-ensemble generation:** We employ score-based diffusion models trained on geostatistical realizations to enable the generation of thousands of permeability field samples. This overcomes the computational bottleneck of traditional methods by providing large ensembles (5000 members) for statistical estimation without running expensive forward simulations.
2. **Enhanced covariance estimation via ML proxies:** By training fast surrogate models on the working ensemble and applying them to large super-ensembles (5000 members) from the diffusion model, we improve covariance estimation accuracy without additional forward simulations.
3. **Adaptive localization respecting geology:** The ML-enhanced covariances capture how information propagates through specific geological structures, enabling localization that adapts to channel connectivity rather than imposing arbitrary geometric constraints.

We test our methodology on 2D channelized systems with pressure observations. We do not address real-time monitoring, dynamic model updating, or non-channelized geological settings, each worthy of dedicated investigation building on the present integration of score-based diffusion priors with ML-enhanced localization.

2 Theoretical background

This section presents the theoretical foundations of our integrated framework, covering score-based diffusion models, ensemble data assimilation with localization, and machine learning-enhanced covariance estimation.

2.1 Score-Based diffusion models: Mathematical foundation

Score-based diffusion models add noise to data through a forward process and then learn to reverse it to generate new samples [24, 25]. Score-based and latent diffusion models denote different families: score-based models, as used here, operate directly in the original data (permeability) space and learn the Stein score $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ of a noise-perturbed density; latent diffusion models [39] first encode the data into a low-dimensional latent space with an autoencoder and then train a diffusion process on those latents. The two formulations share the denoising-by-score-matching principle but differ in where the diffusion acts and in the associated training cost. These methods evolved from foundational work on deep unsupervised learning using nonequilibrium thermodynamics [43] and denoising diffusion probabilistic models [44], and have demonstrated better performance when compared to GANs (Generative Adversarial Networks) and variational autoencoders for high-quality sample generation [45]. Figure 1 illustrates the complete forward and reverse diffusion process, showing how channelized permeability fields are progressively transformed into noise and then reconstructed. The forward process transforms data samples into random noise through a continuous-time stochastic process defined by a stochastic differential equation (SDE):

$$d\mathbf{x}_t = g(t)d\mathbf{W}_t, \quad t \in [0, T] \quad (1)$$

where the initial condition, $\mathbf{x}_0 \in \mathbb{R}^D$, is a sample from the data distribution representing the permeability field in our context. The diffusion coefficient $g(t)$ controls the noise injection rate, $\mathbf{W}_t$ is the Wiener process (Brownian motion), and $t \in [0, T]$ is the time variable. As time progresses from $t = 0$ to $t = T$, $\mathbf{x}_t$ gradually transitions from a data distribution to a noise distribution.

In this work, we adopt the Score Matching with Langevin Dynamics (SMLD) variant of the Variance Exploding (VE) SDE formulation [25]. In this framework, the forward diffusion is defined as $d\mathbf{x} = \sigma^t d\mathbf{W}$, where $\sigma = 25.0$

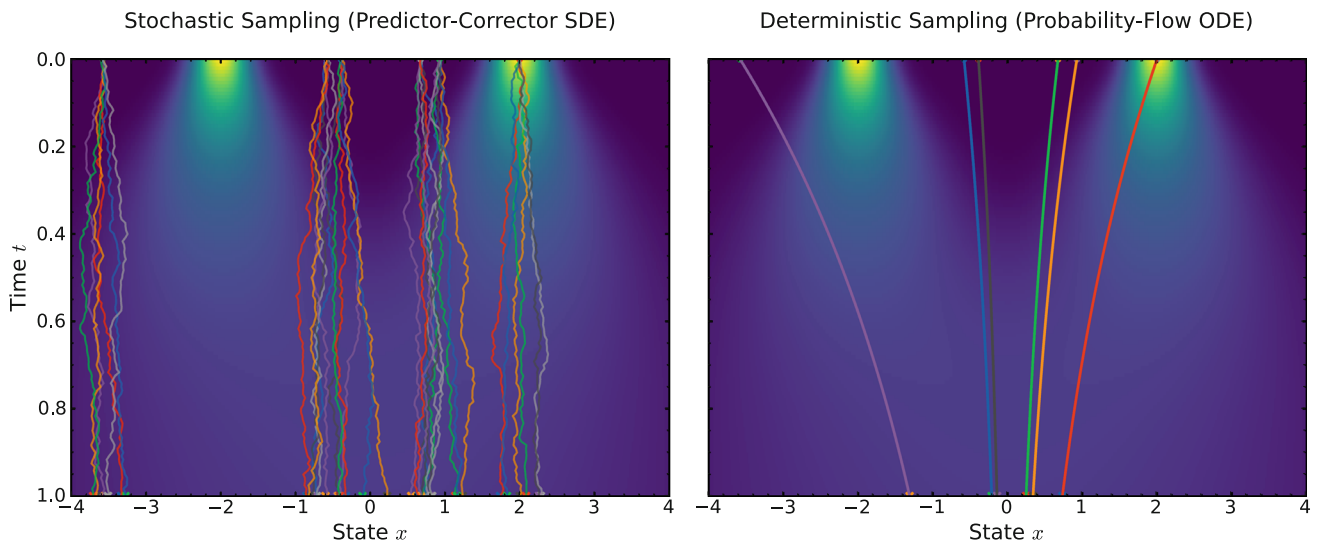


Fig. 1 Score-based diffusion process. Forward diffusion (top) transforms channelized permeability to noise via VE-SDE. Reverse process (bottom) uses learned score function $s_\theta(\mathbf{x}_t, t)$ for denoising with SDE (stochastic) or ODE (deterministic) sampling. In the middle/bottom panels, each coloured line is an individual trajectory of the reverse process: solid blue lines are independent SDE trajectories (reverse-time stochastic) and solid orange lines are ODE trajectories (deterministic probability flow), both starting from the same set of terminal noise samples $\mathbf{x}_T$. Because the SDE retains a Brownian term, two SDE trajectories with identical initial noise explore different regions of the learned density and may land in visibly different modes; the ODE, being deterministic, maps each terminal point to a unique sample. The two samplers share the same marginal distribution at every intermediate time, but not the same pathwise realisations, which is why SDE endpoints can lie far from the corresponding ODE endpoint

is the noise scale parameter. The diffusion coefficient is given by $g(t) = \sigma^t$, and the marginal probability standard deviation is:

$$\sigma_t = \sqrt{\frac{\sigma^{2t} - 1}{2 \ln \sigma}} \quad (2)$$

This formulation ensures that the variance of the noisy data grows exponentially with time, providing a transition from structured channelized patterns to isotropic Gaussian noise. The choice of $\sigma = 25.0$ was empirically determined to provide sufficient noise levels to destroy geological structures while maintaining numerical stability during training. The conditional distribution $p_t(\mathbf{x}_t|\mathbf{x}_0)$ follows a Gaussian distribution:

$$p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma_t^2 \mathbf{I}) \quad (3)$$

The key insight is that the time-reversed SDE can generate samples from the data distribution:

$$d\mathbf{x}_t = -g^2(t) \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) dt + g(t) d\bar{\mathbf{W}}_t, \quad t \in [T, 0] \quad (4)$$

where $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is the score function — the gradient of the log-probability density. By solving this reverse SDE from $t = T$ to $t = 0$ with initial condition $\mathbf{x}_T \sim \mathcal{N}(0, \sigma_T^2 \mathbf{I})$, we obtain samples from the data distribution.

The score function is approximated using a neural network $s_\theta(\mathbf{x}_t, t) \approx \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$, trained using denoising score matching [46]:

$$\mathcal{L}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0, T)} \mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}} \mathbb{E}_{\mathbf{x}_t \sim p_t(\mathbf{x}_t|\mathbf{x}_0)} \|s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{x}_0)\|^2 \quad (5)$$

The trained diffusion model can also be adapted for posterior sampling to condition generated realizations on direct permeability observations. This approach follows recent advances in diffusion posterior sampling for general noisy inverse problems [47], which demonstrated how diffusion models can be adapted for Bayesian inference tasks.

The reverse diffusion process in Eq. 4 can be solved using two distinct approaches: stochastic (SDE) or deterministic (ODE) sampling. The probability flow ODE, which by construction induces, at every time t , the same single-time marginal density $p_t(\mathbf{x}_t)$ as the forward/reverse SDE (while individual ODE and SDE trajectories are different), is given by:

$$\frac{d\mathbf{x}_t}{dt} = -\frac{1}{2}g^2(t)\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t), \quad t \in [T, 0] \quad (6)$$

We tested both SDE and ODE formulations for ensemble generation in data assimilation, and both approaches yield good results for generating geologically realistic channelized permeability fields. The choice between methods depends on the specific application requirements: SDE sampling provides more stochastic variability and exploration of the learned distribution, while ODE sampling offers greater control and reproducibility. For our application in ensemble generation for data assimilation, we adopt the ODE formulation for several practical advantages. The deterministic nature ensures reproducibility. Given its deterministic nature and a fixed starting noise, the ODE always produces the same realization, enabling controlled experiments and sensitivity analyses. The ODE formulation typically requires fewer discretization steps than the SDE for comparable quality, improving computational efficiency. The smooth trajectories of the ODE solver also facilitate better preservation of geological structures during the denoising process. However, both the reverse-time SDE and the probability-flow ODE can generate large ensembles of geologically realistic samples suitable for enhanced covariance estimation in our setting, since they both are able to generate realistic samples.

2.2 Ensemble-Based data assimilation and localization

The Ensemble Smoother with Multiple Data Assimilation (ESMDA) [15] provides a framework for integrating dynamic observations into reservoir models. Extending fundamental ensemble Kalman filter theory [12, 14] as described in the unified data assimilation framework of Evensen et al. [10], the method performs multiple assimilation steps with *inflated* observation-error covariances, thereby linearizing the nonlinear inverse problem through iteration. Here, inflation means that the observation-error covariance matrix $\mathbf{C}_D$ is multiplied by a scalar factor $\alpha_i > 1$ at each step, and the corresponding random perturbations of the data are scaled by $\sqrt{\alpha_i}$. This distributes a single Bayesian update over several smaller ones, which improves linearization and reduces sampling noise.

For an ensemble of model parameters $\{\mathbf{z}_j\}_{j=1}^{N_e}$, where $\mathbf{z} \in \mathbb{R}^{N_z}$ represents the log-permeability field, the ESMDA update equation at iteration i is:

$$\mathbf{z}_j^a = \mathbf{z}_j^f + \mathbf{C}_{zD}^f \left(\mathbf{C}_{DD}^f + \alpha_i \mathbf{C}_D \right)^{-1} \left(\mathbf{d}_{\text{obs}} + \sqrt{\alpha_i} \boldsymbol{\epsilon}_j - \mathbf{d}_j^f \right) \quad (7)$$

where superscripts f and a denote forecast (prior) and analysis (posterior) quantities respectively, $\mathbf{C}_{zD}^f$ is the cross-covariance between parameters and predicted data, $\mathbf{C}_{DD}^f$ is the auto-covariance of predicted data, $\mathbf{C}_D$ is the observation error covariance matrix, α_i is the inflation coefficient for iteration i , $\mathbf{d}_{\text{obs}}$ represents the observed data, $\mathbf{d}_j^f = \mathbf{G}(\mathbf{z}_j^f)$ is the predicted data from forward model $\mathbf{G}$, and $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \mathbf{C}_D)$ are random perturbations.

However, for practical ensemble sizes (typically $N_e \sim 100$), the sample covariance matrices contain large errors, leading to spurious correlations and subsequent ensemble collapse [26]. The industry standard solution to deal with these spurious correlations is distance-based localization, which modifies the Kalman gain through element-wise multiplication with a localization matrix $\mathbf{L}$. This matrix typically uses correlation functions such as the fifth-order piecewise rational function of Gaspari and Cohn [20], which tapers correlations to zero beyond a specified distance, in other words, it masks part of the error covariance matrix. The GC cutoff radius was kept

fixed across all ensemble sizes in the experiments reported in Section 4, so that differences in posterior behaviour are attributable to the localization scheme and not to retuning of the cutoff.

$$\mathbf{z}_j^a = \mathbf{z}_j^f + \mathbf{L} \circ \left[\mathbf{C}_{zD}^f \left(\mathbf{C}_{DD}^f + \alpha_i \mathbf{C}_D \right)^{-1} \right] \left(\mathbf{d}_{\text{obs}} + \sqrt{\alpha_i} \boldsymbol{\epsilon}_j - \mathbf{d}_j^f \right) \quad (8)$$

where $\circ$ denotes the Schur (element-wise) product.

Pseudo-optimal localization (PO), as we use it in the present study, denotes a taper derived by minimizing the expected Frobenius-norm error between the true and localized covariance matrices, following the formulation of [21]. This yields data-driven coefficients that balance the removal of spurious correlations against the preservation of true correlations. Operationally, the PO taper for entry (i, j) of the Kalman-update matrix has the form $L_{ij} = c_{ij}^2 / (c_{ij}^2 + (c_{ii}^2 + c_{jj}^2) / N_e)$, where c_{ij} is a plug-in estimate of the (i, j) cross-covariance entry. The only difference between the baseline PO used in Section 4 and the ML variants in Eq. 14 is the source of that plug-in estimate: for the baseline PO the c_{ij} come from the working-ensemble sample covariance, whereas for ML-linear, ML-RF and ML-XGBoost they come from the super-ensemble covariance $\tilde{\mathbf{C}}_{zD}^{\text{ML}}$ evaluated with the corresponding ML proxy. All ML variants therefore use Eq. 14 with the same functional form; they differ only in the proxy f^{ML} feeding the super-ensemble.

2.3 Machine learning-enhanced localization

Distance-based localization methods have become the industry standard for ensemble-based data assimilation, with implementations ranging from simple cutoff functions, like Gaspari and Cohn [20], to adaptive correlation-based localization models [48] and Schur product-based approaches [26]. These methods have been successfully deployed in operational weather forecasting, oceanography, and petroleum reservoir history matching for over two decades. However, while effective for many applications, they fail to capture the complex, non-local relationships between channelized permeability structures and pressure observations. In channelized systems, where pressure information travels along high-permeability pathways over long distances, distance-only tapering can be misleading, while correlation-based schemes can follow the geology, provided the estimated correlations are reliable. The challenge is that accurate correlation estimates typically require large ensembles, which are rarely affordable in practice. To address this, we use machine learning to augment classical localization: a fast proxy is trained on the working ensemble and then applied to a large super-ensemble generated by the diffusion model to produce more stable, correlation-aware covariance estimates even when the operational ensemble N_e is small [27–29]. We adopt the formulation of Silva et al. [27, 28] without modification; the only deviations are (a) the prior ensemble, which here is generated with the score-based diffusion model rather than FLUVSIM directly, and (b) the fact that the ML proxy f^{ML} is retrained at every ESMDA iteration so that the super-ensemble predictions track the current state of the working ensemble. Implementation details (model classes, 80/20 split, default scikit-learn / XGBoost hyperparameters) are repeated in Section 3 for completeness and are identical to those in Silva et al. [27].

Our ML-enhanced localization framework extends the approach proposed by Silva et al. [27, 28] by training a fast proxy model $f^{\text{ML}} : \mathbf{z} \rightarrow \mathbf{d}$ using the working ensemble, then applying this proxy to a large super-ensemble, of around five thousand members, generated by the diffusion model. The enhanced covariance estimate becomes:

$$\tilde{\mathbf{C}}_{zD}^{\text{ML}} = \frac{1}{N_s - 1} \sum_{k=1}^{N_s} (\mathbf{z}_k^s - \bar{\mathbf{z}}^s) (f^{\text{ML}}(\mathbf{z}_k^s) - \bar{\mathbf{d}}^{\text{ML}})^T \quad (9)$$

where $N_s \gg N_e$ is the super-ensemble size, and $\mathbf{z}_k^s$ are samples from the diffusion model. This approach increases the sample size for covariance estimation while maintaining computational feasibility.

Localization is effective when the taper $\mathbf{L}$ keeps parameter–data entries that represent true physical influence and zeros those that are spurious. With small ensembles, the sample cross-covariance $\mathbf{C}_{zD}$ is noisy: Monte Carlo error

scales like $O(1/\sqrt{N_e})$, so many weak-but-real correlations are drowned out by sampling noise. Our approach trains a fast proxy $f^{ML} : \mathbf{z} \mapsto \mathbf{d}$ on the working ensemble and then evaluates it on a large super-ensemble ($N_s \gg N_e$) drawn from the diffusion model.

The super-ensemble proxy covariance $\tilde{\mathbf{C}}_{zD}^{ML}$ (Eq. 9) does *not* enter the ESM DA update equation in Eq. 11 directly; instead, it is used exclusively to build the ML localization mask $\mathbf{L}^{ML}$ through Eq. 14. The update then applies the mask to the Kalman-gain factor $\mathbf{C}_{zD}^f (\mathbf{C}_{DD}^f + \alpha_i \mathbf{C}_D)^{-1}$, in which $\mathbf{C}_{zD}^f$ and $\mathbf{C}_{DD}^f$ are always the simulator-based sample covariances of the working ensemble of size N_e (see Algorithm 3, lines 4 and 14). In short, the proxy covariance shapes *where* the update acts through $\mathbf{L}^{ML}$, while the magnitude of the update is driven entirely by simulator covariances. This separation is the reason the method tolerates imperfect proxies.

The diffusion-based prior generation follows Di Federico and Durlinsky [39] in spirit, but we train in pixel space rather than in a VAE latent space and we use variance-exploding SMLD rather than their latent DDPM formulation; quantities reused unchanged (ODE sampler with 1000 midpoint steps, ConvNeXt UNet with Linformer attention) are flagged as such in Section 3. The ML-localization component is identical to Silva et al. [27, 28]. The only genuinely new element of this paper is the combination of the two, together with the diagnostic of variance preservation for channelized CO₂-storage ensembles.

The original consistent-localization formulation of Houtekamer and Mitchell [26] applies the taper separately to $\mathbf{C}_{zD}$ and $\mathbf{C}_{DD}$, whereas Eq. 11 applies the Schur product to the Kalman-gain matrix as a whole. For the low-dimensional observation space used here ($\mathbf{d} \in \mathbb{R}^{96}$) the two formulations are numerically very close, and the matrix-level form is retained for simplicity.

3 Methods and implementation

This section presents the technical details of our integrated framework, including the channelized reservoir model, diffusion model implementation, ESM DA configuration, and machine learning-enhanced localization approach.

3.1 Channelized reservoir generation and CO₂ injection modeling

The permeability realizations that feed the diffusion model are produced with FLUVSIM [8], an object-based algorithm that places sinuous channel geometries into the background. Sequential indicator simulation (SIS) is not used to generate the facies template; we only refer to it later in the paper to contrast the bimodal, channel/non-channel facies representation obtained from the object-based realizations with what would be obtained by a two-point SIS workflow. Permeability values inside channels (2000 mD) and background (50 mD) are assigned deterministically by facies, and no further geostatistical simulation is carried out on top of the FLUVSIM facies.

We employ a 2D channelized reservoir model representative of fluvial depositional environments commonly encountered in CO₂ storage formations. The model domain consists of a 64×64 grid with 5-meter cell dimensions, resulting in a 320×320 meter area. Channelized permeability fields are generated using the FLUVSIM algorithm [8], which simulates fluvial channel systems through object-based modeling.

The FLUVSIM parameters are sampled from normal distributions $\mathcal{N}(\mu, \sigma)$ to ensure geological diversity across the training dataset. Table 1 summarizes the key parameters with their mean (μ) and standard deviation (σ) values.

This stochastic parameterization generates diverse channelized patterns while maintaining geological plausibility. The resulting permeability fields exhibit bimodal distributions with high-permeability channels (2000 mD) embedded in low-permeability background facies (50 mD), creating preferential flow paths characteristic of fluvial deposits. Figure 2 shows representative examples of the channelized permeability fields generated using FLUVSIM.

