Correlation-Based Localization for Subsurface Applications

Alexandre A. Emerick & Vinicius Luiz S. Silva

Petrobras Research, Development, and Innovation Center

21th International EnKF Workshop (2026), Rosendal, Norway

Introduction

- Subsurface data assimilation typically requires localization to mitigate spurious correlations

Introduction

- Subsurface data assimilation typically requires localization to mitigate spurious correlations
- Distance-based localization
  - Simple and widely effective in practice
  - Requires empirical tuning of localization radii
  - Spatial distance may not reflect correlations induced by flow dynamics
  - Not well-defined for non-spatial parameters and data

Introduction

- Subsurface data assimilation typically requires localization to mitigate spurious correlations
- Distance-based localization
  - Simple and widely effective in practice
  - Requires empirical tuning of localization radii
  - Spatial distance may not reflect correlations induced by flow dynamics
  - Not well-defined for non-spatial parameters and data
- Correlation-based localization
  - Does not rely on spatial distance
  - Exploits the statistical reliability of sample correlations
  - Still less robust than distance-based approaches in practice

Correlation-based localization can sometimes feel like chasing unreliable correlations

Objectives

Emerick, A.A. & Silva, V.L.S.: *Statistical Tapers for Correlation-Based Localization in Ensemble Data Assimilation*, `arXiv:2605.29922v1 [stat.ME]`, 2026[1]

Objectives

Emerick, A.A. & Silva, V.L.S.: *Statistical Tapers for Correlation-Based Localization in Ensemble Data Assimilation*, `arXiv:2605.29922v1 [stat.ME]`, 2026[1]

1. Propose statistically motivated correlation-based tapers and compare them with existing formulations

Objectives

Emerick, A.A. & Silva, V.L.S.: *Statistical Tapers for Correlation-Based Localization in Ensemble Data Assimilation*, `arXiv:2605.29922v1 [stat.ME]`, 2026[1]

1. Propose statistically motivated correlation-based tapers and compare them with existing formulations
2. Evaluate correlation-based localization in problems with non-local parameter-data relationships

Objectives

Emerick, A.A. & Silva, V.L.S.: *Statistical Tapers for Correlation-Based Localization in Ensemble Data Assimilation*, `arXiv:2605.29922v1 [stat.ME]`, 2026[1]

1. Propose statistically motivated correlation-based tapers and compare them with existing formulations
2. Evaluate correlation-based localization in problems with non-local parameter-data relationships
3. Assess whether correlation-based localization can replace distance-based localization in subsurface data assimilation

Correlation-Based Tapers

- Let
  - ρ_{ij}: actual (unknown) correlation coefficient between i-th model parameter and j-th data point
  - $\tilde{\rho}_{ij}$: correlation coefficient estimated with N_e ensemble members

Correlation-Based Tapers

- Let
  - ρ_{ij}: actual (unknown) correlation coefficient between i-th model parameter and j-th data point
  - $\tilde{\rho}_{ij}$: correlation coefficient estimated with N_e ensemble members

- Define a taper function $r(\tilde{\rho}_{ij}) \in [0, 1]$ such that

$$\tilde{\rho}_{ij}^{\text{loc}} = r(\tilde{\rho}_{ij})\, \tilde{\rho}_{ij}$$

  is a better approximation of ρ_{ij}

Correlation-Based Tapers

- Let
  - ρ_{ij}: actual (unknown) correlation coefficient between i-th model parameter and j-th data point
  - $\tilde{\rho}_{ij}$: correlation coefficient estimated with N_e ensemble members
- Define a taper function $r(\tilde{\rho}_{ij}) \in [0, 1]$ such that

$$\tilde{\rho}_{ij}^{\text{loc}} = r(\tilde{\rho}_{ij})\,\tilde{\rho}_{ij}$$

  is a better approximation of ρ_{ij}
- Ideally, we reduce the effect of spurious correlations while preserving statistically meaningful parameter-data relationships

