High-dimensional Ensemble Kalman Filter with Localization, Inflation and Iterative Updates

Hao-Xuan Sun

Joint with Shouxia Wang, Xiaogu Zheng and Song Xi Chen

July 14, 2024 @JCSDS

Data Assimilation

Target: produce high-quality estimates of state variables in a dynamic system based on a numerical model and observations (Talagrand, 1997)

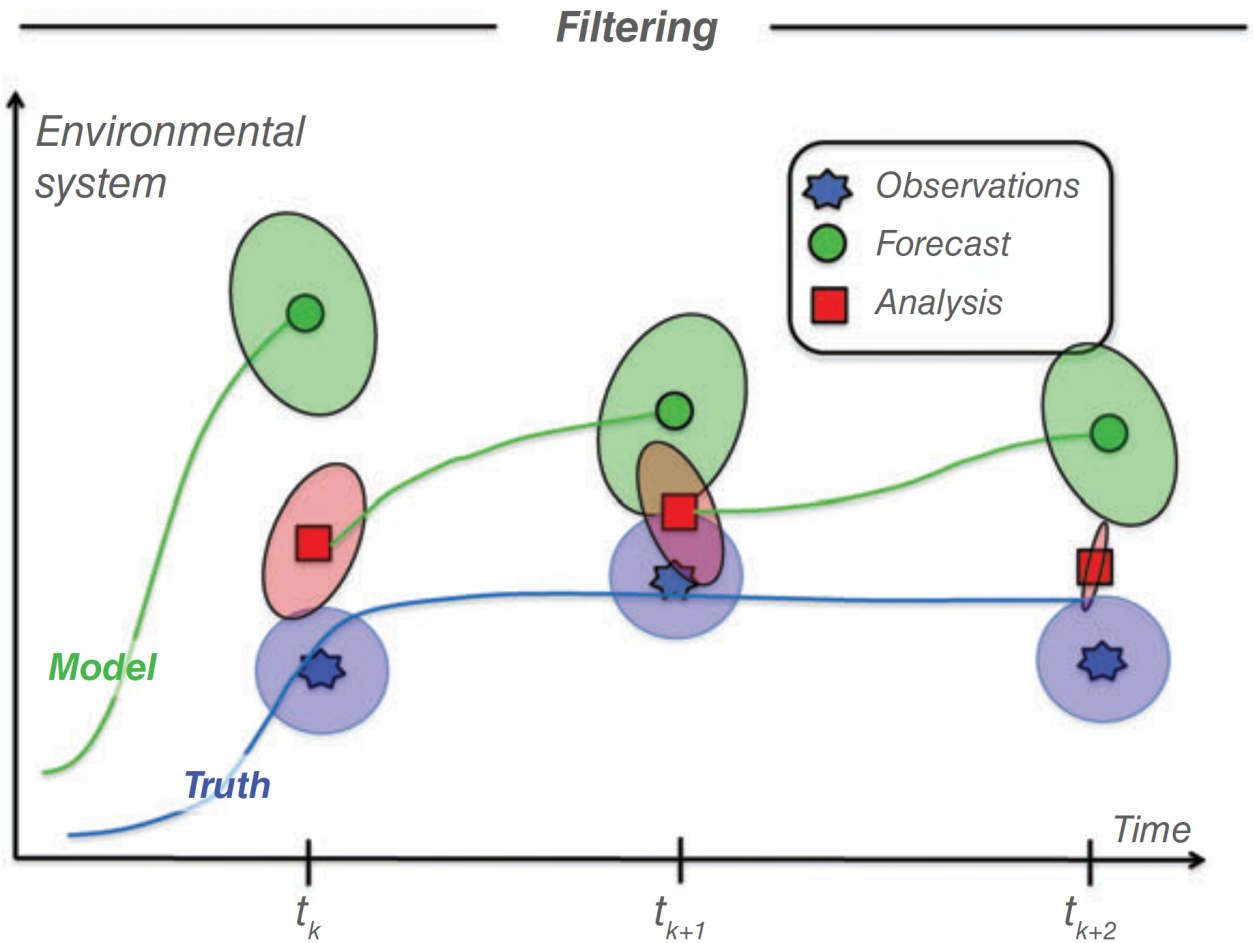


Figure 1: Illustration of filtering in data assimilation (Carrassi et al., 2018).