The CO₂ injection and storage simulations are performed using the Delft Advanced Research Terra Simulator (DARTS), a compositional flow simulator specifically developed for subsurface flow applications [49]. DARTS uses operator-based linearization for efficient solution of complex multiphase flow problems [50], and has been

Table 1 FLUVSIM parameters for channelized reservoir generation

Parameter	Mean $\pm$ Std. Dev.	Description
Channel orientation (degrees)	90 ± 30	Primary flow direction
Channel amplitude (m)	250 ± 10	Variable sinuosity
Channel wavelength (m)	2000 ± 50	Meandering patterns
Width-to-thickness ratio	50 ± 5	Channel aspect ratio
Channel proportion	0.40 ± 0.05	Net-to-gross ratio
Undulation wavelength (m)	250 ± 10	Thickness/width variations

validated against benchmark cases including the FluidFlow experiment [51]. The simulator has been successfully applied to optimize CO₂ injection using multi-scale reconstruction [52] and to model enhanced CO₂ dissolution trapping in saline aquifers [53], demonstrating its capabilities for complex CO₂ storage scenarios. The reservoir model employs a structured Cartesian grid with dimensions of $64 \times 64 \times 1$ cells, representing a two-dimensional horizontal layer. Each grid cell measures 5 meters in length and width, with a thickness of 10 meters, resulting in a computational domain of 320×320 meters horizontally. The simulated reservoir layer is situated at a depth of 2,000 meters, reflecting typical geological conditions for CO₂ storage, with an initial pressure of 150 bar and temperature of 50°C.

The reservoir simulations involve a two-component, two-phase system comprising carbon dioxide (CO₂) and water (brine with 100,000 ppm salinity). The permeability field is generated using FLUVSIM as previously described, exhibiting bimodal distribution with high-permeability channels (2000 mD) embedded in low-permeability background (50 mD). The porosity remains constant throughout the reservoir at 0.3.

CO₂ is injected through a single vertical well located centrally within the grid at a consistent rate of 1000 m³/year over a two-year period. The injected fluid consists of nearly pure CO₂ into the saline aquifer. Pressure measurements are recorded monthly at four monitoring wells positioned in a five-spot pattern, each located at equal distances from the injection point in the cardinal directions. These observations provide spatial data points for subsequent data assimilation, with measurement errors of 1% reflecting realistic field monitoring uncertainties.

The relative permeability in the porous medium is modeled using the Brooks-Corey formulation [54], with exponents $n_w = 2.0$ for water and $n_g = 1.5$ for CO₂. Residual saturations are set to $S_{wc} = 0.25$ and $S_{gc} = 0.1$, with end-point relative permeability values of $k_{rw} = k_{rg} = 1.0$. Capillary pressure is neglected due to its limited influence within the two-dimensional modeling framework. Diffusion and hysteresis effects, while potentially influential on CO₂ migration and trapping, are excluded to maintain consistency with the study's focus on large-scale reservoir dynamics. This modeling approach is considered sufficient to capture the key multiphase flow mechanisms governing CO₂ migration and entrapment in porous media.

DARTS uses operator-based linearization for efficient solution of the coupled multiphase flow and transport equations, accurately capturing the complex physics of CO₂ injection including phase behavior and gravitational segregation. The simulator's computational efficiency enables the thousands of forward model evaluations required for ensemble-based data assimilation.

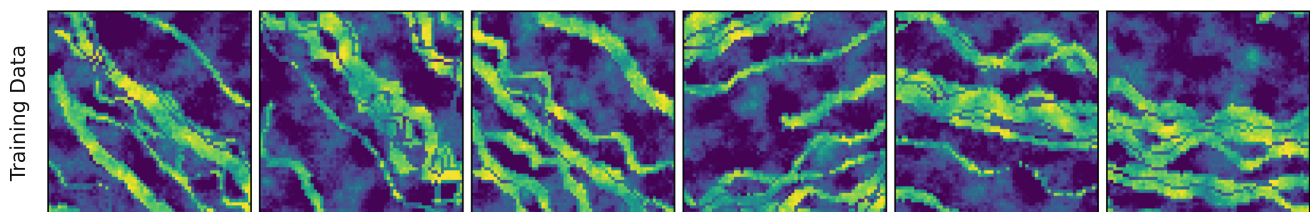


Fig. 2 Representative FLUVSIM-generated channelized permeability fields with bimodal distribution: channels (2000 mD) in background (50 mD)

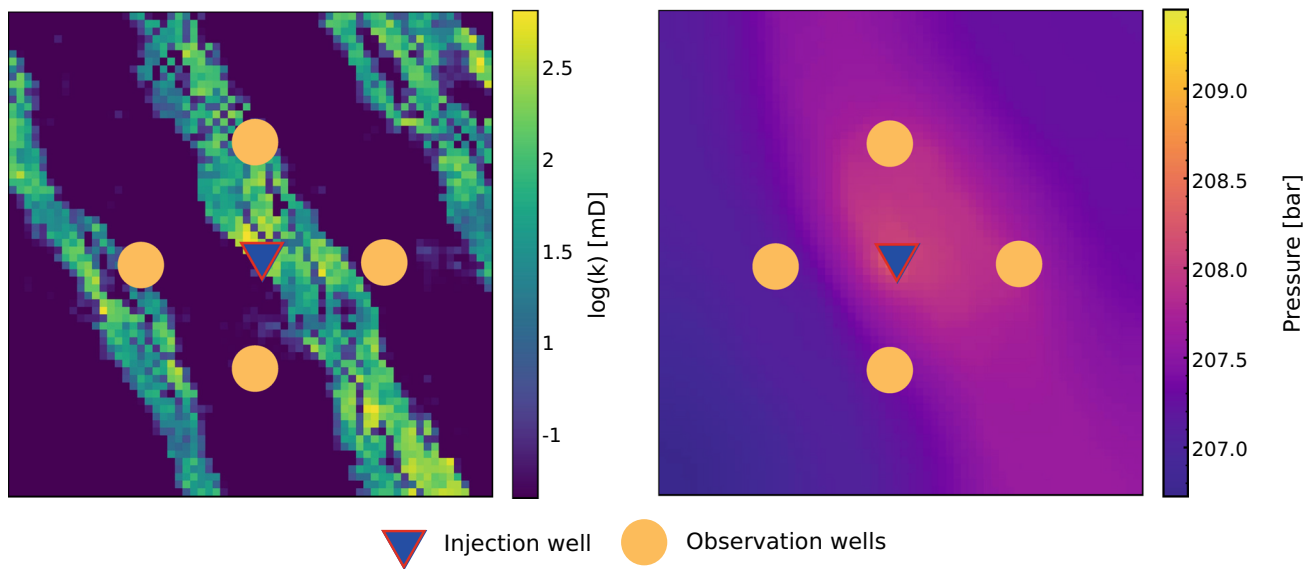


Fig. 3 Channelized reservoir model (left) and DARTS-simulated pressure field (right) with injection well (triangle) and monitoring wells (circles). The pressure distribution correlates with the high-permeability channels and the injector–producer configuration: it is strongest along the dominant connected channel, with weaker propagation along the two minor channels

Figure 3 illustrates the channelized reservoir model and the corresponding pressure field simulated with DARTS. The left panel shows the log-permeability field with high-permeability channels, while the right panel displays the pressure distribution after injection. The injection well (triangle) and four monitoring wells (circles) are indicated. The pressure field clearly follows the channel structure, with higher pressures propagating preferentially along the high-permeability pathways.

3.2 Diffusion model implementation

The score-based diffusion model is implemented to learn the distribution of channelized permeability fields and generate new realizations that maintain geological realism.

3.2.1 Neural network architecture

We employ a hybrid UNet-Transformer [55] architecture that combines convolutional neural networks with efficient attention mechanisms for learning the score function of channelized permeability distributions. The architecture is detailed in Fig. 4. We use a U-Net with attention blocks to capture multi-scale features in permeability fields. The network begins with a time embedding module that uses Gaussian Fourier projections to encode the continuous diffusion time into high-dimensional representations. This temporal signal allows the network to adapt its predictions to the current noise level during the diffusion process.

The encoder path consists of four blocks that progressively downsample the input while expanding the feature channels. Each block uses ConvNeXt-style residual blocks [56]: a 7×7 convolution to capture broader spatial context, followed by 3×3 convolutions and a residual connection; depthwise separable convolutions are not used unless explicitly stated. The time embeddings are injected at each resolution level, enabling the network to modulate its feature extraction based on the diffusion timestep.

At the bottleneck and intermediate resolutions, we incorporate Spatial Linformer attention modules [57]. Unlike standard self-attention which scales quadratically with input size, Linformer achieves linear complexity through low-rank approximations. This builds on the transformer architecture's success in capturing long-range dependencies [58] while addressing computational constraints. This is valuable for geological applications where

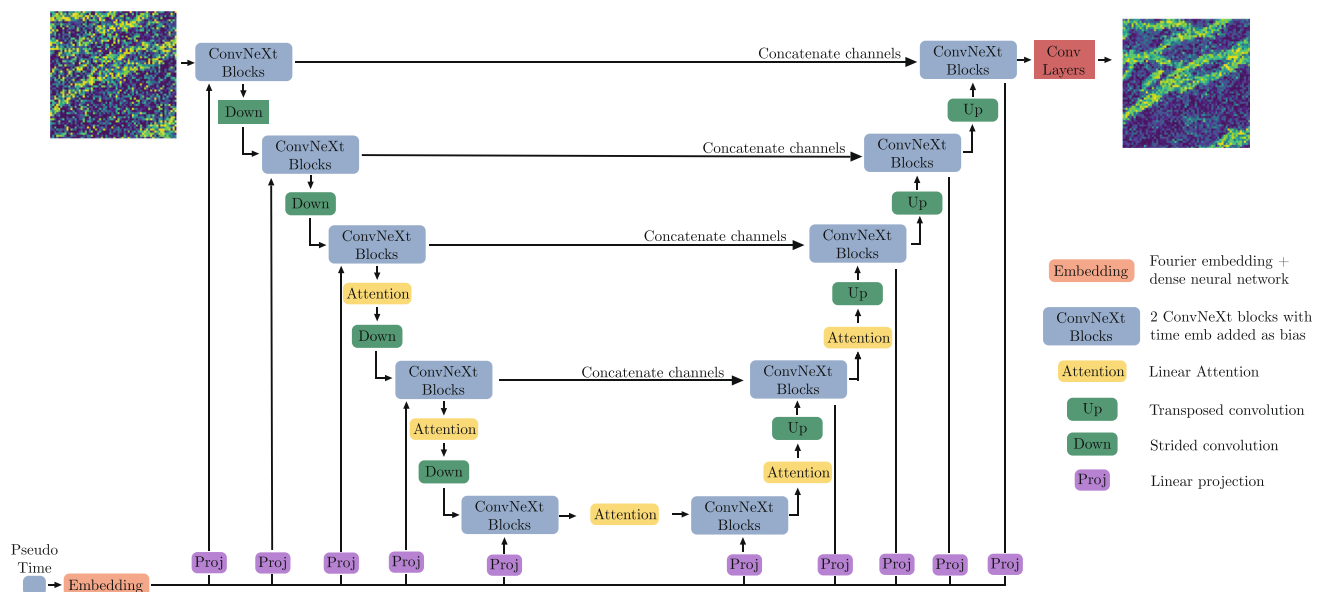


Fig. 4 Visualization of UNet architecture with attention

maintaining long-range spatial dependencies—such as channel connectivity across the domain—is important for realistic permeability field generation.

The decoder path mirrors the encoder with skip connections that preserve fine-scale details during upsampling. Through transposed convolutions and ConvNextBlock refinements, the network progressively reconstructs the score function at the original resolution. The final output represents $s_\theta(\mathbf{x}_t, t)$, the estimated score function that guides the reverse diffusion process.

3.2.2 Training dataset and procedure

The training dataset comprises 3,242 channelized permeability realizations generated using sequential indicator simulation with systematically varied geostatistical parameters. This ensures a comprehensive coverage of plausible geological patterns, with each 64×64 field representing a different realization of channelized formations with varying orientations, widths, and connectivity patterns.

The model training employs standard practices for deep generative models. We use the Adam optimizer with a cosine annealing learning rate schedule that gradually reduces from 10^{-4} to 10^{-6} , promoting stable convergence. The training process minimizes the denoising score matching objective shown in Algorithm 1, which teaches the network to predict the gradient of the log-density at various noise levels.

The model converged in 525 epochs. The final loss of 308.76 represents an 83.6% reduction from the initial value of 1,882.96. The quick convergence shows that channelized patterns are well-structured and the UNet-Transformer architecture captures these geological features effectively.

3.2.3 Sampling strategy

The trained diffusion model supports two complementary sampling approaches for generating permeability fields. The Predictor-Corrector (PC) sampler combines Euler-Maruyama steps for the reverse SDE with Langevin MCMC corrections, producing high-quality samples with stochastic diversity. This approach excels in exploring the full distribution of channelized patterns learned during training.

Algorithm 1 Training score-based diffusion models.

Input : Dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$, neural network s_θ , noise schedule $\sigma(t)$
Output : Trained score network s_θ

```

1 for  $epoch = 1$  to  $N_{epochs}$  do
2   for  $batch \mathcal{B} \subset \mathcal{D}$  do
3     Sample  $t \sim \mathcal{U}(0, T)$  uniformly;
4     Sample  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ ;
5     Compute  $\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon$ ;
6     Compute loss:  $\mathcal{L} = \|\epsilon + \sigma_t s_\theta(\mathbf{x}_t, t)\|^2$ ;
7     Update  $\theta$  using gradient descent on  $\mathcal{L}$ ;
8   end for
9 end for
10 return trained  $s_\theta$ ;

```

For our ESM DA application, however, we primarily employ the probability flow ODE formulation:

$$\frac{d\mathbf{x}}{dt} = -\frac{1}{2}g^2(t)s_\theta(\mathbf{x}_t, t) \quad (10)$$

Deterministic sampling helps, because we can reproduce results exactly for ensemble-based data assimilation. Most importantly, it ensures reproducibility, which is essential for scientific validation and debugging. Given the same initial noise vector, the ODE always produces identical realizations. This property is used for debugging, validation, and ensuring consistent results across different computational environments. The ODE solver employs the midpoint method with 1000 integration steps, balancing numerical accuracy with computational efficiency.

Generated samples are not modified beyond reversing the preprocessing used for training. Specifically, inputs were standardized (per-pixel z-scoring of $\log k$) to ease optimization; at generation time we denormalize the network outputs using the training statistics and invert the log transform to recover permeability in millidarcies. We then compute and report simple diagnostics—channel proportion (0.40 ± 0.05) and value-range checks ($k \in [10, 10^4]$ mD)—but we do not filter or discard samples based on these checks.

Figure 5 illustrates the generation process, showing how both sampling methods progressively transform Gaussian noise into realistic channelized permeability fields while preserving distinct geological features.

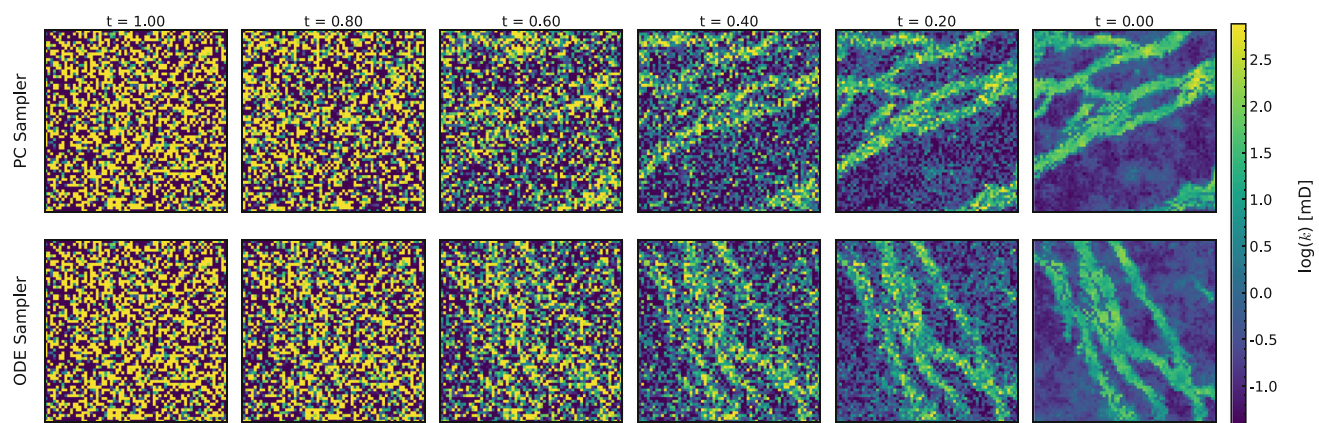


Fig. 5 Evolution of channelized structures during generation. ODE (top) and PC (bottom) samplers progressively denoise from Gaussian to realistic permeability fields

3.3 ESM DA configuration

The diffusion model was trained for 525 epochs on 3,242 channelized realizations (Section 3) and thereafter generates samples via the 1000-step midpoint ODE sampler. A single DARTS simulation takes of the order of a few minutes of wall-clock time per realization on one modern CPU core, so N_e forward simulations dominate every ESM DA iteration regardless of ensemble size. Training the ML proxy at each iteration adds the times reported in Table 2 (fractions of a second for linear regression up to approximately 107 s for Random Forest at $N_e = 1000$), which is small relative to the N_e simulator runs needed in the same iteration.

The data assimilation employs ESM DA with four iterations ($N_a = 4$), using uniform inflation coefficients of $\alpha_i = 4$ for all iterations. This choice balances computational efficiency with adequate linearization of the nonlinear inverse problem. The observation error is set to 1% of the true values, consistent with typical pressure measurement uncertainties. To maintain physical plausibility, parameter bounds constrain the log-permeability values to the range $\log_{10}(k) \in [1, 4]$, corresponding to permeability values between 10 and 10,000 mD. We evaluate the framework performance across multiple ensemble sizes (50, 100, 200, 500, and 1000 members) to understand scaling behavior and practical requirements for operational deployment.

With the inflation schedule $\{\alpha_i\}$ and observation-error covariance $\mathbf{C}_D$ defined above, the localized ESM DA analysis step at iteration i reads:

$$\mathbf{z}_j^a = \mathbf{z}_j^f + \mathbf{L} \circ \left[\mathbf{C}_{zD}^f \left(\mathbf{C}_{DD}^f + \alpha_i \mathbf{C}_D \right)^{-1} \left(\mathbf{d}_j - \mathbf{G}(\mathbf{z}_j^f) \right) \right] \quad (11)$$

where $\circ$ denotes the Schur (element-wise) product, $\mathbf{L}$ is the localization matrix, and $\mathbf{d}_j = \mathbf{d}_{\text{obs}} + \sqrt{\alpha_i} \boldsymbol{\epsilon}_j$ with $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \mathbf{C}_D)$.

3.4 Machine learning-enhanced localization

Our ML localization works by training a fast proxy model on the working ensemble to improve covariance estimation. Adapting the machine learning-based distance-free localization introduced in recent studies [27, 28], we tailor these methods for channelized reservoir characterization. Given ensemble pairs $\{(\mathbf{z}_j, \mathbf{d}_j)\}_{j=1}^{N_e}$ from DARTS simulations, we train:

$$f^{ML} : \mathbf{z} \rightarrow \mathbf{d} \quad (12)$$

where $\mathbf{z} \in \mathbb{R}^{4096}$ is the flattened log-permeability field and $\mathbf{d} \in \mathbb{R}^{96}$ contains pressure observations.