Mean Squared Error Taper

$$\arg\min_r \mathbb{E}\left[(r\tilde{\rho} - \rho)^2\right] \quad \Rightarrow \quad r^\star = \frac{\rho^2}{\rho^2 + \sigma^2} \approx \frac{\tilde{\rho}^2}{\tilde{\rho}^2 + \sigma^2}$$

where $\sigma \propto \frac{1}{\sqrt{N_e}}$

Mean Squared Error Taper

$$\arg\min_{r} \mathbb{E}\left[(r\tilde{\rho} - \rho)^2\right] \quad \Rightarrow \quad r^\star = \frac{\rho^2}{\rho^2 + \sigma^2} \approx \frac{\tilde{\rho}^2}{\tilde{\rho}^2 + \sigma^2}$$

where $\sigma \propto \frac{1}{\sqrt{N_e}}$

- Define the standardized correlation

$$t = \frac{|\tilde{\rho}|}{\sigma}$$

Mean Squared Error Taper

$$\arg\min_r \mathbb{E}\left[(r\tilde{\rho} - \rho)^2\right] \quad \Rightarrow \quad r^\star = \frac{\rho^2}{\rho^2 + \sigma^2} \approx \frac{\tilde{\rho}^2}{\tilde{\rho}^2 + \sigma^2}$$

where $\sigma \propto \frac{1}{\sqrt{N_e}}$

- Define the standardized correlation

$$t = \frac{|\tilde{\rho}|}{\sigma}$$

- Then

$$r(t) = \frac{t^2}{t^2 + 1}$$

- The MSE taper is too permissive in practice (fails to identify spurious correlations)

Power-Law Taper

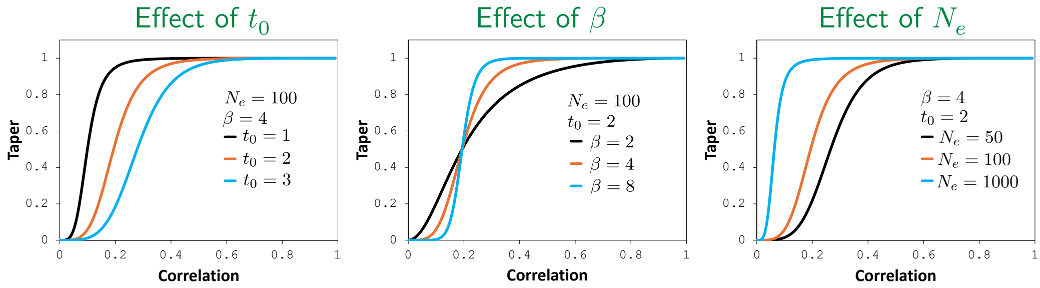
- Generalizes the MSE taper into a more flexible form:

$$r(t) = \frac{t^2}{t^2 + 1} \quad \Rightarrow \quad r(t) = \frac{t^\beta}{t^\beta + t_0^\beta}$$

- $\beta \geq 2$ controls the sharpness of the transition
- t_0 controls the location of the transition $r(t_0) = \frac{1}{2}$

Power-Law Taper

$$r(t) = \frac{t^{\beta}}{t^{\beta} + t_0^{\beta}}$$



Bayesian Tapering: Gaussian Prior

- Gaussian prior:

$$\rho \sim \mathcal{N}(0, \upsilon^2)$$

Bayesian Tapering: Gaussian Prior

- Gaussian prior:

$$\rho \sim \mathcal{N}(0, \upsilon^2)$$

- Likelihood model:

$$\tilde{\rho} \mid \rho \sim \mathcal{N}(\rho, \sigma^2)$$

Bayesian Tapering: Gaussian Prior

- Gaussian prior:
$$\rho \sim \mathcal{N}(0, \upsilon^2)$$
- Likelihood model:
$$\tilde{\rho} \mid \rho \sim \mathcal{N}(\rho, \sigma^2)$$
- Posterior mean:
$$\tilde{\rho}^{\text{loc}} = \mathbb{E}[\rho \mid \tilde{\rho}] = \frac{\upsilon^2}{\upsilon^2 + \sigma^2}\,\tilde{\rho}$$

Bayesian Tapering: Gaussian Prior

- Gaussian prior:
$$\rho \sim \mathcal{N}(0, \upsilon^2)$$
- Likelihood model:
$$\tilde{\rho} \mid \rho \sim \mathcal{N}(\rho, \sigma^2)$$
- Posterior mean:
$$\tilde{\rho}^{\text{loc}} = \mathbb{E}[\rho \mid \tilde{\rho}] = \frac{\upsilon^2}{\upsilon^2 + \sigma^2}\,\tilde{\rho}$$
- Let $\tau = \upsilon/\sigma$
$$r = \frac{\tau^2}{\tau^2 + 1}$$
- Similar to the MSE taper