Ensemble Kalman Filter (EnKF)

Consider a possibly nonlinear discrete-time forecast and linear observational system (Ide et al., 1997)

$$\begin{cases} \mathbf{x}_i^t = \mathcal{M}_{i-1}\left(\mathbf{x}_{i-1}^t\right) + \boldsymbol{\eta}_i; \\ \mathbf{y}_i^o = \mathbf{H}_i \mathbf{x}_i^t + \boldsymbol{\varepsilon}_i, \end{cases} \tag{1}$$

where

- $\mathbf{x}_i^t$: p-dimensional true state vector at time step i;
- $\mathcal{M}_{i-1}$: nonlinear forecast operator at time step $i-1$;
- $\mathbf{y}_i^o$: q_i-dimensional observation vector;
- $\mathbf{H}_i$: linear observation operator;
- $\boldsymbol{\eta}_i$, $\boldsymbol{\varepsilon}_i$: model error and the observation error vectors, respectively, which are assumed to be independent, mean-zero and have covariance matrices $\mathbf{Q}_i$ and $\mathbf{R}_i$.

GOAL: find an analysis state $\mathbf{x}_i^a$ that is close to $\mathbf{x}_i^t$!

CHALLENGE: *limited ensemble size* and *model mis-specification*.

EnKF Algorithm (Evensen, 1994)

Step(1) Run the full model forward in time to get the perturbed forecast states:

$$\mathbf{x}_{i,j}^f = \mathcal{M}_i\left(\mathbf{x}_{i-1,j}^a\right); \quad \mathbf{x}_i^f = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_{i,j}^f. \tag{2}$$

Step(2.1) Compute the sample forecast error covariance matrix

$$\mathbf{S}_{n,i} \equiv \frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{x}_{i,j}^f - \mathbf{x}_i^f\right)\left(\mathbf{x}_{i,j}^f - \mathbf{x}_i^f\right)^{\mathrm{T}}, \tag{3}$$

and the perturbed observation-minus-forecast residuals

$$\mathbf{d}_{i,j} = \mathbf{y}_{o,i} + \varepsilon_{i,j}' - \mathbf{H}_i\mathbf{x}_{i,j}^f; \quad \mathbf{d}_i = \frac{1}{n}\sum_{j=1}^{n}\mathbf{d}_{i,j}, \tag{4}$$

where $\varepsilon_{i,j}'$ are sampled from $\mathcal{N}_{q_i}(\mathbf{0}_{q_i}, \mathbf{R}_i)$.

Step(2.2) Calculate the forecast error covariance matrix as $\hat{\mathbf{P}}_i = \mathbf{S}_{n,i}$ and inflate it by a factor $\hat{\lambda}_i$.

Step(2.3) Compute the perturbed analysis states

$$\mathbf{x}_{i,j}^a = \mathbf{x}_{i,j}^f + \hat{\mathbf{P}}_i\mathbf{H}_i^{\mathrm{T}}\left(\mathbf{H}_i\hat{\mathbf{P}}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)^{-1}\mathbf{d}_{i,j}. \tag{5}$$

Step(3) If i is not the ending time of data assimilation, set $i = i+1$ and repeat Step (1)-(2). Otherwise, the filtering ends.

Inflation

Additive inflation: set $\hat{\mathbf{P}}_i = \hat{\mathbf{P}}_i + \zeta_i$ (e.g., Hamill and Whitaker 2005 and Constantinescu et al. 2007).

Muliplicative inflation: set $\hat{\mathbf{P}}_i = \lambda_i \hat{\mathbf{P}}_i$, with $\lambda_i > 1$ in usual.

- First order estimator (Wang and Bishop, 2003)

$$\lambda_i = \frac{\left(\mathbf{R}_i^{-\frac{1}{2}}\mathbf{d}_i\right)^{\mathrm{T}}\left(\mathbf{R}_i^{-\frac{1}{2}}\mathbf{d}_i\right) - p}{\mathrm{tr}\left(\mathbf{R}_i^{-\frac{1}{2}}\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}\mathbf{R}_i^{-\frac{1}{2}}\right)} \tag{6}$$

- Second order estimator (Wu et al., 2013)

$$\lambda_i = \frac{\mathrm{tr}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}\left(\mathbf{d}_i\mathbf{d}_i^{\mathrm{T}} - \mathbf{R}_i\right)\right)}{\mathrm{tr}\left(\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}\mathbf{H}_i\mathbf{P}_i\mathbf{H}_i^{\mathrm{T}}\right)} \tag{7}$$