Table 2 Machine Learning Model Performance for Localization Coefficient Prediction

Model	Number of Ensemble Members (N_e)				
	50	100	200	500	1000
(a) Root Mean Square Error (RMSE)					
Linear Regression	2.351	2.828	0.681	1.262	0.268
Random Forest	0.254	0.175	0.191	0.177	0.159
XGBoost	0.310	0.198	0.216	0.188	0.159
(b) Training Time (seconds)					
Linear Regression	0.016	0.027	0.067	0.141	0.40
Random Forest	3.528	6.238	15.678	38.890	106.8
XGBoost	1.301	1.849	4.372	15.767	70.9

3.4.1 Enhanced covariance estimation

The enhanced covariance estimation uses a super-ensemble of $N_s = 5000$ realizations:

$$\tilde{\mathbf{C}}_{zD}^{ML} = \frac{1}{N_s - 1} (\mathbf{Z}_s - \bar{\mathbf{Z}}_s) (\mathbf{D}_s^{ML} - \bar{\mathbf{D}}_s^{ML})^T \quad (13)$$

where $\mathbf{D}_s^{ML} = [f^{ML}(\mathbf{z}_1^s), \dots, f^{ML}(\mathbf{z}_{N_s}^s)]$ contains ML predictions.

For the ML-enhanced localization, we compute localization coefficients using a modified pseudo-optimal approach. When model parameters lack direct spatial relationships with observations, conventional distance-based schemes become inadequate. We therefore employ an adaptation of the pseudo-optimal taper function introduced by Rukhin [21], selected for its rigorous theoretical foundation and computational efficiency.

Following Rukhin [21], we employ a pseudo-optimal Schur taper that minimizes the expected Frobenius-norm (matrix Euclidean norm) error between the true and the localized covariance. We adapt the weights to the parameter–data cross-covariance $(\mathbf{z}, \mathbf{d})$ and replace unknown population quantities with ML-enhanced plug-in estimates computed from the super-ensemble. For entry (i, j) of the localization matrix we set

$$\mathbf{L}_{ij}^{ML} = \frac{(\tilde{c}_{ij}^{ML})^2}{(\tilde{c}_{ij}^{ML})^2 + \frac{(\tilde{c}_{ij}^{ML})^2 + \tilde{c}_{ii}^{ML} \tilde{c}_{jj}^{ML}}{N_e}}, \quad (14)$$

where $\tilde{c}_{ij}^{ML}$ is the ML-enhanced covariance between parameter i and datum j , and N_e is the ensemble size used in the ESMDA update. To avoid retaining near-zero, noise-dominated terms, we apply a practical thresholding rule: set $\mathbf{L}_{ij}^{ML} = 0$ whenever $|\tilde{c}_{ij}^{ML}| < \eta \sqrt{\tilde{c}_{ii}^{ML} \tilde{c}_{jj}^{ML}}$ with $\eta = 10^{-3}$. This plug-in adaptation yields an effective suppression of spurious correlations in high-dimensional, small-ensemble settings.

Algorithm 2 ML-Enhanced localization.

Input : Ensemble $\{\mathbf{z}_j, \mathbf{d}_j\}_{j=1}^{N_e}$, ML model type, super-ensemble size N_s
Output : Localization matrix $\mathbf{L}^{ML}$

```

1 // Step 1: Train ML proxy
2 Train  $f^{ML}$  using  $\{(\mathbf{z}_j, \mathbf{d}_j)\}_{j=1}^{N_e}$  with 80/20 train-validation split;
3 // Step 2: Generate super-ensemble
4 for  $k = 1$  to  $N_s$  do
5   | Sample  $\mathbf{z}_k^s$  from prior (diffusion model);
6   | Predict  $\mathbf{d}_k^{ML} = f^{ML}(\mathbf{z}_k^s)$ ;
7 end for
8 // Step 3: Compute enhanced covariance
9 Calculate  $\tilde{\mathbf{C}}_{zD}^{ML}$  using Eq. 13;
10 // Step 4: Calculate localization
11 for each parameter  $i$  and observation  $j$  do
12   | Compute  $\mathbf{L}_{ij}^{ML}$  using Eq. 14;
13 end for
14 return  $\mathbf{L}^{ML}$ ;
```

3.4.2 Machine learning models for state estimation

We capture the relationship between permeability fields and pressure responses in reservoir simulation in a tabular dataset, where each grid cell's permeability serves as a feature and well pressures constitute the targets. This structure, combined with recent evidence that classical machine learning methods can outperform deep learning on tabular data [59–61], guides our model selection.

We use three different ML models to approximate the nonlinear forward model. Linear regression provides a computationally efficient baseline, establishing whether nonlinearity is essential for accurate predictions. Despite its simplicity, linear models struggle with the complex multiphase flow physics inherent in CO₂ injection scenarios.

Random forest models [62] address these limitations through ensemble decision trees trained on bootstrap samples. The method naturally handles the sparse influence patterns characteristic of reservoir flow, where distant grid cells often have negligible effect on local pressure observations. By averaging predictions across multiple trees and using random feature selection at each split, random forests capture nonlinear relationships while maintaining robustness against overfitting. XGBoost [63] extends the ensemble approach through gradient boosting, where each successive tree corrects errors from previous iterations.

The choice of tree-based methods for our application comes from two data characteristics identified by Grinsztajn et al. [61]. First, many grid cells represent uninformative features for specific pressure observations due to flow barriers and distance effects. Second, pressure responses can exhibit non-smooth behavior with sharp transitions at permeability contrasts and channel boundaries.

These properties explain the performance gap observed in our experiments, where tree-based methods achieve prediction errors 10–15 times lower than linear regression while preserving geological patterns in the localization coefficients. All models employ default hyperparameters from their respective implementations (scikit-learn [64] and XGBoost library [63]), reflecting operational constraints and ensuring reproducibility.

3.5 Integrated workflow

The workflow consists of two phases: (1) initial ensemble generation using geostatistical models, and (2) iterative ESM DA with ML-enhanced localization. Algorithm 3 details the entire process.

The framework's modular design allows flexibility in component selection. Different ML models can be employed based on problem characteristics—Random Forest for capturing local effects, XGBoost for handling complex nonlinearities, or Linear Regression when computational resources are extremely limited. Similarly, the diffusion model architecture can be adapted to different geological environments by training on appropriate datasets, be they from channelized, fractured, or layered systems (Fig. 6).

4 Results and analysis

This section presents the evaluation of our integrated framework for uncertainty quantification in geological carbon storage. We assess the diffusion model's performance, analyze machine learning proxy models for localization coefficient prediction, and evaluate the data assimilation results across different ensemble sizes. The focus is on how well the framework captures channelized permeability structures and improves pressure predictions through ML-enhanced localization. To quantify variance preservation, we report the normalized variance (NV), defined as the ratio between the posterior and prior ensemble variances of log permeability: $NV = \text{tr}(\mathbf{C}_{\text{post}})/\text{tr}(\mathbf{C}_{\text{prior}})$, where $\mathbf{C}_{\text{post}}$ and $\mathbf{C}_{\text{prior}}$ are the ensemble covariance matrices over all grid cells. An NV of 1 indicates that there is no variance loss, while smaller values indicate that there is a reduced variance. Throughout this work the RMSE

Algorithm 3 Integrated diffusion-ML-ESMDA workflow.

Input : Ensemble size N_e , observations $\mathbf{d}_{\text{obs}}$, trained diffusion model, ML model type

Output : Posterior ensemble $\{\mathbf{z}_j^a\}_{j=1}^{N_e}$

```

1 // Phase 1: Initial Ensemble Generation
2 Generate initial ensemble  $\{\mathbf{z}_j^0\}_{j=1}^{N_e}$  using geostatistical models (e.g., FLUVSIM);
3 // Phase 2: ESMDA with ML-Enhanced Localization
4 for  $i = 1$  to  $N_a$  iterations do
5   // Forward simulations
6   for  $j = 1$  to  $N_e$  do
7     Run DARTS:  $\mathbf{d}_j = \mathbf{G}(\mathbf{z}_j^{i-1})$ ;
8   end for
9   // ML proxy training
10  Train ML model:  $f^{ML} = \text{fit}(\{\mathbf{z}_j^{i-1}\}, \{\mathbf{d}_j\})$ ;
11  // Super-ensemble generation
12  Generate  $\{\mathbf{z}_k\}_{k=1}^{N_s}$  using diffusion model;
13  Predict:  $\mathbf{d}_k^{ML} = f^{ML}(\mathbf{z}_k)$  for  $k = 1, \dots, N_s$ ;
14  // Enhanced localization
15  Compute  $\tilde{\mathbf{C}}_{zD}^{ML}$  using super-ensemble;
16  Calculate localization matrix  $\mathbf{L}^{ML}$  (Algorithm 2);
17  // ESMDA update
18  for  $j = 1$  to  $N_e$  do
19    Perturb observations:  $\mathbf{d}_j^{\text{pert}} = \mathbf{d}_{\text{obs}} + \sqrt{\alpha_i} \boldsymbol{\epsilon}_j$ ;
20    Apply localized update to get  $\mathbf{z}_j^i$ ;
21  end for
22 end for
23 return Posterior ensemble  $\{\mathbf{z}_j^{N_a}\}_{j=1}^{N_e}$ ;

```

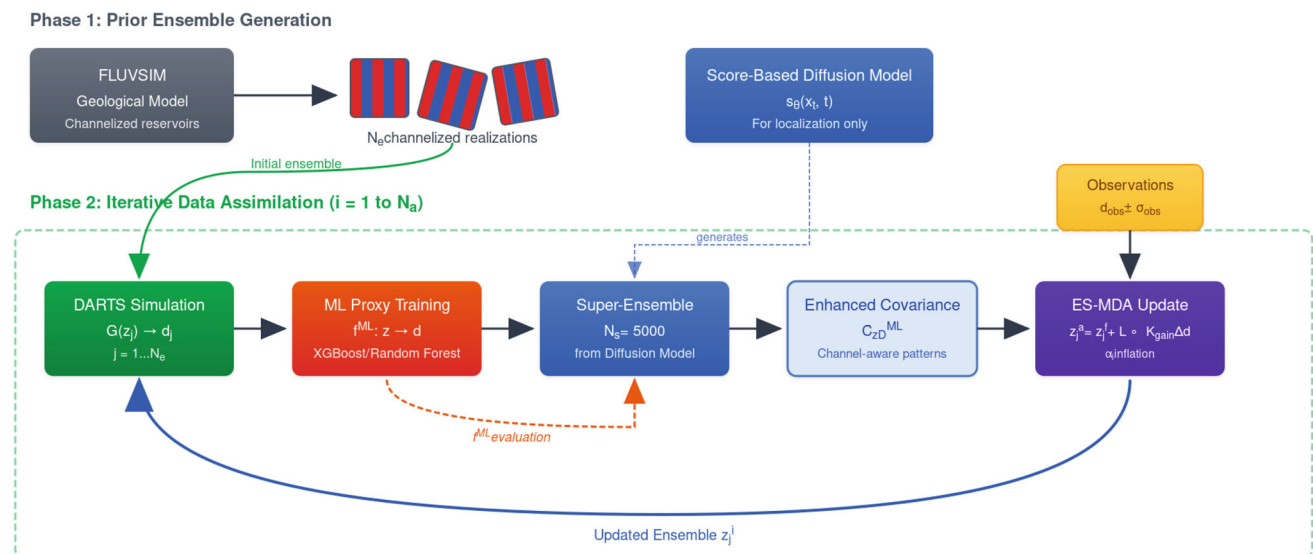


Fig. 6 Integrated workflow combining score-based diffusion models, ML-based localization, and ESMDA. Phase 1: prior ensemble generation via geostatistical models. Phase 2: iterative data assimilation with ML-enhanced localization

metric used for data-match evaluation is the Data-Assimilation-style, noise-normalized root-mean-square misfit

$$\text{RMSE} = \sqrt{\frac{1}{N_d N_e} \sum_{j=1}^{N_e} \sum_{k=1}^{N_d} \left(\frac{d_{\text{obs},k} - G(\mathbf{z}_j)_k}{\sigma_k} \right)^2},$$

where N_d is the number of scalar observations (pressure samples across wells and times), σ_k is the observation standard deviation (1% of the true value in this study), and $\mathbf{z}_j$ is a posterior realization. The RMSE reported for the ML-proxy regression in Table 2 is the standard regression RMSE in the original units of the target pressure, with no normalization by σ_k ; the two quantities are distinct and are labelled accordingly in each case.

When we write that a method achieves “acceptable” variance preservation, the benchmark is the large- N_e no-localization run, for which the NV value is reported in Table 3. Values comparable to or above this reference indicate near-complete variance preservation; values substantially below it indicate collapse.

In what follows, we evaluate each localization method using three metrics: (i) the normalized variance (NV) of log-permeability—our primary metric, which indicates how well the ensemble avoids collapse; (ii) data-match accuracy measured by the RMSE between simulated and observed pressures; and (iii) spatial diagnostics (localization maps and posterior permeability fields) to assess geological realism of the updates.

4.1 Generating geological models with diffusion

Section 3 has already introduced the training set, architecture and sampling procedure; here we focus on the resulting behaviour rather than restating those settings.

The diffusion model converged quickly during training on 3,242 channelized permeability realizations. Starting from an initial loss of 1,882.96, the model achieved stable convergence at 308.76 after only 525 epochs—representing an 83.6% reduction in loss and early termination compared to the planned 5,000 epochs. The optimal model checkpoint, captured at epoch 515 with a loss of 306.46, was selected for all subsequent experiments.

Figure 7 demonstrates the quality of generated permeability fields compared to the original FLUVSIM training data. The comparison shows realizations generated using both ODE and SDE sampling approaches alongside true training samples.

Both samplers generate geologically plausible variations without replicating training samples. In practice, they produce diverse permeability realizations that capture channel geometry and connectivity, which we then use to sample geological models for the data-assimilation experiments.

Beyond the visual comparison in Fig. 7, we rely on two diagnostics of the generated ensemble already described in Section 3: the channel proportion and the value-range check ($k \in [10, 10^4]$ mD, channel proportion 0.40 ± 0.05 matching the FLUVSIM prior). Any residual overestimation of extreme-high permeabilities visible in Fig. 7 does not propagate into the localization matrix, because the proxy f^{ML} is used only to estimate the cross-covariance structure, which is insensitive to isolated outlier pixels. The generated ensemble is therefore fit for purpose as a super-ensemble for covariance estimation, which is the only role it plays in the localization workflow.

Table 3 Normalized Variance (NV) results for different localization methods across ensemble sizes

Method	$N_e=50$	$N_e=100$	$N_e=200$	$N_e=500$	$N_e=1000$
No Localization	0.400	0.632	0.714	0.801	0.765
Pseudo-Optimal	0.577	0.740	0.797	0.835	0.787
Gaspari-Cohn	0.749	0.857	0.904	0.927	0.882
ML-Linear	0.570	0.708	0.780	0.825	0.791
ML-Random Forest	0.864	0.885	0.871	0.861	0.811
ML-XGBoost	0.821	0.872	0.858	0.857	0.805

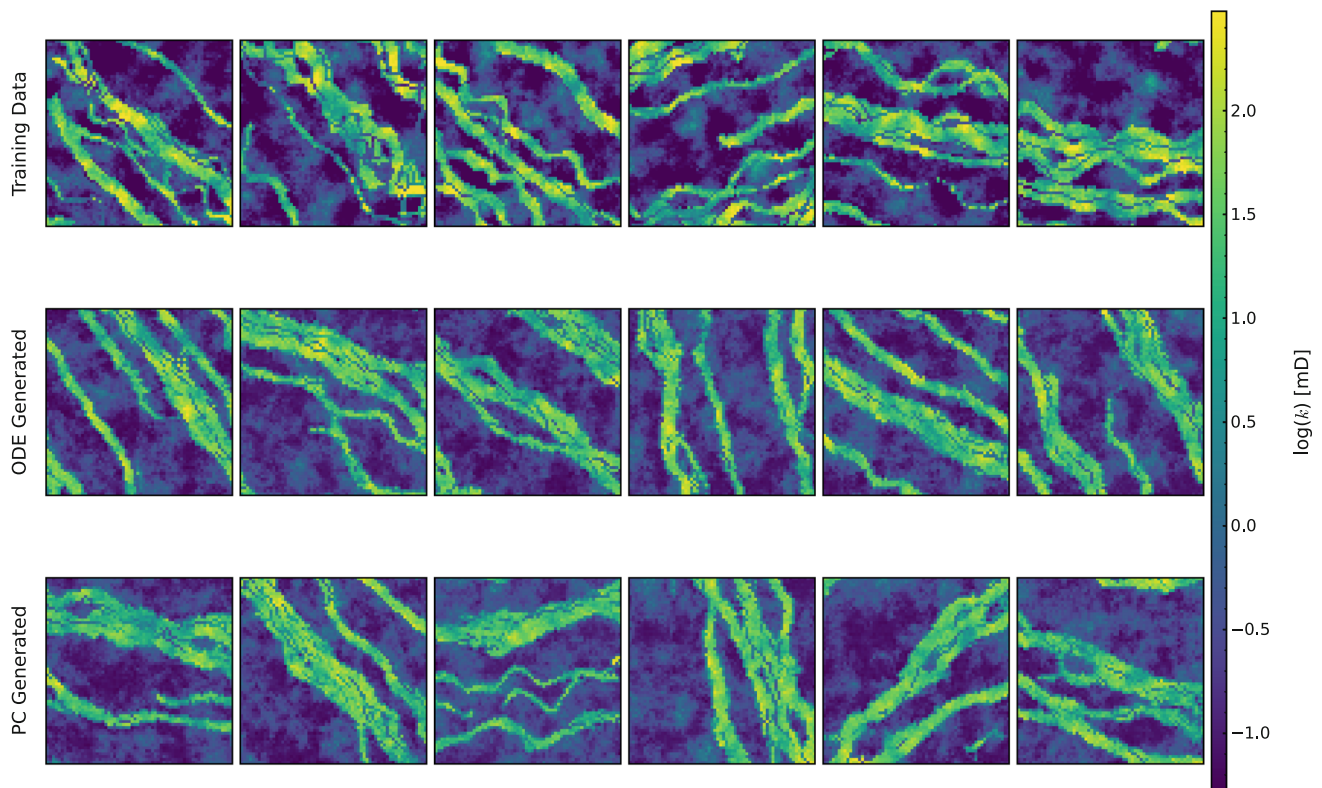


Fig. 7 Visual comparison of training data (top row) with generated realizations using ODE sampling (middle row) and SDE sampling (bottom row)

4.2 Machine learning proxy for localization

We evaluate three machine learning algorithms for predicting localization coefficients: linear regression, random forest, and XGBoost. Table 2 summarizes their performance across different ensemble sizes, revealing distinct patterns that inform our framework design.

Tree-based methods (Random Forest and XGBoost) consistently generate better results than linear regression, achieving RMSE values below 0.31 across the full range of ensemble sizes. The much less consistent RMSE values of the linear regression (RMSE values ranging from 0.268 to 2.828) illustrate that this method is not as effective in capturing the pressure response in the strongly heterogeneous channelized systems.

While Random Forest achieves the lowest prediction errors for most ensemble sizes, computational cost analysis reveals important practical considerations. Training time for Random Forest increases with ensemble size, reaching 106.8 seconds for $N_e = 1000$. XGBoost offers an attractive alternative, providing comparable accuracy (RMSE within 0.02 of Random Forest) with 33% lower computational cost for large ensembles.

The tree-based proxies in Table 2 achieve RMSE values below 0.31 across the full range of ensemble sizes, whereas the linear proxy ranges from 0.268 to 2.828. The localization matrix in Eq. 14 depends on the proxy only through the relative structure of the parameter to data covariance, not on absolute prediction accuracy. The tree-based proxy accuracy is therefore sufficient for the localization task; the linear proxy is included in the comparison as a stress test, and its larger RMSE is consistent with the visibly noisier ML-linear localization patterns later in the paper.

4.3 Data assimilation performance

ML localization's main benefit shows up when examining the spatial patterns of parameter updates. Figure 8 illustrates representative localization patterns for medium ($N_e = 200$) and large ($N_e = 500$) ensemble sizes, demonstrating how ML-enhanced methods adapt to the channelized reservoir structure.

A comparison across all ensemble sizes and methods is presented in Fig. 9.