Bayesian Tapering: Spike-and-Slab Prior

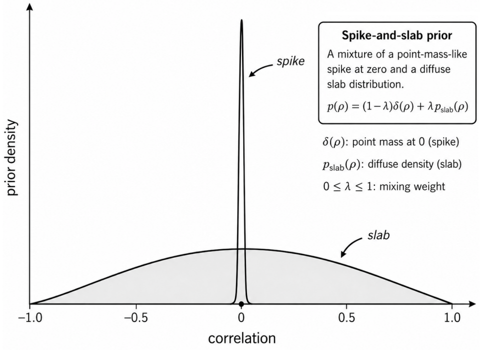
- Spike-and-slab prior:

$$p(\rho) = (1 - \lambda)\delta(\rho) + \lambda\, p_{\text{slab}}(\rho)$$

- Posterior:

$$p(\rho \mid \tilde{\rho}) = (1 - f(\tilde{\rho}))\delta(\rho) + f(\tilde{\rho}) p_{\text{slab}}(\rho \mid \tilde{\rho})$$

where $f(\tilde{\rho}) = P(\rho \neq 0 \mid \tilde{\rho})$ (probability that ρ belongs to the slab)



Bayesian Tapering: Spike-and-Slab Prior

- Posterior mean:

$$\tilde{\rho}^{\mathrm{loc}} = \mathbb{E}[\rho \mid \tilde{\rho}] = f(\tilde{\rho})\mathbb{E}_{\mathrm{slab}}[\rho \mid \tilde{\rho}]$$

Bayesian Tapering: Spike-and-Slab Prior

- Posterior mean:

$$\tilde{\rho}^{\text{loc}} = \mathbb{E}[\rho \mid \tilde{\rho}] = f(\tilde{\rho})\mathbb{E}_{\text{slab}}[\rho \mid \tilde{\rho}]$$

- Assuming a Gaussian slab $\rho \mid (\rho \neq 0) \sim \mathcal{N}(0, \upsilon^2)$:

$$\tilde{\rho}^{\text{loc}} = \mathbb{E}[\rho \mid \tilde{\rho}] = f(\tilde{\rho})\frac{\upsilon^2}{\upsilon^2 + \sigma^2}\tilde{\rho}$$

where

$$f(\tilde{\rho}) = \left[1 + \frac{1-\lambda}{\lambda}\sqrt{\frac{\sigma^2+\upsilon^2}{\sigma^2}}\exp\left(-\frac{\upsilon^2\tilde{\rho}^2}{2\sigma^2(\sigma^2+\upsilon^2)}\right)\right]^{-1}$$

Bayesian Tapering: Spike-and-Slab Prior

- Taper:

$$r(\tilde{\rho}) = f(\tilde{\rho})\frac{\upsilon^2}{\upsilon^2 + \sigma^2}$$

Bayesian Tapering: Spike-and-Slab Prior

- Taper:

$$r(\tilde{\rho}) = f(\tilde{\rho})\frac{\upsilon^2}{\upsilon^2 + \sigma^2}$$

- Detection component $f(\tilde{\rho})$: determines whether the correlation is likely to be genuine

Bayesian Tapering: Spike-and-Slab Prior

- Taper:

$$r(\tilde{\rho}) = f(\tilde{\rho})\frac{\upsilon^2}{\upsilon^2 + \sigma^2}$$

- Detection component $f(\tilde{\rho})$: determines whether the correlation is likely to be genuine
- Shrinkage component $\frac{\upsilon^2}{\upsilon^2+\sigma^2}$: controls the damping of nonzero correlations

Bayesian Tapering: Loigistic Taper

- Rewrite the taper in terms of t:

$$r(t) = \frac{1}{1 + \exp\left(-c\left(t^2 - t_0^2\right)\right)}$$

Bayesian Tapering: Loigistic Taper

- Rewrite the taper in terms of t:

$$r(t) = \frac{1}{1 + \exp(-c\,(t^2 - t_0^2))}$$

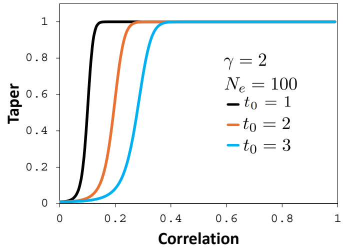
- Add an exponent parameter, γ, to control the shape of the transition:

$$r(t) = \frac{1}{1 + \exp(-c\,(t^\gamma - t_0^\gamma))}$$

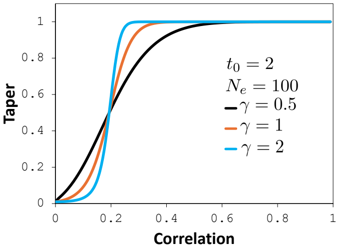
Bayesian Tapering: Logistic Taper

$$r(t) = \frac{1}{1 + \exp(-c\,(t^{\gamma} - t_0^{\gamma}))}$$

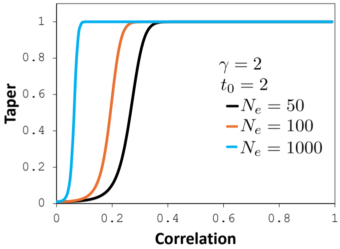
Effect of t_0



Effect of γ



Effect of N_e

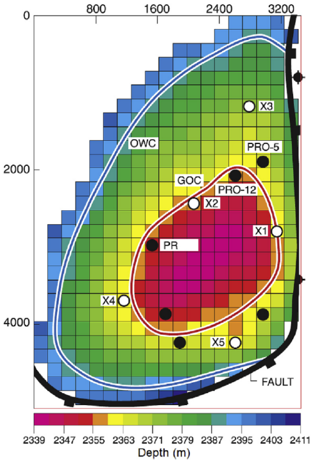


Summary of Tapers

Taper	Expression	Parameters	Reference
MSE	$r = \frac{t^2}{t^2+1}$	–	
Power-law	$r = \frac{t^\beta}{t^\beta+t_0^\beta}$	$\beta \geq 2,\ t_0 \in [1,3]$	
Logistic	$r = \frac{1}{1+\exp[-c(t^\gamma-t_0^\gamma)]}$	$\gamma \in [0.5,2],\ t_0 \in [1,3]$	
Discrepancy	$r = \max(0,\ 1-\frac{\eta}{t})$	$\eta \in [0.1,1]$	
Correlation Gaspari-Cohn (CGC)	$r = f_{\mathrm{GC}}\left(\frac{1-\lvert\widetilde{\rho}\rvert}{1-\theta}\right)$	$\theta \approx 1/\sqrt{N_e}$	Luo and Bhakta [4]
Pseudo-optimal (PO)	$r = \frac{\widetilde{\rho}^2}{\widetilde{\rho}^2+(1+\widetilde{\rho}^2)/N_e}$	–	Furrer and Bengtsson [3]
Modified PO (MPO)	$r = \max\left(0,\ \frac{N_e-\frac{1}{\widetilde{\rho}^2}}{N_e+1}\right)$	–	Ranazzi; Luo, and Sampaio [5]

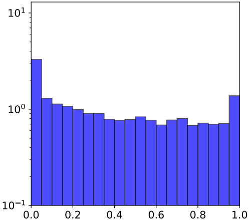
Test Case 1: Scalar Parameters

- PUNQ-S3 problem[2], considering only scalar model parameters
- Goal: Evaluate correlation-based localization in a setting where distance-based localization is not applicable
- 15 scalar parameters with different levels of influence on the reservoir dynamic response
- 5 dummy parameters intentionally defined to be uncorrelated with the simulation responses
- Data assimilation repeated 10 times using ensembles with $N_e = 100$ realizations