- Maximum likelihood estimator (MLE, Liang et al. 2012, Zheng 2009): obtain inflation factor $\hat{\lambda}_i$ by minimizing

$$L(\lambda_i) = \ln\det\left(\mathbf{H}_i\lambda_i\hat{\mathbf{P}}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right) + \mathbf{d}_i^{\mathrm{T}}\left(\mathbf{H}_i\lambda_i\hat{\mathbf{P}}_i\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i\right)^{-1}\mathbf{d}_i. \tag{8}$$

Iterative Update

Key: replace $\mathbf{x}_i^f$ with $\mathbf{x}_i^a$, a better estimation of $\mathbf{x}_i^t$ (Wu et al., 2013).

Step(2.4) Initializes $\mathbf{x}_{i,j}^{a(0)} = \mathbf{x}_{i,j}^a$, $\mathbf{x}_i^{a(0)} = \mathbf{x}_i^a$ and $\hat{\mathbf{P}}_i^{(0)} = \mathbf{S}_{n,i}$. For the rth round, update $\mathbf{S}_{n,i}^{(r)} = \frac{1}{n-1}\sum_{j=1}^n (\mathbf{x}_{i,j}^f - \mathbf{x}_i^{a(r-1)})(\mathbf{x}_{i,j}^f - \mathbf{x}_i^{a(r-1)})^{\mathrm{T}}$ with $\mathbf{x}_i^{a(r-1)} = n^{-1}\sum_{j=1}^n \mathbf{x}_{i,j}^{a(r-1)}$ calculated in the $r-1$th round and estimate $\lambda_i^{(r)}$ as $\hat{\lambda}_i^{(r)}$ that minimizes (8). Generate $\mathbf{x}_{i,j}^{a(r)}$ via (5) and compute $\mathbf{x}_i^{a(r)}$.

Step(2.5) If $L\left(\hat{\lambda}_i^{(r-1)}\right) - L\left(\hat{\lambda}_i^{(r)}\right) > \delta$, set $r = r+1$ and repeat Step(2.4). Otherwise, stop the iteration and update the $\mathbf{x}_{i,j}^a$ as (5) with $\hat{\lambda}_i = \hat{\lambda}_i^{(r-1)}$ and $\hat{\mathbf{P}}_i = \hat{\mathbf{P}}_i^{(r-1)}$.

Iterative updates can be viewed as additive inflations (Zheng et al., 2013).

$$\frac{\hat{\lambda}_i}{n-1}\sum_{j=1}^n \left(\mathbf{x}_{i,j}^f - \mathbf{x}_i^a\right)\left(\mathbf{x}_{i,j}^f - \mathbf{x}_i^a\right)^{\mathrm{T}} =$$

$$\underbrace{\frac{\hat{\lambda}_i}{n-1}\sum_{j=1}^n \left(\mathbf{x}_{i,j}^f - \mathbf{x}_i^f\right)\left(\mathbf{x}_{i,j}^f - \mathbf{x}_i^f\right)^{\mathrm{T}}}_{\text{multiplicative inflation}} + \underbrace{\frac{\hat{\lambda}_i n}{n-1}\left(\mathbf{x}_i^f - \mathbf{x}_i^a\right)\left(\mathbf{x}_i^f - \mathbf{x}_i^a\right)^{\mathrm{T}}}_{\text{additive inflation}}$$

Localization

Diagnose and ignore the unphysical correlation systematically (Hamill et al., 2001, Houtekamer and Mitchell, 1998, 2001).

- Set $\hat{\mathbf{P}}_i = \rho \circ \hat{\mathbf{P}}_i$, where ρ is a manually chosen localization matrix.

Bandable covariance matrix class for some positive ϵ, α, C:

$$\mathcal{U}(\epsilon, \alpha, C) = \Big\{ \mathbf{\Sigma} = \left[\sigma_{\ell_1 \ell_2}\right]_{p \times p} : \quad \text{(i)} \quad \max_{\ell_2} \sum_{k_{\ell_1 \ell_2} > k} |\sigma_{\ell_1 \ell_2}| \leq C k^{-\alpha} \text{ for all } k > 0;$$

$$\text{(ii)} \quad 0 < \epsilon \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq \epsilon^{-1} \Big\}, \quad (9)$$

where $k_{\ell_1 \ell_2}$ is the distance between ℓ_1th and ℓ_2th grids.