Traditional distance-based localization (Gaspari-Cohn) produces circular update regions centered on observation wells, indiscriminately modifying both channel and background permeabilities. This approach fails to respect the fundamental geological structure where pressure information propagates preferentially along high-permeability channels. In contrast, ML-enhanced methods generate updates that follow channel geometries, concentrating modifications where they are geologically meaningful. This channel-aligned updating is maintained consistently across all ensemble sizes, demonstrating the robustness of the ML approach.

Further details and additional localization maps for the remaining ensemble sizes ($N_e = 50, 100, 1000$) are provided in Appendix B, which illustrates the spatial features of the localization coefficients for each case.

The posterior permeability fields for the different localization strategies are illustrated in Fig. 10. The figure displays a representative posterior permeability sample for different ensemble sizes and localization methods. For small ensemble sizes ($N_e = 50$), the absence of localization or the use of pseudo-optimal localization leads to severely blurred permeability fields. The updates affect the background facies indiscriminately, and even the channel structures lose their sharpness. In contrast, the ML-based localization methods (Random Forest and XGBoost) preserve the channel connectivity and overall geological structure remarkably well, even with a small ensemble. The updates exhibit a channel-like form, which is a direct consequence of the geology-aware localization patterns shown on Figs. 8 and 9. The Gaspari-Cohn localization produces updates that are distinctly circular, reflecting its distance-based nature and leading to less geologically consistent modifications. As the ensemble size increases, the benefits of ML-based localization become less pronounced, as expected. However, even for larger ensembles ($N_e = 500, 1000$), the permeability fields generated with ML-localization exhibit the channelized structure more realistically compared to other methods. For large ensembles, ML-enhanced localization still preserves fine-scale geological detail better than distance-based tapers or no localization, although the margin of improvement is smaller than at low N_e .

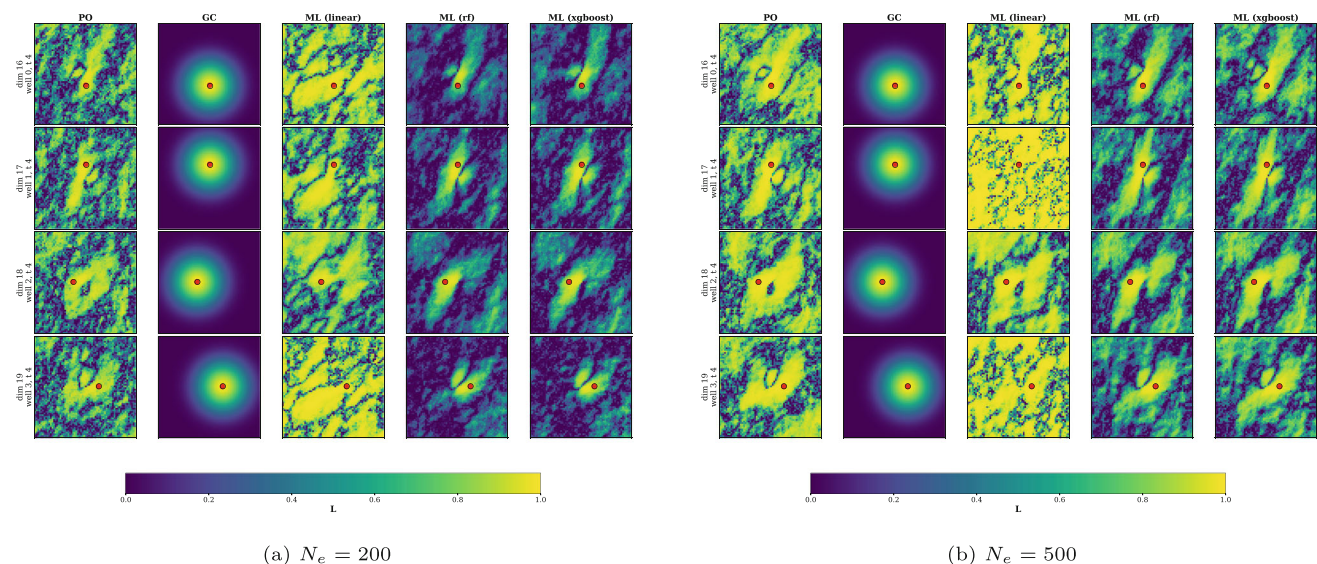


Fig. 8 Representative localization patterns for medium and large ensemble sizes showing channel-aligned parameter updates for different methods. The data point location is marked with a red circle. Unlike traditional distance-based methods (GC) that produce circular patterns, ML-enhanced localization (ML methods) adapts to the channelized permeability structure, concentrating updates along high-permeability pathways

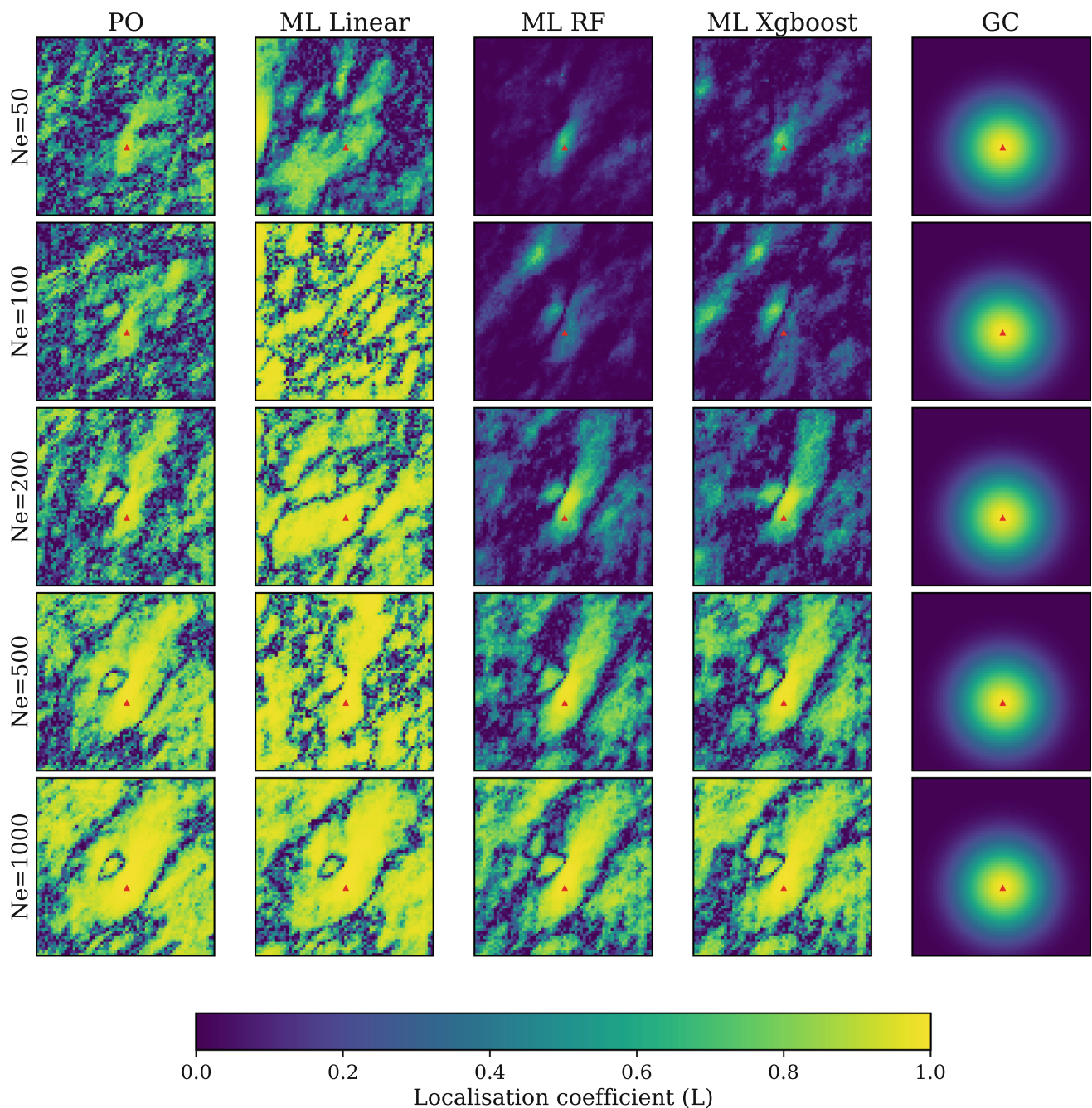


Fig. 9 Localization patterns for different ensemble sizes ($N_e = 50$ –1000) and methods. ML methods produce channel-aligned patterns; traditional methods show circular patterns. Each panel displays the localization mask $\mathbf{L}$ alone, not the localized cross-covariance $\mathbf{L} \circ \mathbf{C}_{zD}$. The mask is bounded between 0 and 1, with 1 indicating that the corresponding parameter to data correlation is fully retained and 0 that it is suppressed. Optimality is therefore not a single value: the desired mask should be close to 1 along physically connected high-permeability paths and close to 0 in the background. The ML masks satisfy this qualitative target; GC, by construction, satisfies it only along radial directions. The localized correlations themselves are summarized through their effect on the posterior fields (Fig. 10), the per-cell update magnitude (Fig. 11), and the variance preservation metric (Fig. 14)

To visualize *where* and *how strongly* each method actually modifies the model, Fig. 11 shows the absolute change in log-permeability (posterior – prior) for the same realization shown in Fig. 10. The machine-learning localizations concentrate updates along the high-permeability channel that connects to the observation well (green

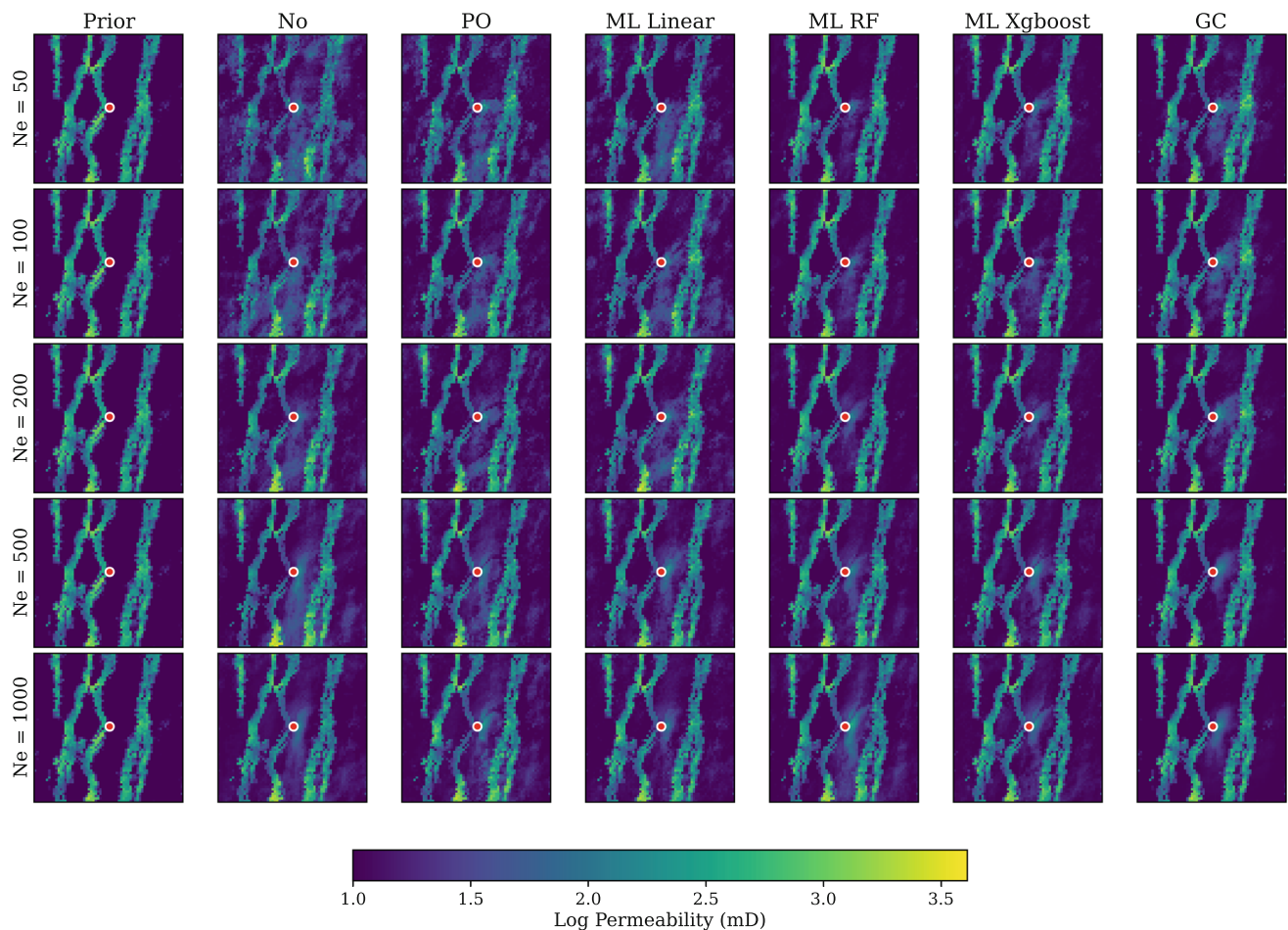


Fig. 10 Sample of posterior permeability fields for different ensemble sizes (N_e) and localization methods. ML-based methods preserve channel structures better than other methods, especially for small ensembles

marker), whereas the un-localised cases disperse changes into the background. This effect is larger for $N_e=50$ and 100 members. In these cases, the Random Forest– and XGBoost–based localizations concentrate updates along connected channels and preserve ensemble variance better than pseudo-optimal, Gaspari–Cohn, or no localization; the linear ML variant performs in between the tree-based and classical methods.

To quantify the quality of data matching across different ensemble sizes, Fig. 12 presents the RMSE values achieved by each method, providing a comprehensive measure of data matching quality.

The RMSE analysis reveals distinct performance patterns across ensemble sizes. For larger ensembles ($N_e \geq 100$), all localization methods achieve acceptable data matching quality. However, very small ensembles ($N_e = 50$) show degraded performance across all methods, suggesting this ensemble size may be insufficient for reliable data assimilation in channelized systems, even with enhanced localization techniques. Despite these limitations at $N_e = 50$, ML-enhanced methods deliver clear gains in variance preservation and in the geological realism of the posterior fields, while data-mismatch RMSE is comparable to, and at small ensemble sizes slightly above, that of GC and PO. We therefore separate the two performance axes throughout the rest of the paper: the ML methods are presented as the preferred option when variance preservation and channel-aware updates are the priority, and not as a uniformly lower-RMSE alternative.

To illustrate the history matching in detail, Fig. 13 presents pressure evolution at four monitoring wells throughout the 10-year CO_2 injection period for the $N_e = 500$ case, comparing different localization methods. All methods track the observed pressure evolution, with ensemble mean predictions closely following the true data.

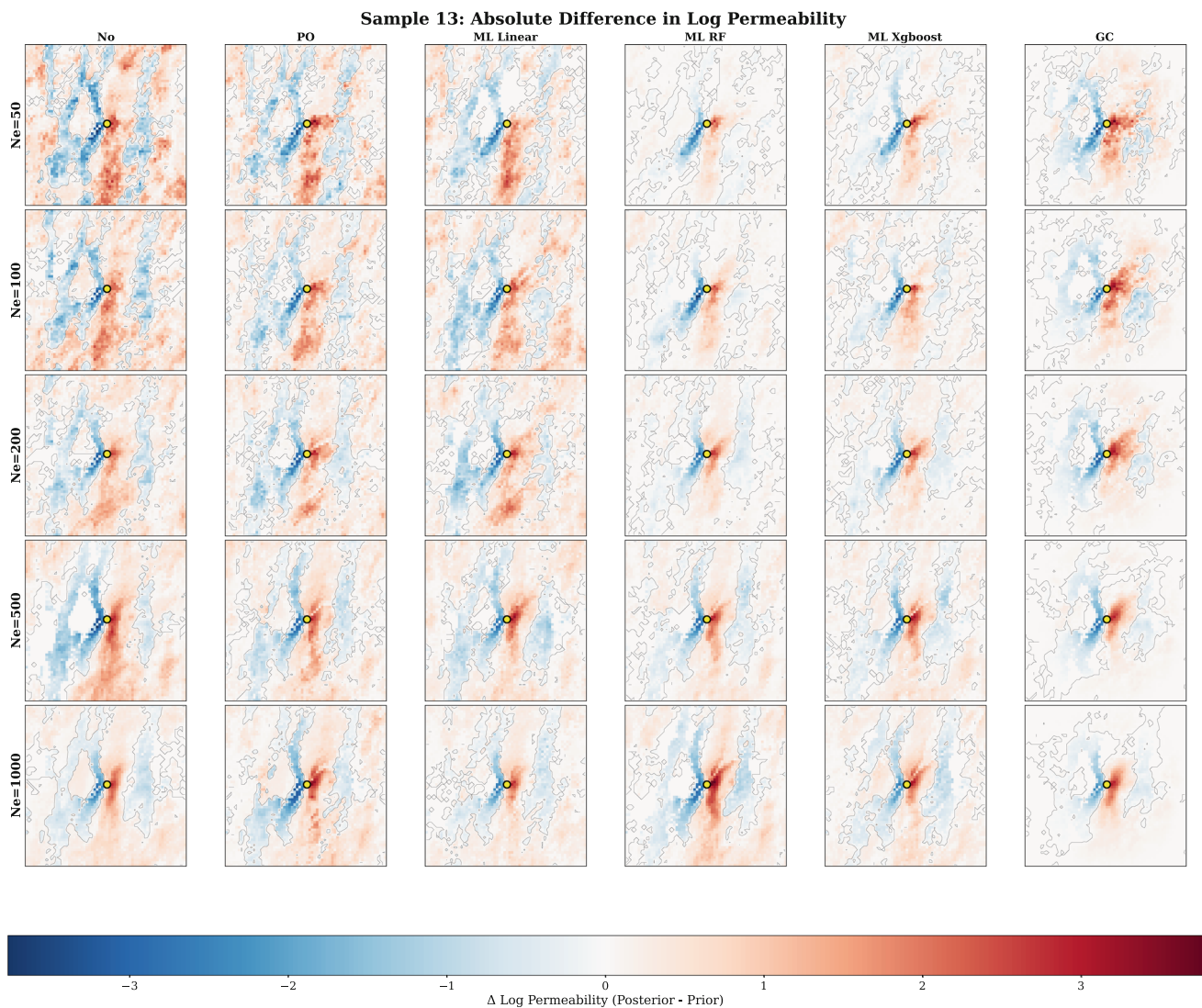


Fig. 11 Log-permeability change (posterior - prior) as a function of ensemble size N_e and localization method for the following methods: PO, linear machine learning (ML-linear), random forest (ML RF), xgboost (ML XGBoost). The single colorbar at the right applies to all panels

The effectiveness of different localization methods in preventing ensemble collapse is quantified through normalized variance (NV), our primary performance metric. The higher the NV, the better the variance is preserved in the data assimilation and the less likely the occurrence of ensemble collapse. Table 3 and Fig. 14 present comprehensive results across all tested configurations.

Random Forest and XGBoost have the highest NV across all ensemble sizes. The improvements are most pronounced for small ensembles where sampling errors are severe. For $N_e = 50$, ML-Random Forest achieves $NV = 0.864$ compared to only 0.400 without localization. This is a 116% improvement that transforms a collapsed ensemble into one that preserves its variance.