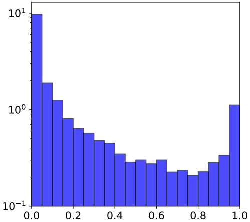


Test Case 1: Taper Values

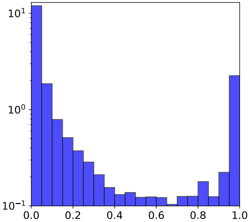
MSE



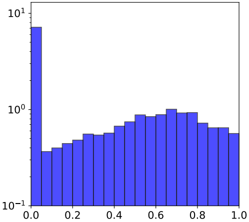
Power-law $(\beta = 3,\, t_0 = 2)$



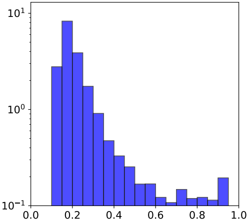
Logistic $(\gamma = 1.5,\, t_0 = 2)$



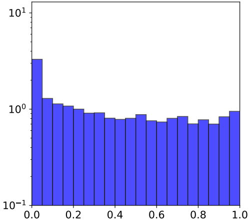
Discrepancy $(\eta = 0.5)$



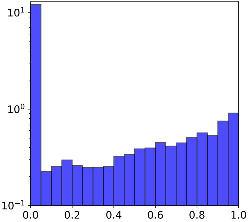
CGC $(\theta = \sigma)$



PO



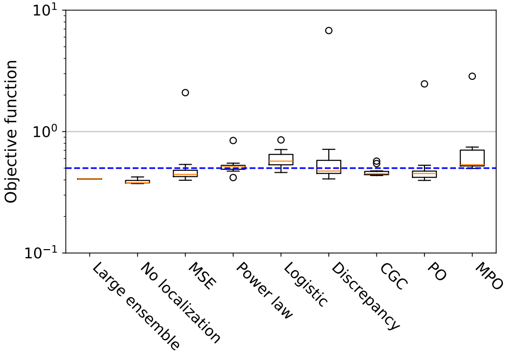
MPO



Test Case 1: Results

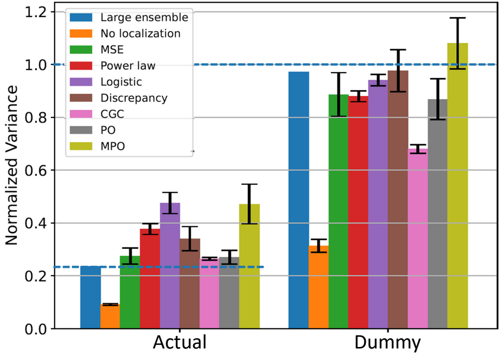
Average objective function

$$\overline{\mathcal{O}_d(\mathbf{m})} = \frac{1}{N_e}\sum_{k=1}^{N_e}\left[\frac{1}{2N_d}\sum_{j=1}^{N_d}\left(\frac{d_{\mathrm{obs},j} - g_j(\mathbf{m}_k)}{\sigma_{e,j}}\right)^2\right]$$



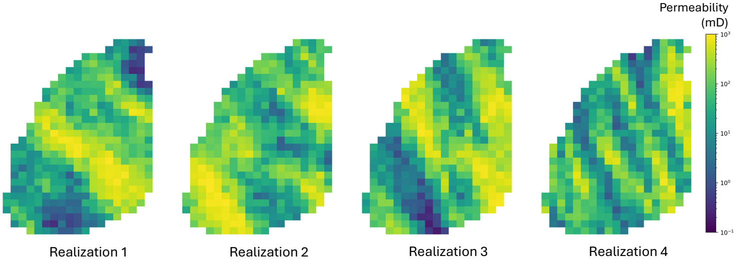
Normalized variance

$$\mathrm{NV} = \frac{1}{N_m}\sum_{i=1}^{N_m}\frac{\mathrm{var}\left[m_{\mathrm{post},i}\right]}{\mathrm{var}\left[m_{\mathrm{prior},i}\right]}$$



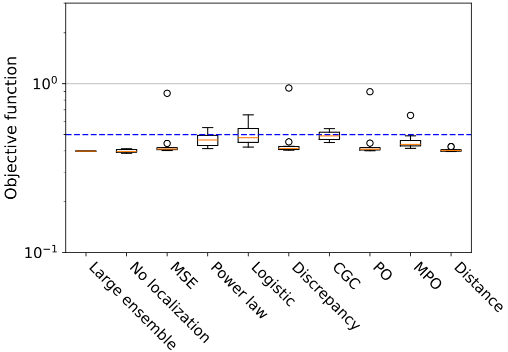
Test Case 2: Grid Parameters

- PUNQ-S3 problem with grid-based model parameters: porosity, horizontal permeability, and vertical permeability for each active gridblock (Total $N_m = 5{,}283$ parameters)
- Data assimilation repeated 10 times using ensembles with $N_e = 100$ realizations
- Goal: Evaluate correlation-based localization in a situation where distance-based is applicable, but not straightforward to define optimally