We believe the parameter-space diagnostics in Figs. 10 and 11 and the data-misfit objective in Fig. 12 evaluate two different aspects of the same assimilation and should not be expected to agree pixel for pixel. Figures 10 and 11 show parameters (posterior log-permeability and per-cell update); their quality is naturally judged by geological realism, channel preservation and update locality. Figure 12 reports the data-misfit objective function between simulated and observed pressure traces, whose value tends to be reduced by any update that better fits the four pressure curves, regardless of how the parameter field looks. It is therefore reasonable to expect the

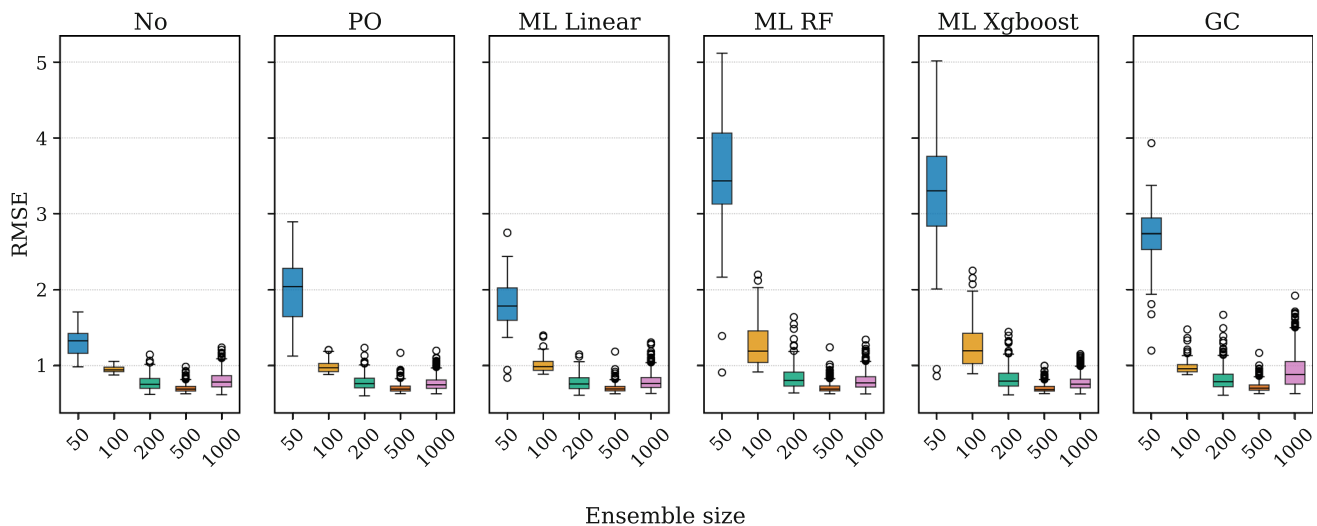


Fig. 12 RMSE comparison across ensemble sizes. The RMSE here is the noise-normalized data-misfit defined at the start of Section 4 and is computed against the observed pressures, not against the prior or true permeability field. It is therefore a measure of data match, not of geological realism, and should be read jointly with Figs. 10, 11 and 14

data-misfit objective to be improved most by aggressive, broadly distributed updates, of the kind produced by No-localization and PO. In our interpretation, they achieve the lowest RMSE by redistributing the parameter update over a larger portion of the domain, including the background, at the cost of inflating the field with noise (visible as diffuse, large-amplitude changes in Fig. 11). The ML-based methods appear to concentrate updates along the connected channel and leave the background almost untouched, which in our view yields posterior fields that look more realistic geologically and higher NV, at the price of slightly larger data-misfit RMSE in the small-ensemble regime. Figure 11 uses a single shared colorbar across all panels. We suggest reading the two evaluation axes (parameters and objective function) jointly rather than as competing claims.

The mild deterioration of GC at $N_e = 1000$ visible in Fig. 12 and the corresponding NV reduction in Fig. 14 do not, in our view, indicate an inadequate retuning. The GC cutoff was kept fixed across all N_e so that differences would be attributable to the localization scheme rather than to per-ensemble parameter sweeps. We believe the reduction at $N_e = 1000$ can be rationalised by two effects acting in opposite directions: first, sample noise in the cross-covariance is small enough that a fixed GC taper may begin to over-suppress real long-range correlations along the channels; second, the larger ensemble samples a wider portion of the prior, which plausibly includes realizations whose pressure response is harder to match within a fixed number of ESM DA iterations. It is reasonable to think that these effects also contribute to the small NV decrease at $N_e = 1000$ for all methods, including the ML variants. A budget-controlled extension of the assimilation schedule (more ESM DA iterations or finer inflation) would likely narrow the gap, but $N_a = 4$ has been kept identical across methods and ensemble sizes so that the comparison is not biased by per-configuration tuning. We therefore suggest reading the mild non-monotonicity at $N_e = 1000$ as consistent with the iteration budget becoming binding for the largest ensemble rather than as an artefact of localization tuning, while acknowledging that a dedicated sensitivity study would be needed to settle the question conclusively.

The framework exhibits distinct variance preservation across different ensemble size regimes:

Small ensembles ($N_e < 100$): ML-enhanced localization may be necessary to prevent variance collapse. Without proper localization, small ensembles lose the channelized geological structure and underestimate uncertainty. ML methods (RF, XGBoost) restore viable uncertainty quantification with minimal computational overhead.

Medium ensembles ($N_e = 100 - 500$): ML methods provide benefits, with variance preservation improvements of 20-40% over no localization. For $N_e = 200$, ML methods achieve NV values of 0.86-0.87 compared to 0.71 without localization, maintaining channelized structures throughout the assimilation process.

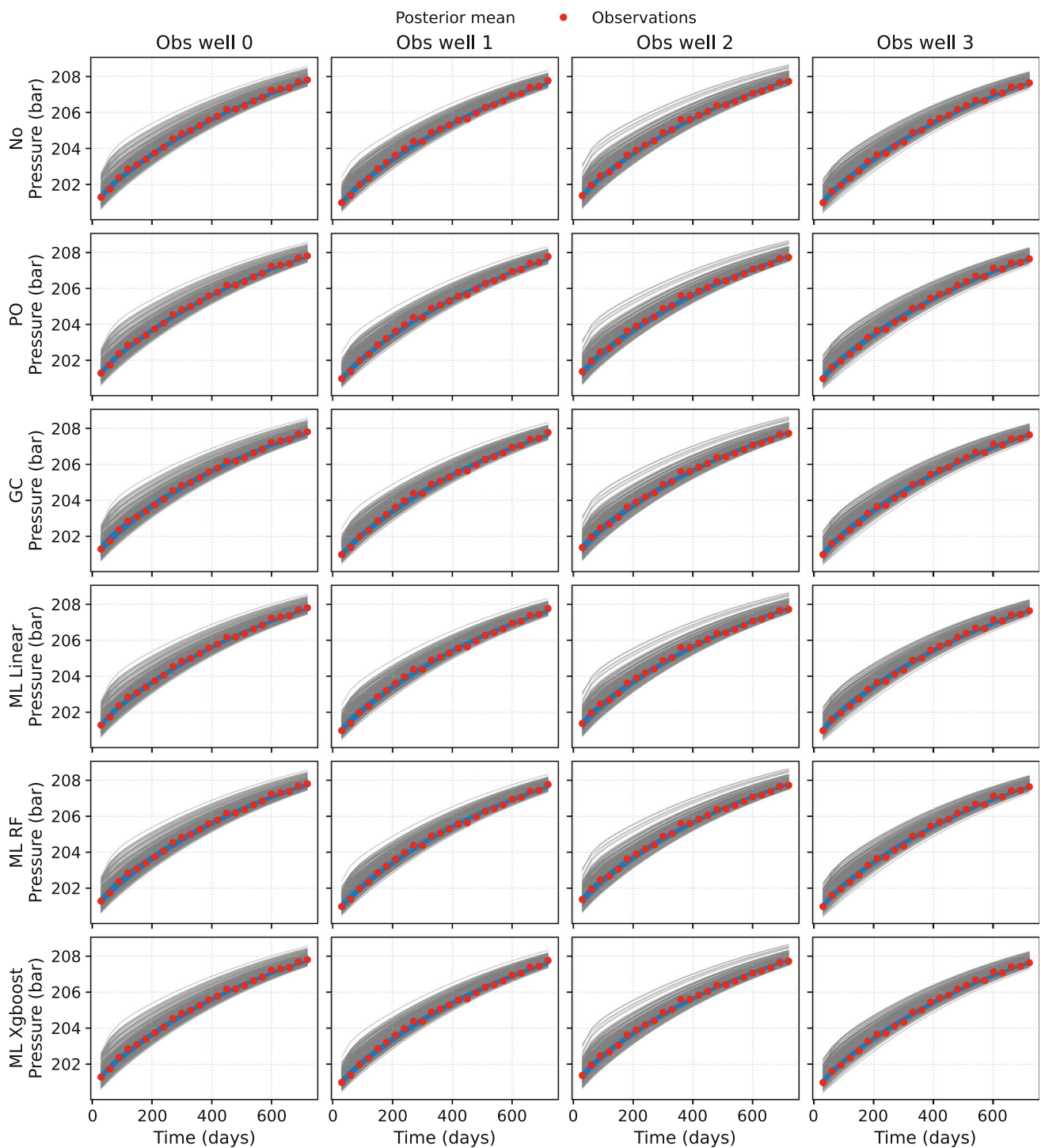
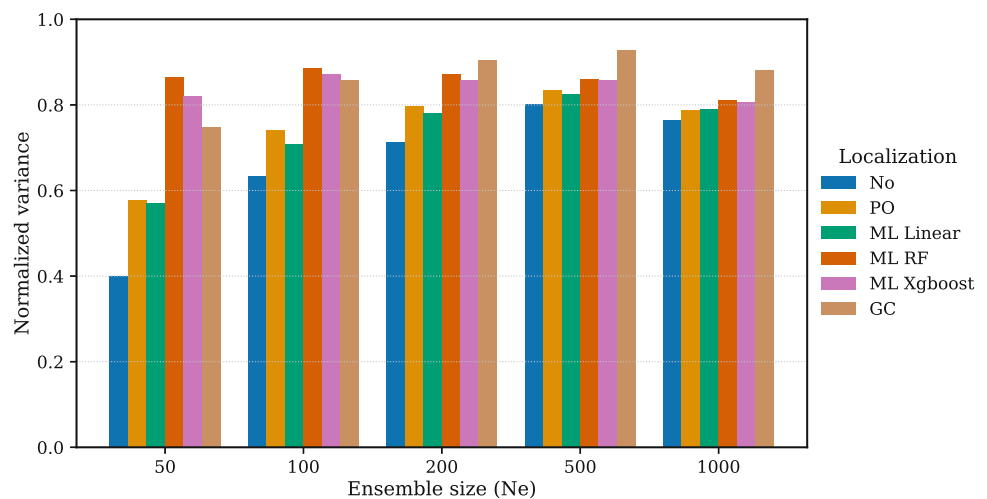


Fig. 13 Data match for $N_e = 500$ across four monitoring wells. True data (red), posterior (blue), prior (gray) curves

Large ensembles ($N_e > 500$): Benefits diminish as natural sampling errors decrease, though ML methods (RF and XGBoost) still show marginal improvements. For $N_e = 1000$, all methods achieve acceptable variance preservation ($NV > 0.78$), with ML approaches reaching 0.81.

Fig. 14 Normalized variance comparison across ensemble size for each of the different localization approaches



5 Discussion

This study investigates whether a combination of score-based diffusion models and machine learning proxies can overcome two persistent drawbacks of ensemble smoothers applied to channelized CO₂ storage: systematic variance loss caused by limited ensemble sizes and the geologically unrealistic permeability patterns imposed by distance-based localization. The numerical experiments confirm that the proposed workflow makes progress on both fronts. This represents a step beyond traditional ensemble methods [15, 65], which can suffer from variance underestimation and spurious correlations due to limited ensemble sizes. By leveraging machine learning proxy models, such as those recently explored by Lacerda et al. [29], Silva et al. [27], and Silva et al. [28], our methodology effectively suppresses spurious correlations in data assimilation.

The variance preservation gain delivered by the Random Forest and XGBoost localizations appears largest in the small-ensemble regime: at $N_e = 50$, NV moves from 0.40 without localization to roughly 0.86 with ML-RF, and at $N_e = 100$ the corresponding values are 0.63 and 0.89. At $N_e = 200$ the ML and GC results are within roughly 3 NV percentage points of each other, and for $N_e \geq 500$ GC reaches the highest NV. We therefore suggest that the main operational message of the variance comparison is: ML-localization looks preferable in the regime in which ensemble methods most need help (50 to 200 members), and the gain over GC appears to close as the ensemble grows. GC may be a reasonable default choice for $N_e \geq 500$ when only NV is considered. A plausible reason for the relative drop of ML methods at large N_e is that the proxy-based super-ensemble is expected to add value mainly when the working-ensemble cross-covariance is noisy, and that benefit naturally shrinks as N_e grows and the working-ensemble covariance itself becomes better estimated.

One point deserves emphasis: in our workflow, the machine-learning proxies are never used as replacements for the flow simulator inside the Kalman update itself; they serve exclusively to estimate the localization matrix. This design choice matters because localization depends only on the relative structure of the parameter–data covariance, not on the exact reproduction of the pressure traces. Consequently, even proxies that would be inadequate as full surrogate models prove perfectly adequate—and far cheaper—to guide localization. By confining the proxy’s role to this task, we avoid the risk of biasing the analysis with approximate physics while still reaping the statistical benefit of a much larger effective ensemble.

Equally important is the qualitative improvement in the spatial structure of the updates. Visual inspection of the localization maps and the posterior permeability fields shows that the machine-learning approach concentrates

corrections along high-permeability channels and generates less perturbations on the low-permeability background. Distance-based kernels, by contrast, produce circular footprints that severely underestimate geological connectivity.

The computational overhead of ML-localization is small. Training the proxy at each assimilation step adds a few seconds for ensembles up to 200 members and less than 30 seconds for 500 members, which is minor relative to even a single forward DARTS run, which for this case takes about 2–3 minutes of wall-clock time per realization on one modern CPU core.

The proxy is retrained at every ESM DA iteration because, although the diffusion model and its super-ensemble are iteration independent, the working ensemble is not: at iteration i , the parameter cloud has been pulled toward the data and its joint distribution with $\mathbf{d}$ is no longer the prior joint distribution. Training the proxy f^{ML} once on the prior-iteration ensemble and freezing it would tie the localization mask to the prior cross-covariance throughout the assimilation, which defeats the purpose of running multiple ESM DA iterations. By retraining f^{ML} on the current $\{\mathbf{z}_j^{i-1}, \mathbf{d}_j\}$ pairs, the super-ensemble predictions inherit the local response surface around the current ensemble, and the localization mask becomes iteration dependent. The cost of this retraining is given by the ML training times in Table 2 and is small relative to the N_e forward simulations required in the same iteration.

ESM DA is a Gaussian, two-point method and its assumptions are indeed violated by channelized priors. It is nonetheless used here for two reasons. First, ESM DA remains the operational baseline in industrial reservoir history matching, and we believe that demonstrating how variance preservation can be improved within this baseline is practically more useful than introducing a fully non-Gaussian sampler. Second, the diffusion-based prior, the super-ensemble, and the ML-enhanced localization are designed to act on the components of ESM DA most affected by non-Gaussianity (the prior shape and the cross-covariance), which may help explain why the method continues to deliver geologically realistic posteriors even though the analytical update remains linear and Gaussian. The augmentations are intended to bring the practical posteriors closer to what a non-Gaussian sampler would produce, at a much smaller cost; a formal benchmark against a full non-Gaussian sampler is left for future work.

Channels are not guaranteed to be preserved bit by bit during the assimilation; the Schur-product update is element-wise linear and can move pixels across the channel/background threshold. The posterior maps in Fig. 10 and the variance metric suggest, however, that the ML-enhanced localizations appear to preserve channel connectivity to a greater extent than distance-based or unlocalized variants in the small-ensemble regime, and that the support of the posterior seems to remain broadly within the channelized class generated by the prior. A formal preservation guarantee would require coupling the diffusion prior to the ESM DA update through a non-linear projection, which is beyond the scope of this paper.

It is a legitimate concern that the diffusion super-ensemble samples regions of parameter space that the working-ensemble proxy was not trained on, and that individual pressure predictions on those samples may therefore be unreliable. We believe this does not invalidate the workflow because the proxy is not used as a forward model: it enters only through the cross-covariance that feeds the localization mask in Eq. 14, which depends on the ranking and relative magnitude of the parameter to data correlations rather than on the absolute pressure values. As long as the proxy preserves the qualitative correlation structure on the support of the diffusion prior, we expect the localization mask to remain informative even if individual pressure predictions are off. The numerical consistency of the resulting masks across ensemble sizes (Fig. 9) is, in our interpretation, consistent with this view. Adding gas-saturation observations from an injection scenario, where surrogate accuracy on out-of-distribution samples would be a much more demanding test, is a natural next step and is identified as future work. For pressure-only data of the kind used here, we believe the design is justified.

For very small working ensembles ($N_e = 50$), Fig. 14 and Table 3 report that the ML-enhanced localization still performs favourably in the metric where the question is operationally most pressing (variance preservation, where NV moves from 0.40 without localization to 0.86 with ML-RF). We believe the underlying mechanism is

statistical rather than physical: the proxy does not need to learn the full non-linear pressure response, but only needs to provide a parameter to data covariance whose structure is plausibly closer to truth than the noisy sample covariance estimated from 50 simulator runs. It seems reasonable to expect that tree-based proxies trained on 50 realizations of a 4096-dimensional input can do this because the relevant signal in the pressure response to a permeability perturbation is likely concentrated on a small number of effective directions (channel connectivity, distance to wells), for which 50 samples may be sufficient. A formal quantification of this effective dimensionality is beyond the scope of this study.

The ML-localization method has limitations: We only tested two-dimensional scenarios, leaving the potential computational complexities of three-dimensional geological formations unexplored. The extension to 3D would require addressing memory constraints for diffusion model training. The current GPU architectures limit practical grid sizes to approximately 128^3 for full-resolution models. Given that most operational reservoir models comprise thousands of grid cells, this limitation needs to be addressed to apply ML-localization in an operational setting. Additionally, our analysis focuses on pressure data integration, neglecting other geophysical data types such as seismic or electromagnetic observations, commonly incorporated in multi-physics integration studies. The reliance on offline training of ML proxy models at each assimilation iteration also poses scalability and flexibility challenges, which asks for further exploration of online learning strategies.

6 Conclusions

In this work, we combine score-based diffusion models, machine-learning proxies, and the Ensemble Smoother with Multiple Data Assimilation (ESMDA) to deliver a variance-preserving, channel-preserving history-matching strategy for channelized CO₂ storage reservoirs. The diffusion model generates thousands of geologically plausible permeability fields at negligible cost once trained. The ML proxies exploit these large synthetic ensembles to build more reliable covariance estimates, and the resulting localization mitigates ensemble collapse without additional flow simulations. In tests with two-dimensional channel systems, the approach preserved up to 40% more variance than standard distance-based tapers, maintained channel connectivity in the posterior fields, and introduced only minor computational overhead.

These results demonstrate that data-driven localization can significantly reduce the gap between small ensembles that are affordable in practice and the large ensembles required for accurate covariance estimation. Extending the framework to three dimensions, adding multi-physics data streams, and adopting online learning for the proxies (i.e., realtime development of ML models as the model runs forward) in a data assimilation framework with ML-localization are the logical next steps toward real-time, uncertainty-aware management of large-scale CO₂ storage projects.

Appendix A: Data match plots for varying ensemble sizes

Figures 15–19 present the history-matching performance for all ensemble sizes considered in this study. Each figure shows the evolution of pressure at the four monitoring wells over the two-year injection period for the corresponding ensemble size N_e .

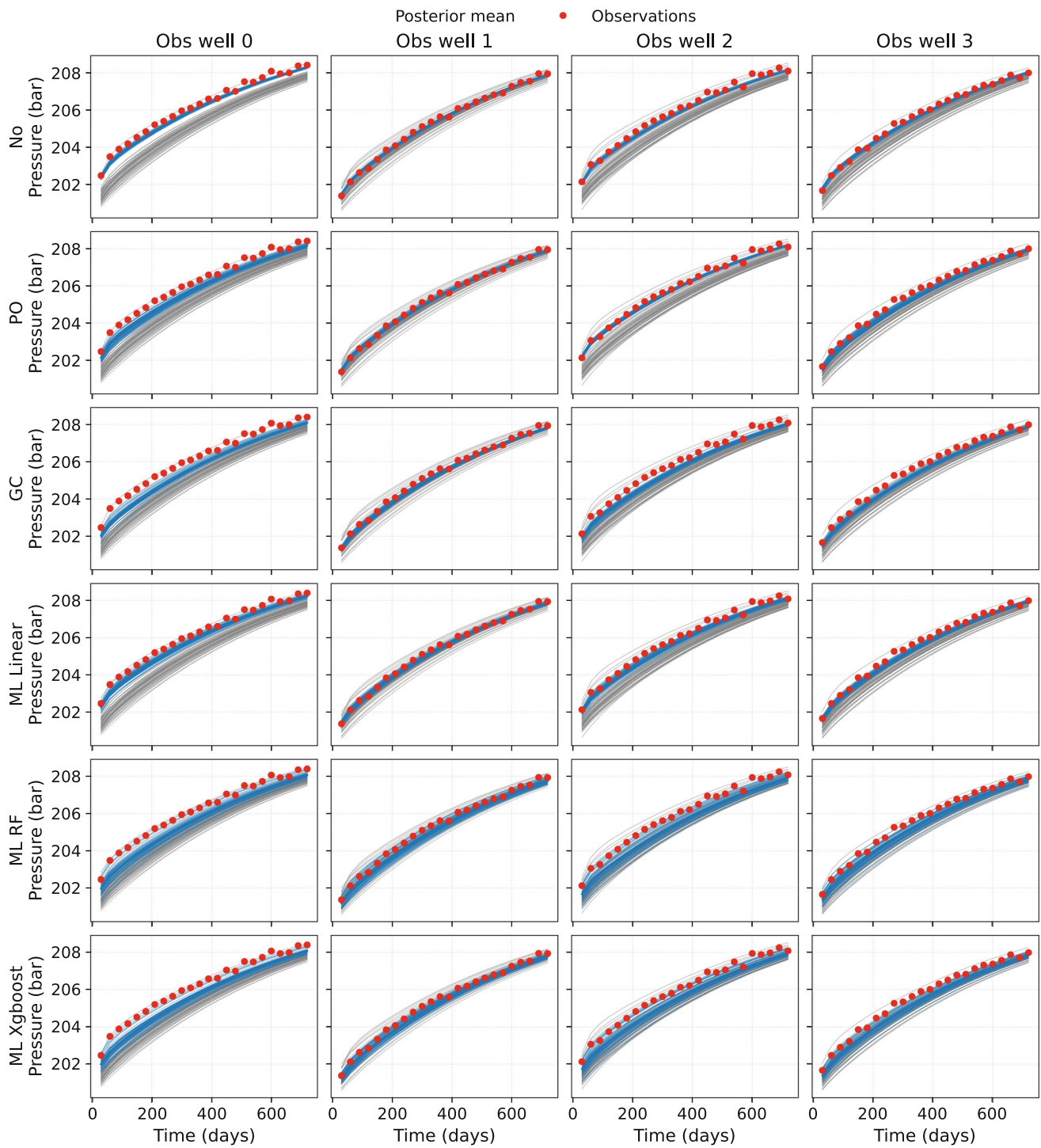


Fig. 15 Data match results for $N_e = 50$

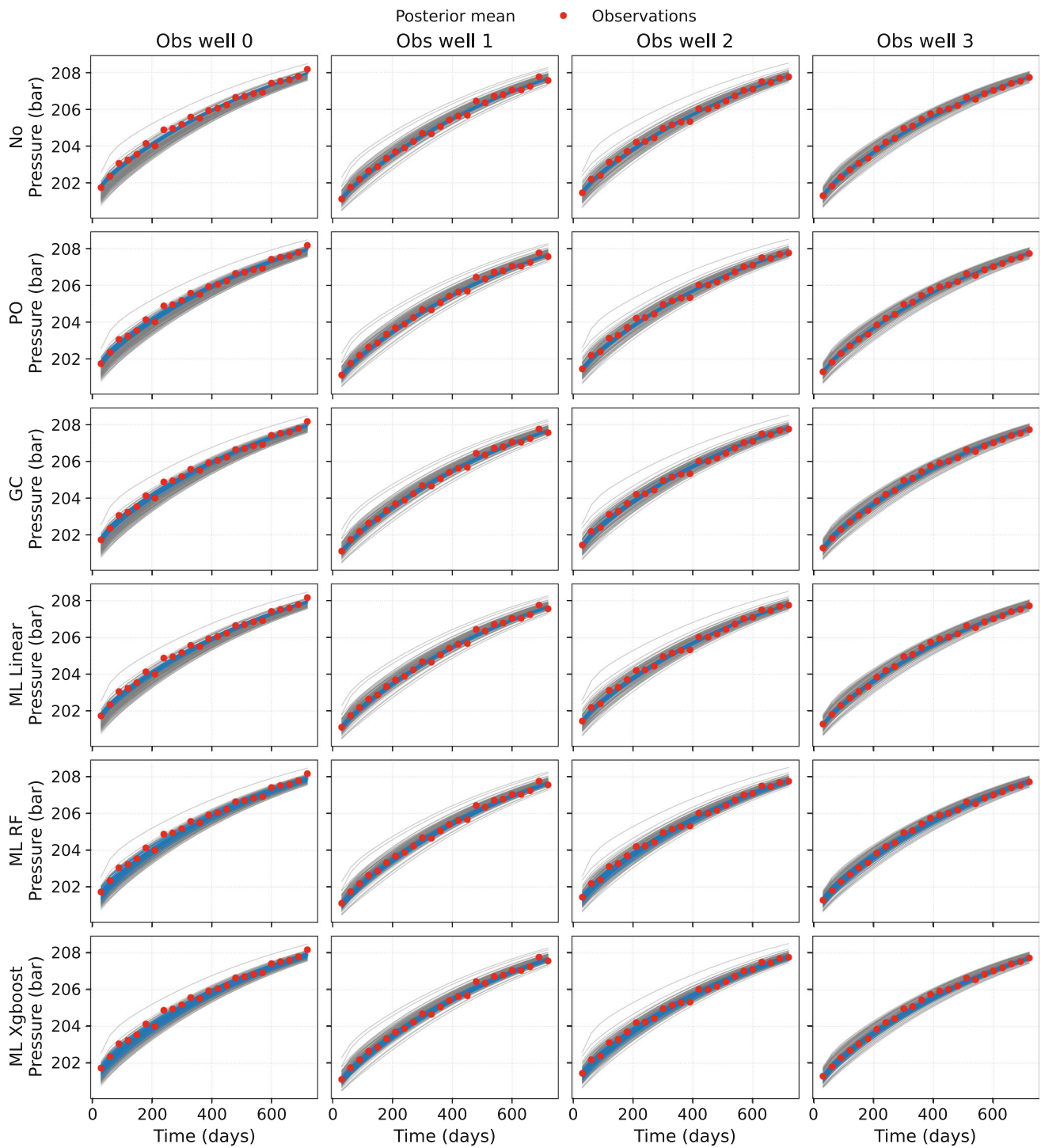


Fig. 16 Data match results for $N_e = 100$

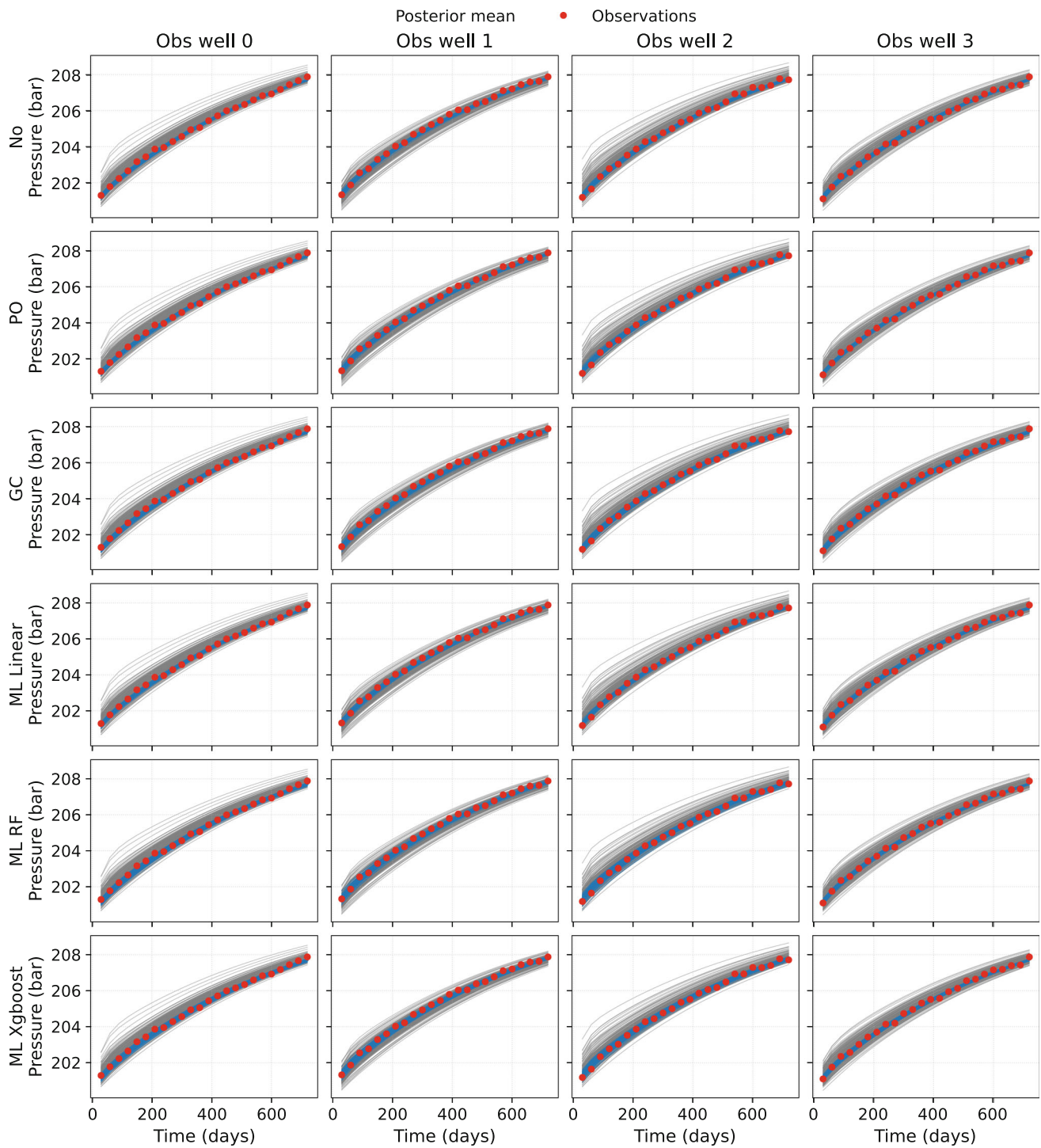


Fig. 17 Data match results for $N_e = 200$

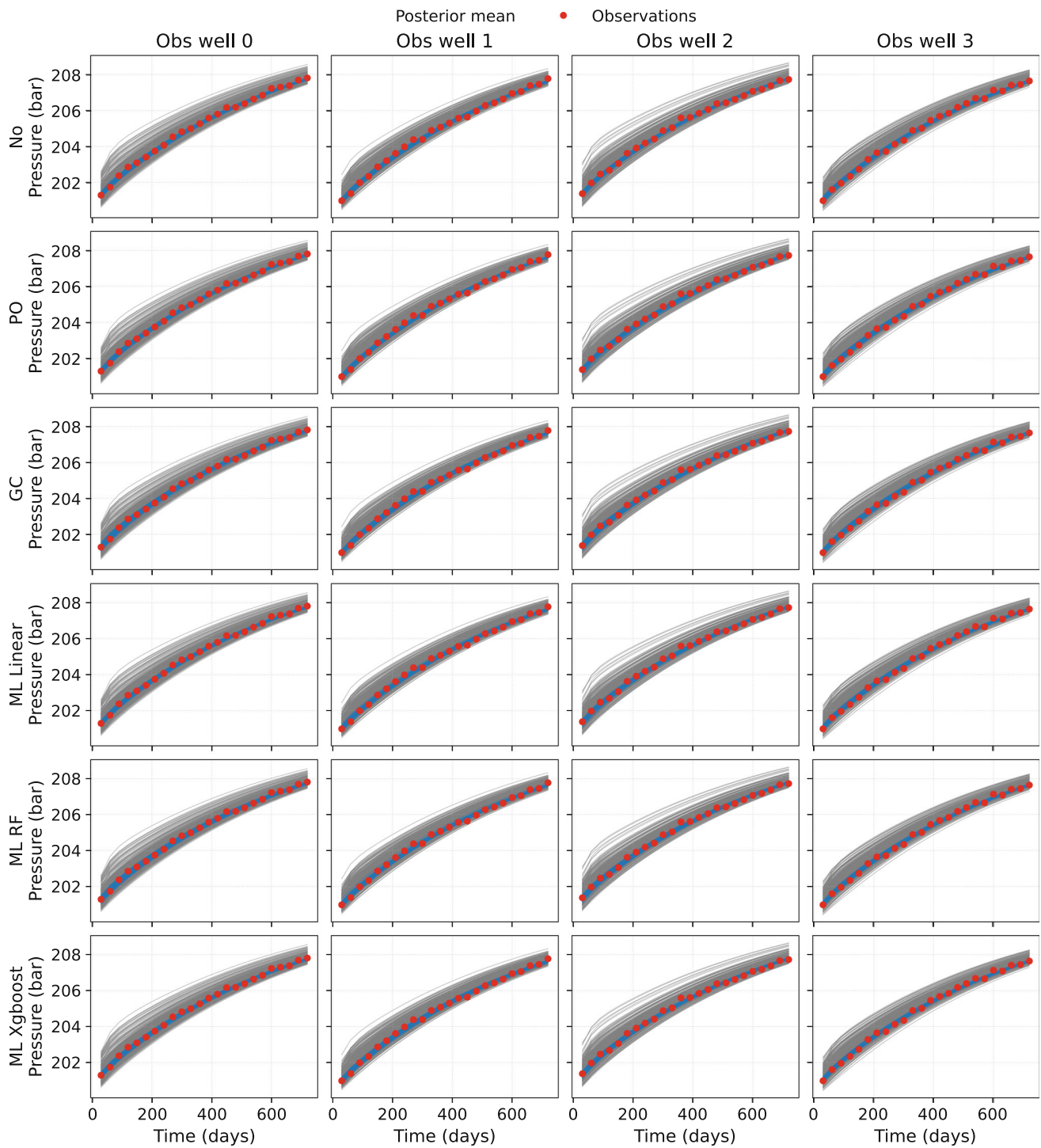


Fig. 18 Data match results for $N_e = 500$

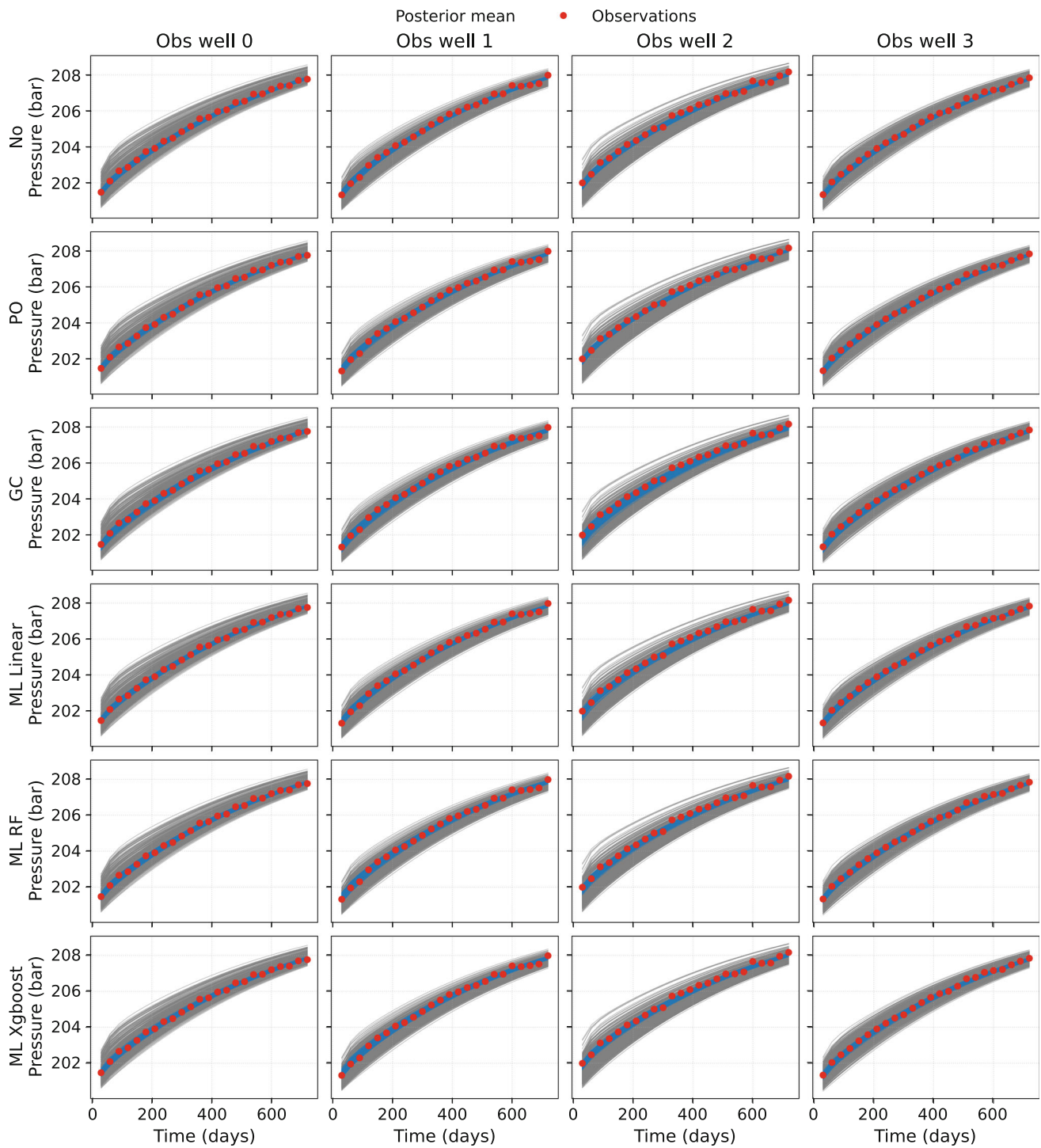


Fig. 19 Data match results for $N_e = 1000$

Appendix B: Localization maps for varying ensemble sizes

This appendix presents the localization maps for all ensemble sizes considered in this study ($N_e = 50, 100, 200, 500, 1000$). Each figure shows the spatial pattern of the localization coefficients for a representative observation well and assimilation time step, illustrating how the localization adapts to the channelized reservoir structure for different ensemble sizes.