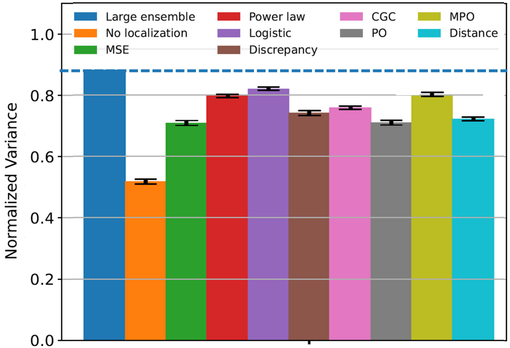


Test Case 2: Results

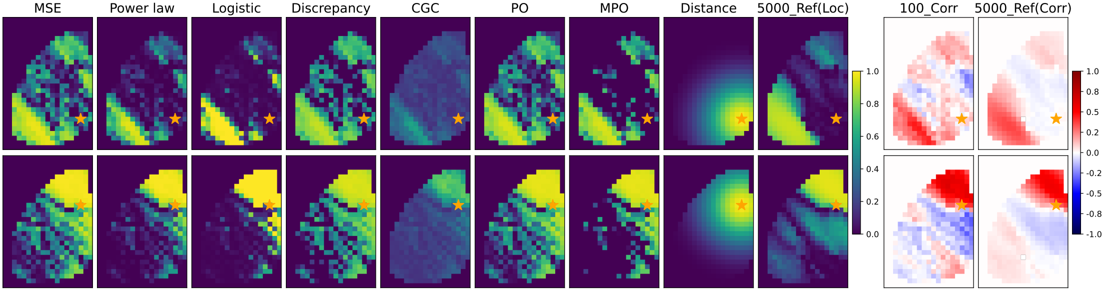
Average objective function



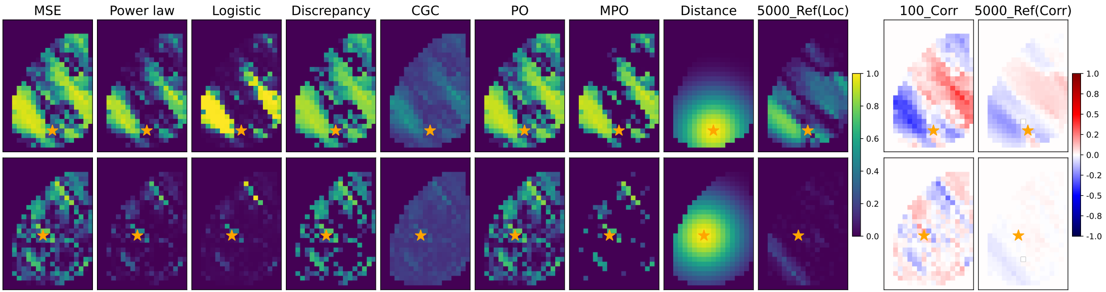
Normalized variance



Test Case 2: Tapers (Pressure)

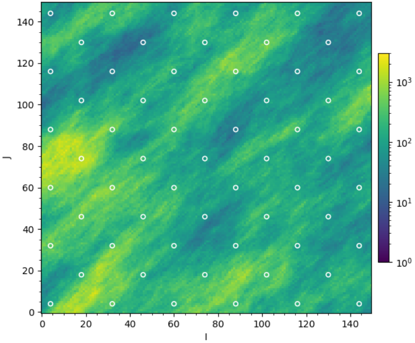


Test Case 2: Tapers (Water cut)



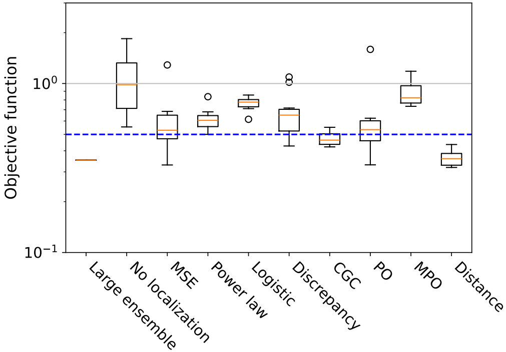
Test Case 3: Localized Updates

- 2D reservoir model with a 150×150 grid and five-spot well patterns
- Objective: evaluate correlation-based localization when updates are expected to be spatially localized
  - Relevant updates are mainly associated with flow paths connecting injectors and producers
  - Favorable case for distance-based localization
- 10 repeated data assimilation runs with $N_e = 200$