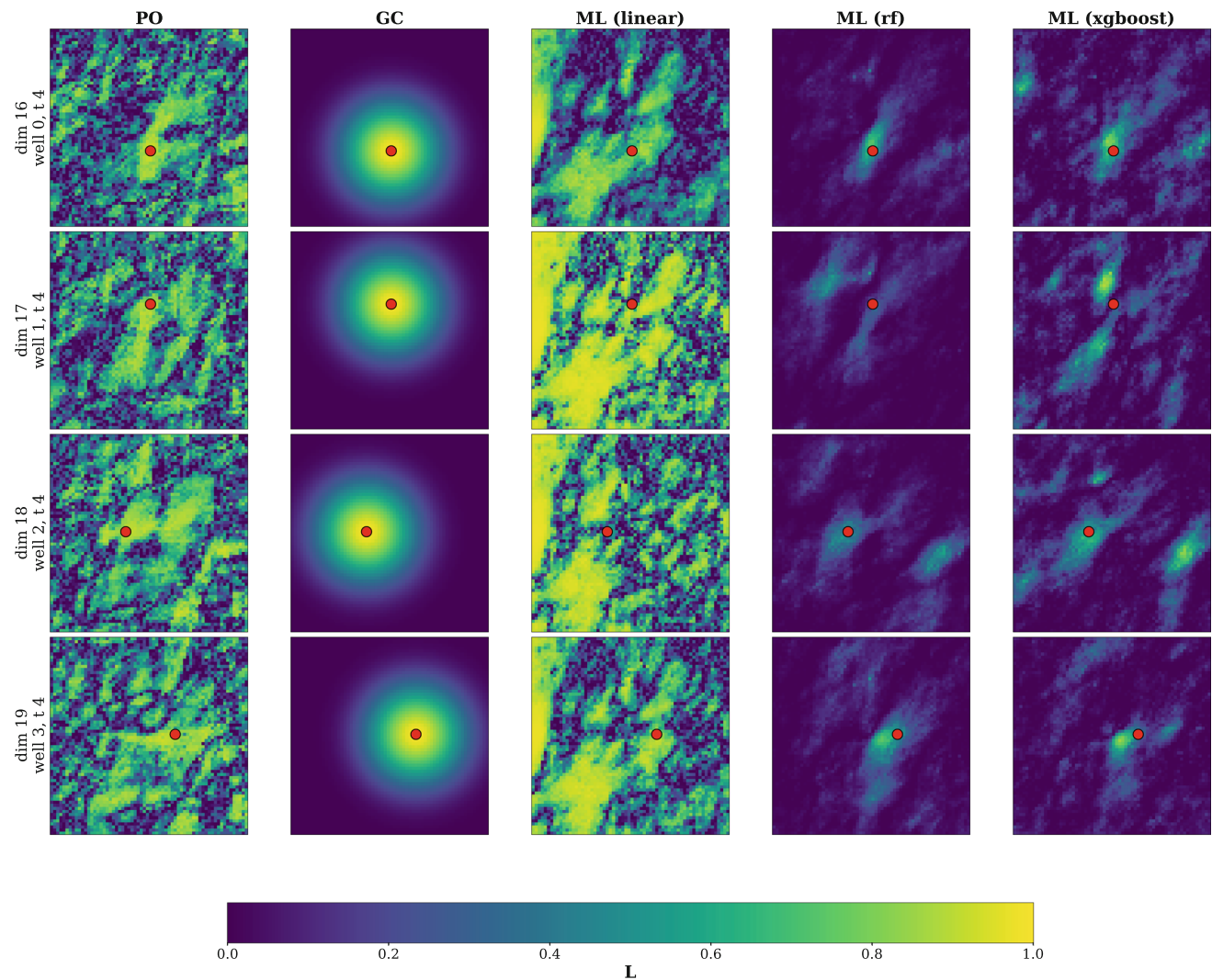


Fig. 20 Localization map for $N_e = 50$

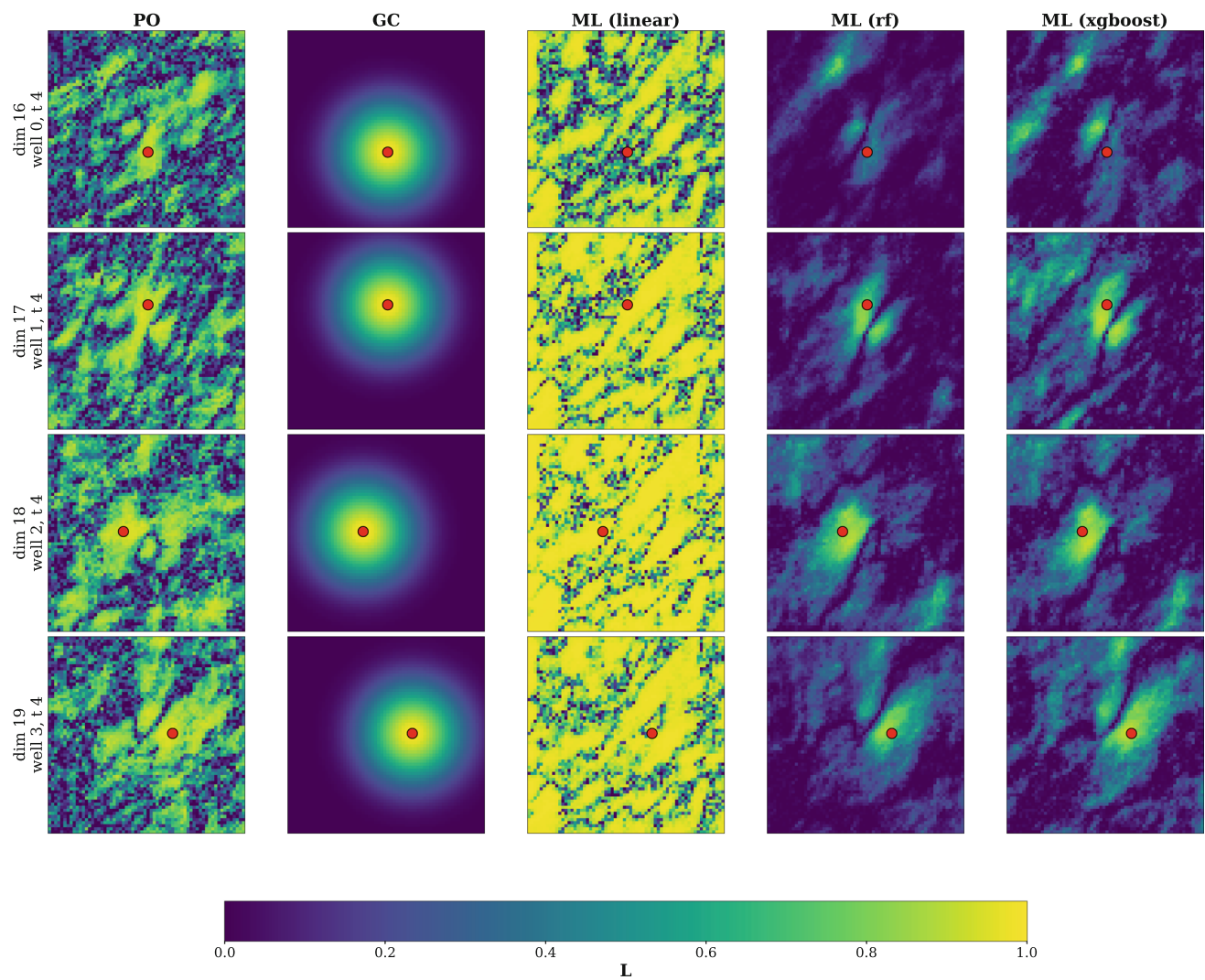


Fig. 21 Localization map for $N_e = 100$

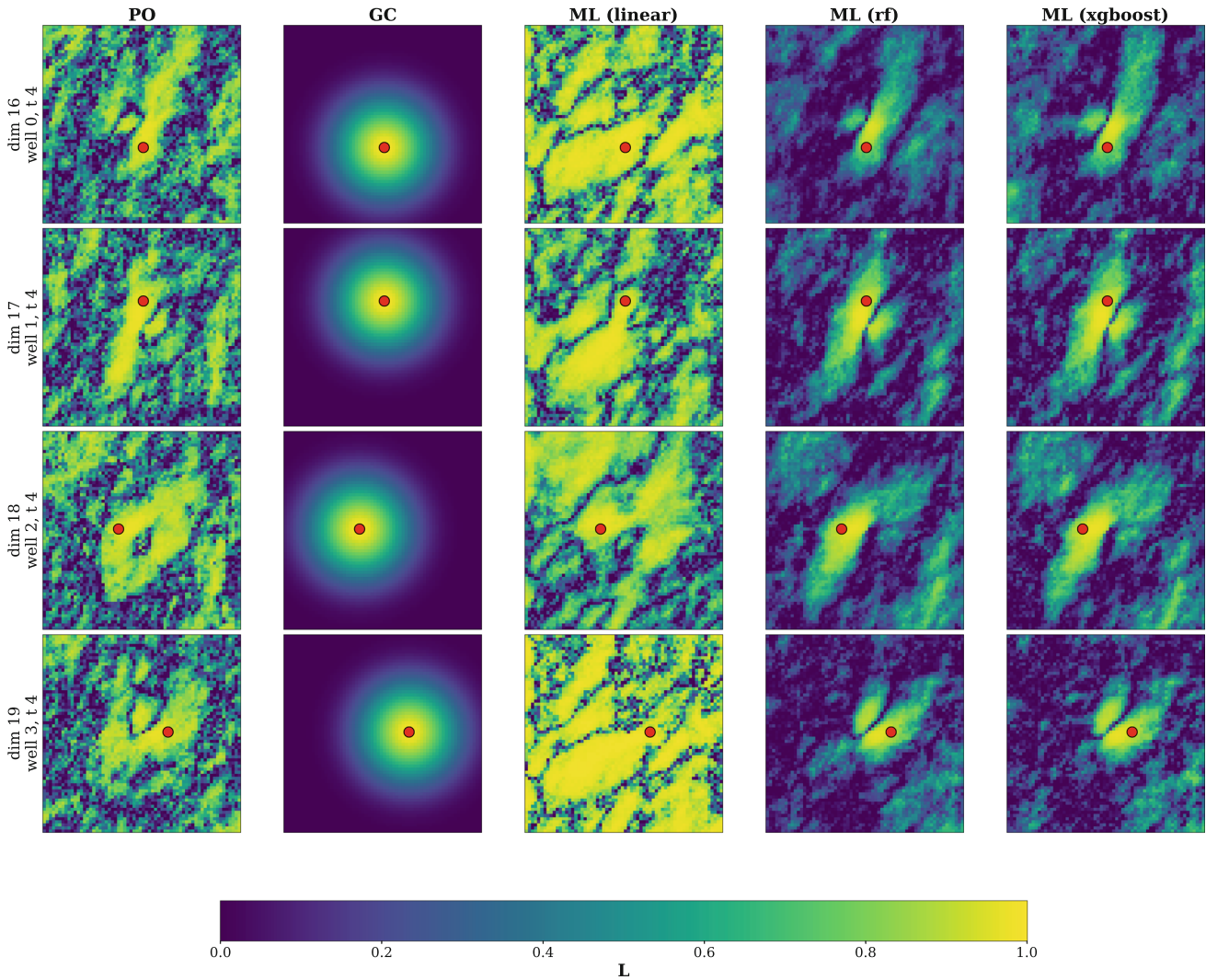


Fig. 22 Localization map for $N_e = 200$

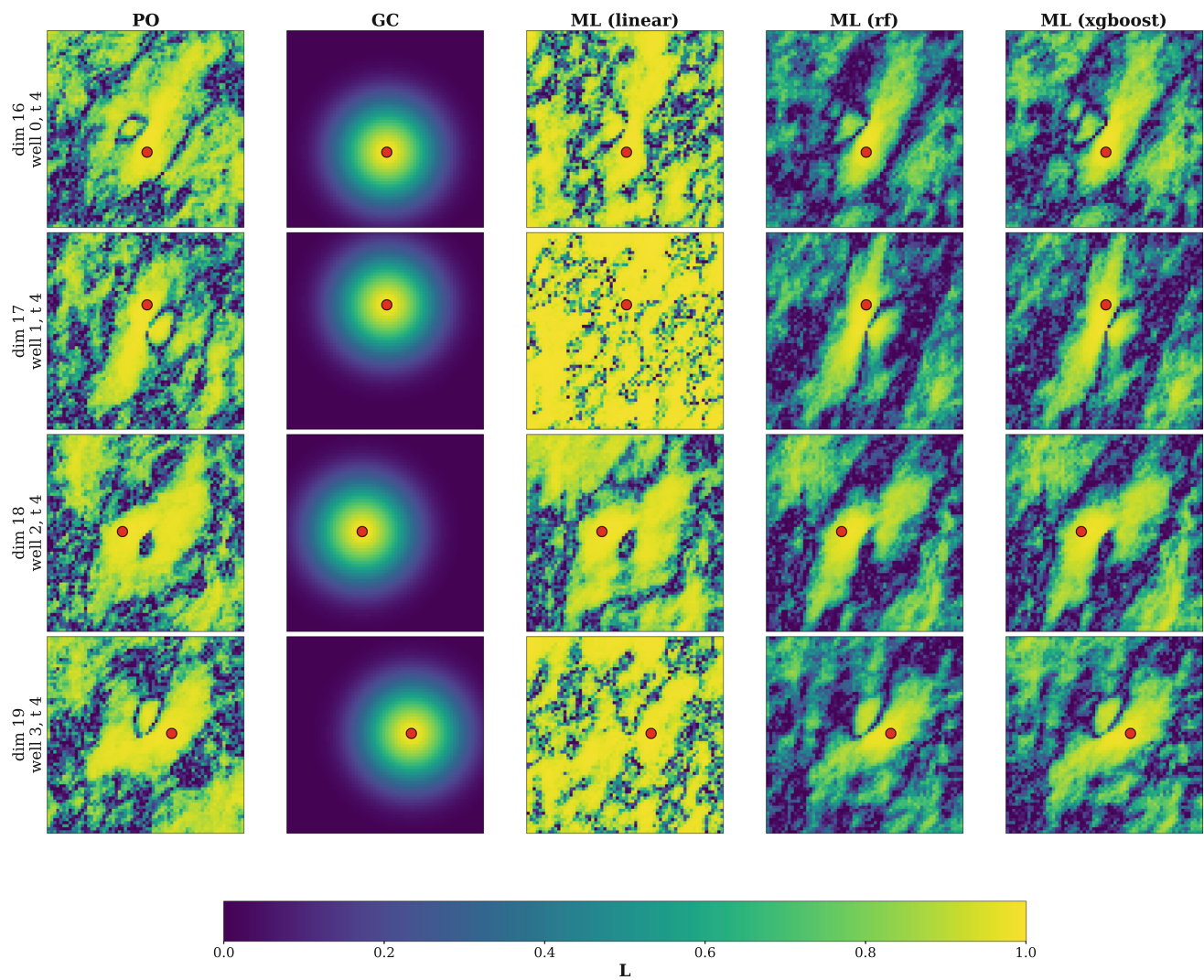


Fig. 23 Localization map for $N_e = 500$

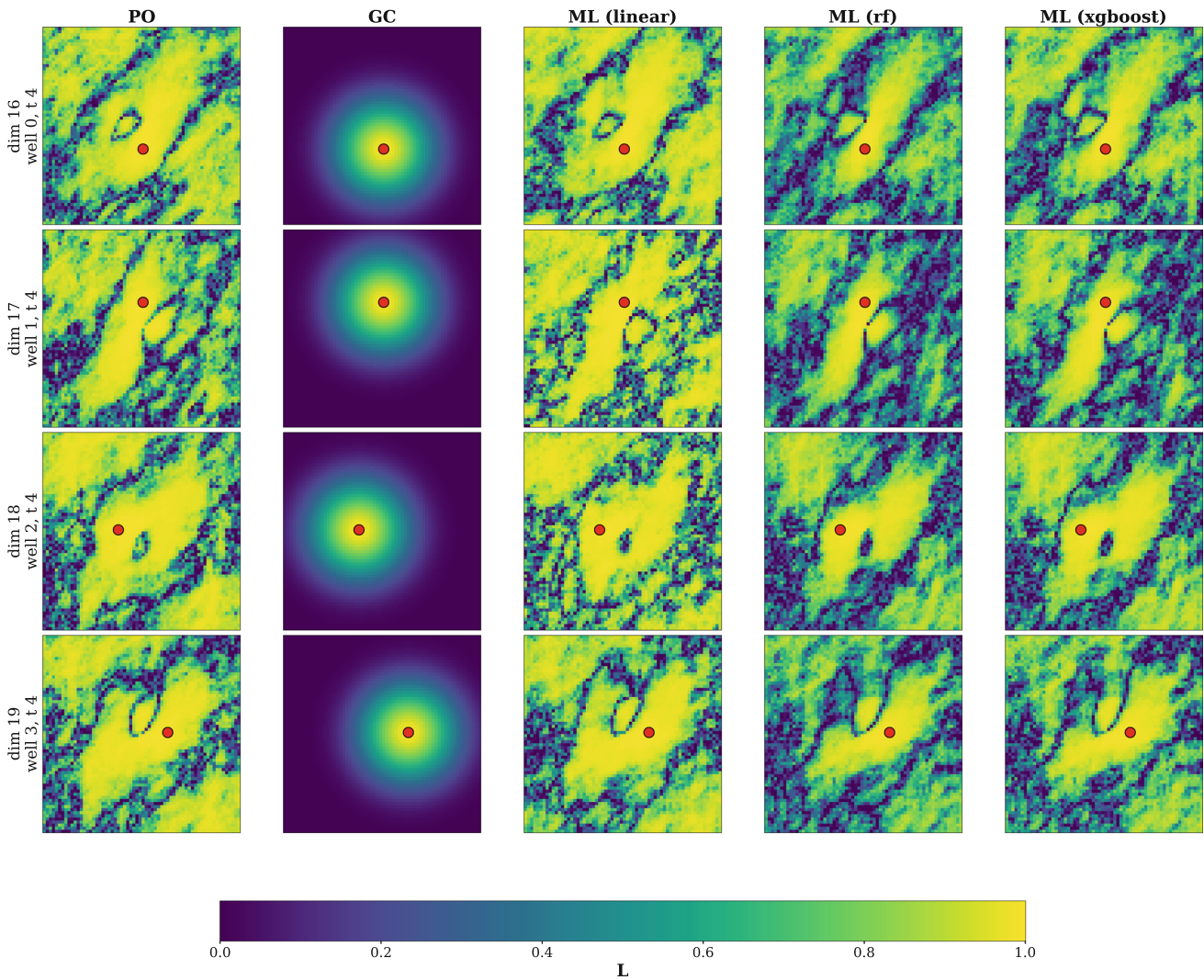


Fig. 24 Localization map for $N_e = 1000$

Author Contributions G.S.S. and N.T.M. contributed to the study conception and design, methodology, software development, validation, data curation, formal analysis, and writing of the original draft. G.S.S. further contributed to the investigation and visualization. V.L.S.S. contributed to the conceptualization, methodology, validation, and review and editing of the manuscript. A.A.E. contributed to the methodology, investigation, and review and editing. D.V. contributed to the conceptualization, supervision, and review and editing. F.C.V. was responsible for the conceptualization, methodology, supervision, project administration, funding acquisition, and review and editing. All authors read and approved the final manuscript.

Funding The authors are grateful to Petróleo Brasileiro S.A. (Petrobras) for sponsoring the doctoral research of Gabriel Serrão Seabra.

Data Availability The source code for all experiments is openly available at <https://github.com/gabrielserrao/DA-with-generative-ml>. The dataset used in this study is archived at <https://doi.org/10.4121/a8ad7808-b923-4335-ba7a-898c8c1232be>.

Declarations

Competing Interests The authors declare no competing interests.

Ethics and Consent to Participate Not applicable.