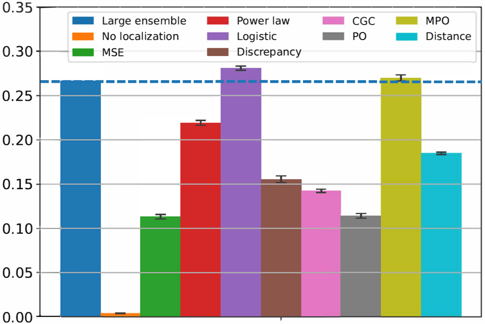


Test Case 3: Results

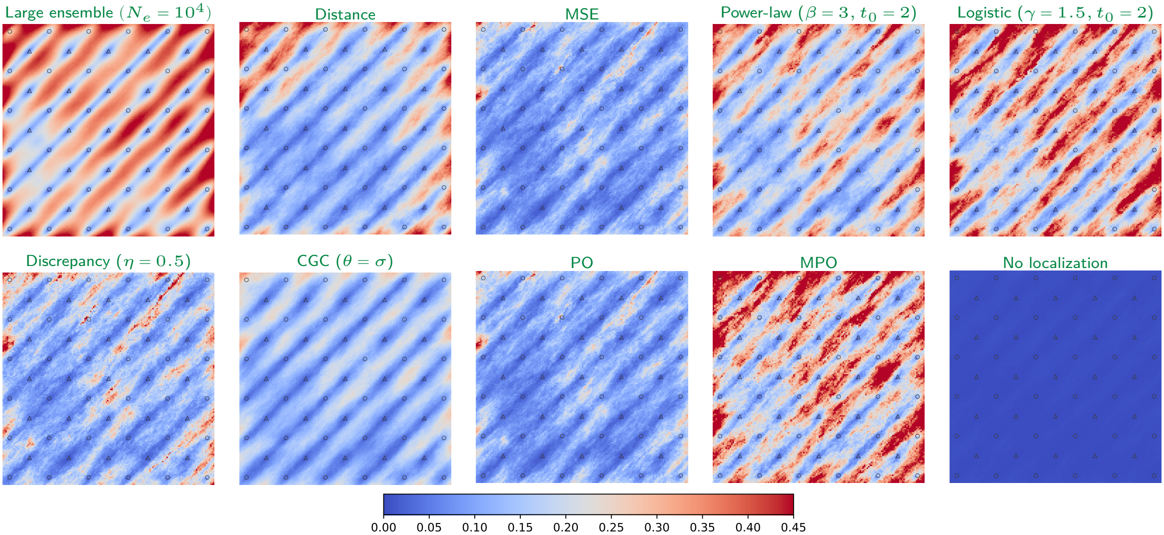
Average objective function



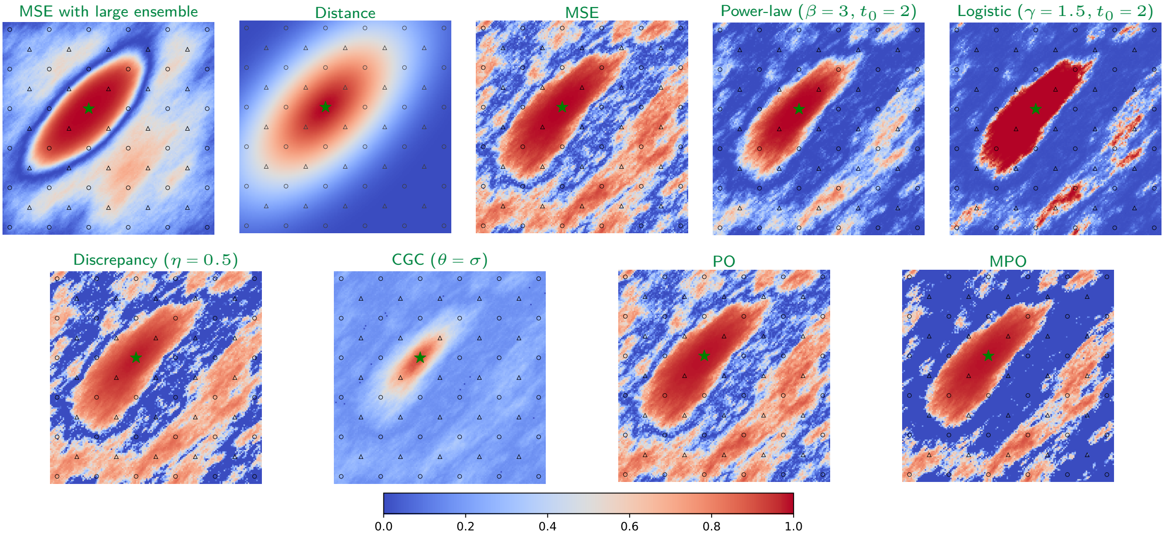
Normalized variance



Test Case 3: Normalized Variance

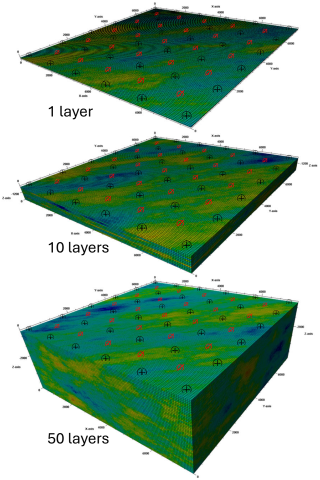


Test Case 3: Localization

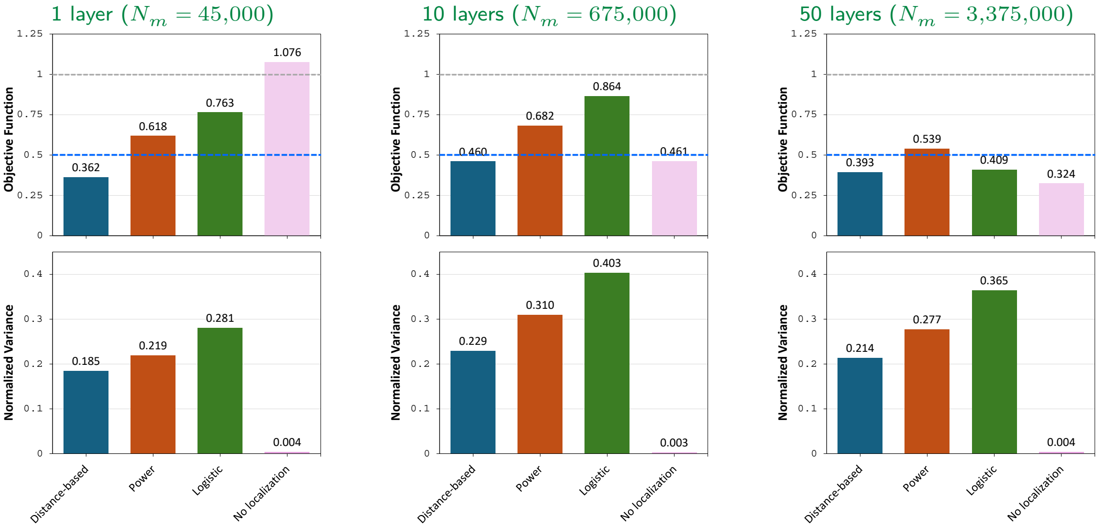


Test Case 4: Increasing Model Dimension

- Designed to test whether correlation-based localization remains selective in high-dimensional settings
- Same reservoir configuration, but with increasingly larger grid models:
  1. 1 layer: $N_m = 45{,}000$
  2. 10 layers: $N_m = 675{,}000$
  3. 50 layers: $N_m = 3{,}375{,}000$
- Question: Does increasing N_m make correlation-based localization less robust?



Test Case 4: Results



Final Comments

Correlation-based localization is a promising distance-free strategy:

- Preserves ensemble variability while maintaining acceptable data matches

Final Comments

Correlation-based localization is a promising distance-free strategy:

- Preserves ensemble variability while maintaining acceptable data matches
- Extends naturally to non-spatial parameters

Final Comments

Correlation-based localization is a promising distance-free strategy:

- Preserves ensemble variability while maintaining acceptable data matches
- Extends naturally to non-spatial parameters
- Adapts to non-local, flow-driven parameter-data relationships

Final Comments

Correlation-based localization is a promising distance-free strategy:

- Preserves ensemble variability while maintaining acceptable data matches
- Extends naturally to non-spatial parameters
- Adapts to non-local, flow-driven parameter-data relationships
- Performs competitively even when distance-based localization is favorable.
  (Further validation is needed in more complex data assimilation problems)

Final Comments

Pow-law and logistic tapers:

- Better discrimination between meaningful and spurious correlations

Final Comments

Pow-law and logistic tapers:

- Better discrimination between meaningful and spurious correlations
- Two free parameters (t_0 and exponent)

Final Comments

Pow-law and logistic tapers:

- Better discrimination between meaningful and spurious correlations
- Two free parameters (t_0 and exponent)
- Interpretation:
  - t_0: significance detection
  - exponent: correlation correction

Final Comments

Pow-law and logistic tapers:

- Better discrimination between meaningful and spurious correlations
- Two free parameters (t_0 and exponent)
- Interpretation:
  - t_0: significance detection
  - exponent: correlation correction
- Introduce a practical trade-off:
  - smoother tapers tend to improve data-match quality
  - sharper tapers tend to preserve ensemble variance more effectively

Final Comments

Computational cost:

- Computing all model-data correlations requires inner products over the ensemble dimension
- The total cost scales as $\mathcal{O}(N_m N_d N_e)$
- Problematic if $N_m \times N_d$ is large; e.g., big models with seismic data

Correlation-based localization appears to be a viable alternative to distance-based localization for subsurface applications

References

Emerick, A. A. and Silva, V. L. S. “Statistical Tapers for Correlation-Based Localization in Ensemble Data Assimilation”. In: *arXiv:2605.29922 [stat.ME]* (2026). DOI: 10.48550/arXiv.2605.29922.

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

Furrer, R. and Bengtsson, T. “Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants”. In: *Journal of Multivariate Analysis* 98.2 (2007), pp. 227–255. DOI: 10.1016/j.jmva.2006.08.003.

Luo, X. and Bhakta, T. “Automatic and adaptive localization for ensemble-based history matching”. In: *Journal of Petroleum Science and Engineering* 184 (2020). DOI: 10.1016/j.petrol.2019.106559.

Ranazzi, P. H.; Luo, X., and Sampaio, M. A. “Covariance scaling: Theory, extension, and applications to ensemble-based history matching”. In: *Computational Geosciences* on line (2026). DOI: 10.1007/s10596-026-10413-w.