Use of Generative AI During manuscript preparation the authors employed OpenAI's ChatGPT strictly to improve wording and clarity, and Grammarly® for grammar checking. After using these tools, the authors reviewed and edited all suggestions to ensure accuracy and originality. The authors take full responsibility for the content of this publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Bui, M., Adjiman, C.S., Bardow, A., Anthony, E.J., Boston, A., Brown, S., Fennell, P.S., Fuss, S., Galindo, A., Hackett, L.A., Hallett, J.P., Herzog, H.J., Jackson, G., Kemper, J., Krevor, S., Maitland, G.C., Matuszewski, M., Metcalfe, I.S., Petit, C., Puxty, G., Reimer, J., Reiner, D.M., Rubin, E.S., Scott, S.A., Shah, N., Smit, B., Martin Trusler, J.P., Webley, P., Wilcox, J., Mac Dowell, N.: Carbon capture and storage (CCS): the way forward. *Energ. Environm. Sci.* **11**(5), 1062–1176 (2018). ISSN 1754-5692, 1754-5706. <https://doi.org/10.1039/C7EE02342A>. URL <https://xlink.rsc.org/?DOI=C7EE02342A>
2. Ringrose, P.: How to store CO₂ underground. SpringerBriefs in Earth Sciences Ser. Springer International Publishing AG, Cham (2020). ISBN 978-3-030-33113-9
3. Furre, A.-K., Eiken, O., Alnes, H., Vevatne, J.N., Kiær, A.F.: 20 years of monitoring CO₂-injection at Sleipner. *Energ. Proced.* **114**, 3916–3926 (2017). ISSN 18766102. <https://doi.org/10.1016/j.egypro.2017.03.1523>. URL <https://linkinghub.elsevier.com/retrieve/pii/S1876610217317174>
4. Chadwick, R.A., British Geological Survey (eds.): Best practice for the storage of CO₂ in saline aquifers: observations and guidelines from the SACS and CO2STORE projects. Number no. 14 in British Geological Survey occasional publication. British Geological Survey, Keyworth, Nottingham (2008). ISBN 978-0-85272-610-5
5. Pyrcz, M., Deutsch, C.V.: Geostatistical reservoir modeling, second edn. Oxford University Press, Oxford, New York (2014). ISBN 978-0-19-973144-2
6. Mariethoz, G., Caers, J.: Multiple-point geostatistics: Stochastic modeling with training images. Wiley, 1 edn (2014). ISBN 978-1-118-66275-5 978-1-118-66295-3. <https://doi.org/10.1002/9781118662953>. <https://onlinelibrary.wiley.com/doi/book/10.1002/9781118662953>
7. Strebelle, S.: Conditional simulation of complex geological structures using multiple-point statistics. *Math. Geol.* **34**(1), 1–21 (2002). ISSN 0882-8121, 1573-8868. <https://doi.org/10.1023/A:1014009426274>. <https://link.springer.com/10.1023/A:1014009426274>

8. Deutsch C.V., Tran T.T.: FLUVSIM: a program for object-based stochastic modeling of fluvial depositional systems. *Comput. Geosci.* **28**(4), 525–535 (2002). ISSN 00983004. [https://doi.org/10.1016/S0098-3004\(01\)00075-9](https://doi.org/10.1016/S0098-3004(01)00075-9). <https://linkinghub.elsevier.com/retrieve/pii/S0098300401000759>
9. Oliver, D.S., Reynolds, A.C., Liu, N.: Inverse theory for petroleum reservoir characterization and history matching. Cambridge University Press, 1 edn (2008). ISBN 978-0-521-88151-7 978-0-511-53564-2. <https://doi.org/10.1017/CBO9780511535642>. <https://www.cambridge.org/core/product/identifier/9780511535642/type/book>
10. Evensen, G., Vossepoel, F.C., Van Leeuwen, P.J.: Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem. Springer Textbooks in Earth Sciences, Geography and Environment. Springer International Publishing, Cham (2022). ISBN 978-3-030-96708-6 978-3-030-96709-3. <https://doi.org/10.1007/978-3-030-96709-3>. <https://link.springer.com/10.1007/978-3-030-96709-3>
11. Tarantola, A.: Inverse problem theory and methods for model parameter estimation. Soc. Industr. Appl. Math. (2005). ISBN 978-0-89871-572-9 978-0-89871-792-1. <https://doi.org/10.1137/1.9780898717921>. <http://epubs.siam.org/doi/book/10.1137/1.9780898717921>
12. Evensen, G.: Data assimilation: The ensemble Kalman filter. Springer, Berlin Heidelberg, Berlin, Heidelberg (2009). ISBN 978-3-642-03710-8 978-3-642-03711-5. <https://doi.org/10.1007/978-3-642-03711-5>. <https://link.springer.com/10.1007/978-3-642-03711-5>
13. Aanonsen, S.I., Nævdal, G., Oliver, D.S., Reynolds, A.C., Vallès, B.: The ensemble Kalman filter in reservoir engineering—a review. *SPE J.* **14**(03), 393–412 (2009). ISSN 1086-055X, 1930-0220. <https://doi.org/10.2118/117274-PA>. <https://onepetro.org/SJ/article/14/03/393/192079/The-Ensemble-Kalman-Filter-in-Reservoir>
14. Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.: Oceans* **99**(C5), 10143–10162 (1994). ISSN 0148-0227. <https://doi.org/10.1029/94JC00572>. URL <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/94JC00572>
15. Emerick, A.A., Reynolds, A.C.: Ensemble smoother with multiple data assimilation. *Comput. Geosci.* **55**, 3–15 (2013). ISSN 00983004. <https://doi.org/10.1016/j.cageo.2012.03.011>. <https://linkinghub.elsevier.com/retrieve/pii/S0098300412000994>
16. Chen, Y., Oliver, D.S.: Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Math. Geosci.* **44**(1), 1–26 (2012). ISSN 1874-8961, 1874-8953. <https://doi.org/10.1007/s11004-011-9376-z>. <http://link.springer.com/10.1007/s11004-011-9376-z>
17. Houtekamer, P.L., Zhang, F.: Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Rev.* **144**(12), 4489–4532 (2016). ISSN 0027-0644, 1520-0493. <https://doi.org/10.1175/MWR-D-15-0440.1>. <https://journals.ametsoc.org/view/journals/mwre/144/12/mwr-d-15-0440.1.xml>
18. Oliver, D.S., Chen, Y.: Recent progress on reservoir history matching: a review. *Computat. Geosci.* **15**(1), 185–221 (2011). ISSN 1420-0597, 1573-1499. <https://doi.org/10.1007/s10596-010-9194-2>. <http://link.springer.com/10.1007/s10596-010-9194-2>
19. Hamill, T.M., Whitaker, J.S., Snyder, C.: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Rev.* **129**(11), 2776–2790 (2001). ISSN 0027-0644, 1520-0493. [https://doi.org/10.1175/1520-0493\(2001\)129<2776:DDFOBE>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2). [http://journals.ametsoc.org/doi/10.1175/1520-0493\(2001\)129<2776:DDFOBE>2.0.CO;2](http://journals.ametsoc.org/doi/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2)
20. Gaspari, G., Cohn, S.E.: Construction of correlation functions in two and three dimensions. *Quart. J. Royal Meteorol. Soc.* **125**(554), 723–757 (1999). ISSN 0035-9009, 1477-870X. <https://doi.org/10.1002/qj.49712555417>. <https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.49712555417>
21. Rukhin, A.L.: Estimating common vector parameters in interlaboratory studies. *J. Multivar. Anal.* **98**(3), 435–454 (2007). ISSN 0047-259X. <https://doi.org/10.1016/j.jmva.2006.09.005>
22. Emerick, A.A., Reynolds, A.C.: History matching a field case using the ensemble Kalman filter with covariance localization. *SPE Reserv. Eval. Eng.* **14**(04), 443–452 (2011). ISSN 1094-6470, 1930-0212. <https://doi.org/10.2118/141216-PA>. <https://onepetro.org/REE/article/14/04/443/198210/History-Matching-a-Field-Case-Using-the-Ensemble>
23. Lacerda, A., Luo, X., Emerick, A.A.: Comparison of adaptive localization methods for updating petrophysical properties and areal parameters. *J. Petroleum Sci. Eng.* **183**, 106362 (2019). <https://doi.org/10.1016/j.petrol.2019.106362>
24. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations (2020). [arxiv:2011.13456](https://arxiv.org/abs/2011.13456)
25. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution (2019). [arxiv:1907.05600](https://arxiv.org/abs/1907.05600)
26. Houtekamer, P.L., Mitchell, H.L.: A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Rev.* **129**(1), 123–137 (2001). ISSN 0027-0644, 1520-0493. [https://doi.org/10.1175/1520-0493\(2001\)129<0123:ASEKFF>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2). [http://journals.ametsoc.org/doi/10.1175/1520-0493\(2001\)129<0123:ASEKFF>2.0.CO;2](http://journals.ametsoc.org/doi/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2)
27. Silva, V.L.S., Seabra, G.S., Emerick, A.A.: Mitigating loss of variance in ensemble data assimilation: machine learning-based and distance-free localization (2025). [arxiv:2506.13362](https://arxiv.org/abs/2506.13362)
28. 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. In: SPE Reservoir Simulation Conference, page D031S012R004, Galveston, Texas, USA (2025). SPE. <https://doi.org/10.2118/223908-MS>. <https://onepetro.org/spersc/proceedings/25RSC/25RSC/D031S012R004/646742>

29. Lacerda, J.M., Emerick, A.A., Pires, A.P.: Using a machine learning proxy for localization in ensemble data assimilation. *Computat. Geosci.* **25**(3), 931–944 (2021). ISSN 1420-0597, 1573-1499. <https://doi.org/10.1007/s10596-020-10031-0>. <https://link.springer.com/10.1007/s10596-020-10031-0>
30. Brajard, J., Carrassi, A., Bocquet, M., Bertino, L.: Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *J. Computat. Sci.* **44**, 101171 (2020). ISSN 18777503. <https://doi.org/10.1016/j.jocs.2020.101171>. <https://linkinghub.elsevier.com/retrieve/pii/S1877750320304725>
31. Bocquet, M., Brajard, J., Carrassi, A., Bertino, L.: Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *Foundat. Data Sci.* **2**(1), 55–80 (2020). <https://doi.org/10.3934/fods.2020004>
32. Buizza, C., Quilodr  n Casas, C., Nadler, P., Mack, J., Marrone, S., Titus, Z., Le Cornec, C., Heylen, E., Dur, T., Baca Ruiz, L., Heaney, C., D  az Lopez, J.A., Kumar, K.S.S., Arcucci R.: Data Learning: Integrating Data Assimilation and Machine Learning. *J. Computat. Sci.* **58**, 101525 (2022). ISSN 18777503. <https://doi.org/10.1016/j.jocs.2021.101525>. <https://linkinghub.elsevier.com/retrieve/pii/S1877750321001861>
33. Cheng, S., Quilodr  n-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., Xiao, D., Janjic, T., Ding, W., Guo, Y., Carrassi, A., Bocquet, M., Arcucci, R.: Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. *IEEE/CAA J. Automat. Sinica* **10**(6), 1361–1387 (2023). ISSN 2329-9266, 2329-9274. <https://doi.org/10.1109/JAS.2023.123537>. <https://ieeexplore.ieee.org/document/10141545/>
34. Farchi, A., Laloyaux, P., Bonavita, M., Bocquet, M.: Using machine learning to correct model error in data assimilation and forecast applications. *Quart. J. Royal Meteorol. Soc.* **147**(739), 3067–3084 (2021). ISSN 0035-9009, 1477-870X. <https://doi.org/10.1002/qj.4116>. <https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.4116>
35. Farchi, A., Bocquet, M., Laloyaux, P., Bonavita, M., Malartic, Q.: A comparison of combined data assimilation and machine learning methods for offline and online model error correction. *J. Comput. Sci.* **55**, 101468 (2021). ISSN 18777503. <https://doi.org/10.1016/j.jocs.2021.101468>. <https://linkinghub.elsevier.com/retrieve/pii/S1877750321001435>
36. Brajard, J., Carrassi, A., Bocquet, M., Bertino, L.: Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophic. Trans. Royal Soc. A: Math. Phys. Eng. Sci.* **379**(2194), 20200086 (2021). ISSN 1471-2962. <https://doi.org/10.1098/rsta.2020.0086>. <http://dx.doi.org/10.1098/rsta.2020.0086>
37. Wikner, A., Pathak, J., Hunt, B.R., Szunyogh, I., Girvan, M., Ott, E.: Using data assimilation to train a hybrid forecast system that combines machine-learning and knowledge-based components. *Chaos: Interdiscipl. J. Nonlin. Sci.* **31**(5), 053114 (2021). ISSN 1054-1500, 1089-7682. <https://doi.org/10.1063/5.0048050>. <https://pubs.aip.org/cha/article/31/5/053114/569378/Using-data-assimilation-to-train-a-hybrid-forecast>
38. Arcucci, R., Zhu, J., Hu, S., Guo, Y.-K.: Deep data assimilation: integrating deep learning with data assimilation. *Appl. Sci. (Switzerland)* **11**(3), 1–21 (2021). ISSN 2076-3417. <https://doi.org/10.3390/app11031114>
39. Di Federico, G., Durlofsky, L.J.: Latent diffusion models for parameterization and data assimilation of facies-based geomodels (2024). [arxiv:2406.14815](https://arxiv.org/abs/2406.14815)
40. Fan, M., Liu, Y., Lu, D., Wang, H., Zhang, G.: A novel conditional generative model for efficient ensemble forecasts of state variables in large-scale geological carbon storage. *J. Hydrol.* **648**, 132323 (2025). ISSN 0022-1694. <https://doi.org/10.1016/j.jhydrol.2024.132323>
41. Gahlot, A.P., Orozco, R., Yin, Z., Bruer, G., Herrmann, F.J.: An uncertainty-aware digital shadow for underground multimodal CO₂ storage monitoring. *Geophys. J. Intern.* **242**(1), ggaf176 (2025). ISSN 0956-540X, 1365-246X. <https://doi.org/10.1093/gji/ggaf176>. <https://academic.oup.com/gji/article/doi/10.1093/gji/ggaf176/8142541>
42. Teng, W., Durlofsky, L.J.: Likelihood-free inference and hierarchical data assimilation for geological carbon storage. *Adv. Water Res.* **201**, 104961 (2025). ISSN 03091708. <https://doi.org/10.1016/j.advwatres.2025.104961>. <https://linkinghub.elsevier.com/retrieve/pii/S0309170825000752>
43. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *Proceedings of the 32nd international conference on machine learning*, pp. 2256–2265. PMLR (2015). <https://proceedings.mlr.press/v37/sohl-dickstein15.html>
44. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020). [arxiv:2006.11239](https://arxiv.org/abs/2006.11239)
45. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis (2021). [arxiv:2105.05233](https://arxiv.org/abs/2105.05233)
46. Vincent, P.: A connection between score matching and denoising autoencoders. *Neural Computat.* **23**(7), 1661–1674 (2011). ISSN 0899-7667, 1530-888X. https://doi.org/10.1162/NECO_a_00142. <https://direct.mit.edu/neco/article/23/7/1661-1674/7677>
47. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems (2022). [arxiv:2209.14687](https://arxiv.org/abs/2209.14687)
48. Vossepoul, F.C., Evensen, G., Van Leeuwen, P.J.: Adaptive correlation- and distance-based localization for iterative ensemble smoothers in a coupled nonlinear multiscale model. *Monthly Weather Rev.* (2025). ISSN 0027-0644, 1520-0493. <https://doi.org/10.1175/MWR-D-24-0269.1>. <https://journals.ametsoc.org/view/journals/mwre/aop/MWR-D-24-0269.1/MWR-D-24-0269.1.xml>

49. Voskov, D.V.: Operator-based linearization approach for modeling of multiphase multi-component flow in porous media. *J. Computat. Phys.* **337**, 275–288 (2017). ISSN 00219991. <https://doi.org/10.1016/j.jcp.2017.02.041>. <https://linkinghub.elsevier.com/retrieve/pii/S0021999117301444>
50. Khait, M., Voskov, D.V.: Operator-based linearization for general purpose reservoir simulation. *J. Petroleum Sci. Eng.* **157**, 990–998 (2017). ISSN 09204105. <https://doi.org/10.1016/j.petrol.2017.08.009>. <https://linkinghub.elsevier.com/retrieve/pii/S0920410516310907>
51. Wapperom, M., Tian, X., Novikov, A., Voskov, D.: FluidFlow benchmark: lessons learned from the perspective of subsurface simulation. *Transport Porous Med.* **151**(5), 1033–1052 (2024). ISSN 0169-3913, 1573-1634. <https://doi.org/10.1007/s11242-023-01984-8>. <https://link.springer.com/10.1007/s11242-023-01984-8>
52. Chen, Y., Voskov, D.: Optimization of CO₂ injection using multi-scale reconstruction of composition transport. *Computat. Geosci.* **24**(2), 819–835 (2020). ISSN 1420-0597, 1573-1499. <https://doi.org/10.1007/s10596-019-09841-8>. <http://link.springer.com/10.1007/s10596-019-09841-8>
53. Lyu, X., Voskov, D.: Advanced modeling of enhanced CO₂ dissolution trapping in saline aquifers. *Intern. J. Greenhouse Gas Contr.* **127**, 103907 (2023). ISSN 17505836. <https://doi.org/10.1016/j.ijggc.2023.103907>. <https://linkinghub.elsevier.com/retrieve/pii/S1750583623000774>
54. Brooks, R.H., Corey, A.T.: Hydraulic properties of porous media and their relation to drainage design. *Trans. ASAE* **7**(1), 0026–0028 (1964). ISSN 2151-0059. <https://doi.org/10.13031/2013.40684>. <http://elibrary.asabe.org/abstract.asp?JID=3&AID=40684&CID=1964&v=7&i=1&T=1>
55. Li, G., Jin, D., Yu, Q., Qi, M.: IB-TransUNet: Combining information bottleneck and Transformer for medical image segmentation. *J. King Saud Univ. - Comput. Inf. Sci.* **35**(3), 249–258 (2023). ISSN 13191578. <https://doi.org/10.1016/j.jksuci.2023.02.012>. <https://linkinghub.elsevier.com/retrieve/pii/S1319157823000411>
56. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11966–11976. New Orleans, LA, USA (2022). IEEE. ISBN 978-1-6654-6946-3. <https://doi.org/10.1109/CVPR52688.2022.01167>. <https://ieeexplore.ieee.org/document/9879745/>
57. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: self-attention with linear complexity (2020). [arxiv:2006.04768](https://arxiv.org/abs/2006.04768)
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
59. Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. *Inf. Fus.* **81**, 84–90 (2022). ISSN 15662535. <https://doi.org/10.1016/j.inffus.2021.11.011>. <https://linkinghub.elsevier.com/retrieve/pii/S1566253521002360>
60. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **35**(6), 7499–7519 (2024). ISSN 2162-237X, 2162-2388. <https://doi.org/10.1109/TNNLS.2022.3229161>. <https://ieeexplore.ieee.org/document/9998482/>
61. Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why do tree-based models still outperform deep learning on tabular data? (2022). [arxiv:2207.08815](https://arxiv.org/abs/2207.08815)
62. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001). ISSN 0885-6125, 1573-0565. <https://doi.org/10.1023/A:1010933404324>. URL <https://link.springer.com/10.1023/A:1010933404324>
63. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 785–794. San Francisco California USA (2016). ACM. ISBN 978-1-4503-4232-2. <https://doi.org/10.1145/2939672.2939785>. <https://dl.acm.org/doi/10.1145/2939672.2939785>
64. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011). ISSN 1532-4435
65. Evensen, G., Raanes, P.N., Stordal, A.S., Hove, J.: Efficient implementation of an iterative ensemble smoother for data assimilation and reservoir history matching. *Front. Appl. Math. Stat.* **5**, 47 (2019). ISSN 2297-4687. <https://doi.org/10.3389/fams.2019.00047>. <https://www.frontiersin.org/article/10.3389/fams.2019.00047/full>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Gabriel Serrão Seabra^{1,2} · Nikolaj T. Mücke^{1,3} · Vinicius Luiz Santos Silva^{2,4} · Alexandre A. Emerick² · Denis Voskov^{1,5} · Femke C. Vossepoel¹

✉ Gabriel Serrão Seabra
g.serraoseabra@tudelft.nl

Nikolaj T. Mücke
nikolaj.mucke@cw.nl

Vinicius Luiz Santos Silva
v.santos-silva19@alumni.imperial.ac.uk

Alexandre A. Emerick
alexandre.emerick@petrobras.com.br

Denis Voskov
voskov@tudelft.nl

Femke C. Vossepoel
f.c.vossepoel@tudelft.nl

¹ Faculty of Civil Engineering and Geosciences, TU Delft, Stevinweg 1, 2628 CN Delft, Netherlands

² Petroleo Brasileiro S.A. (Petrobras), Rio de Janeiro, Brazil

³ Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, Netherlands

⁴ Novel Reservoir Modelling and Simulation Group, Imperial College London, London, UK

⁵ Department of Energy Resources Engineering, Stanford University, Stanford, CA, USA