Tapering Estimator: let $g(z)$ be a tapering function, which is non-increasing, non-negative and satisfies $g(0) = 1$, $g(z) = 0$ for $z > 1$ and $g(z) > 0$ for $z \in (0, 1)$. Then, under mild conditions, a statistically consistent estimator is

$$\mathcal{T}_g(\mathbf{\Sigma}, k_g) \equiv \left[\sigma_{\ell_1 \ell_2} g(k_{\ell_1 \ell_2} / k_g)\right]_{p \times p},$$

where k_g is localization length scale.

Localization Length Scale Selection

Choose the localization parameter by minimizing the expectation of the standardized square Frobenius norm (Qiu and Chen, 2015).

$$L_{g,i}(k) \equiv p^{-1}\mathbb{E}\|\mathcal{T}_g(\mathbf{S}_{n,i}, k_{g,i}) - \mathbf{P}_i\|_{\mathrm{F}}^2 = \frac{1}{p}\mathrm{tr}(\mathbf{P}_i^2) + \frac{1}{p}\left(1 - n^{-1}\right)\tilde{L}_{g,i}(k), \tag{10}$$

where

$$\tilde{L}_{g,i}(k) = \sum_{k_{\ell_1\ell_2} \le k_{g,i}} \left[\left\{ g(k_{\ell_1\ell_2}/k_{g,i})^2 - 2g(k_{\ell_1\ell_2}/k_{g,i}) \right\} \sigma_{\ell_1\ell_2}^2 + n^{-1} g(k_{\ell_1\ell_2}/k_{g,i})^2 \sigma_{\ell_1\ell_1}\sigma_{\ell_2\ell_2} \right]. \tag{11}$$

Estimate k_g by minimizing (11) with the imputation

$$\begin{aligned}
\hat{\sigma}_{\ell_1\ell_2}^2 &= \frac{1}{A_n^2}\sum_{j_1,j_2}^{\star} \mathbf{x}_{j_1}(\ell_1)\mathbf{x}_{j_1}(\ell_2)\mathbf{x}_{j_2}(\ell_1)\mathbf{x}_{j_2}(\ell_2) - 2\frac{1}{A_n^3}\sum_{j_1,j_2,j_3}^{\star} \mathbf{x}_{j_1}(\ell_1)\mathbf{x}_{j_2}(\ell_2)\mathbf{x}_{j_3}(\ell_1)\mathbf{x}_{j_3}(\ell_2) \\
&\quad + \frac{1}{A_n^4}\sum_{j_1,j_2,j_3,j_4}^{\star} \mathbf{x}_{j_1}(\ell_1)\mathbf{x}_{j_2}(\ell_2)\mathbf{x}_{j_3}(\ell_1)\mathbf{x}_{j_4}(\ell_2), \text{ and} \\
\widehat{\sigma_{\ell_1\ell_1}\sigma_{\ell_2\ell_2}} &= \frac{1}{A_n^2}\sum_{j_1,j_2}^{\star} \left(\mathbf{x}_{j_1}(\ell_1)\right)^2 \left(\mathbf{x}_{j_2}(\ell_2)\right)^2 - \frac{1}{A_n^3}\sum_{j_1,j_2,j_3}^{\star} \mathbf{x}_{j_1}(\ell_1)\mathbf{x}_{j_2}(\ell_2)\left(\mathbf{x}_{j_3}(\ell_1)\right)^2 \\
&\quad - \frac{1}{A_n^3}\sum_{j_1,j_2,j_3}^{\star} \left(\mathbf{x}_{j_1}(\ell_1)\right)^2 \mathbf{x}_{j_2}(\ell_1)\mathbf{x}_{j_3}(\ell_2) + \frac{1}{A_n^4}\sum_{j_1,j_2,j_3,j_4}^{\star} \mathbf{x}_{j_1}(\ell_1)\mathbf{x}_{j_2}(\ell_2)\mathbf{x}_{j_3}(\ell_1)\mathbf{x}_{j_4}(\ell_2),
\end{aligned}$$

where $\sum^{\star}$ denotes the summation over different subscripts and $A_n^c = n!/(n-c)!$.

High-dimension EnKF (HD-EnKF)

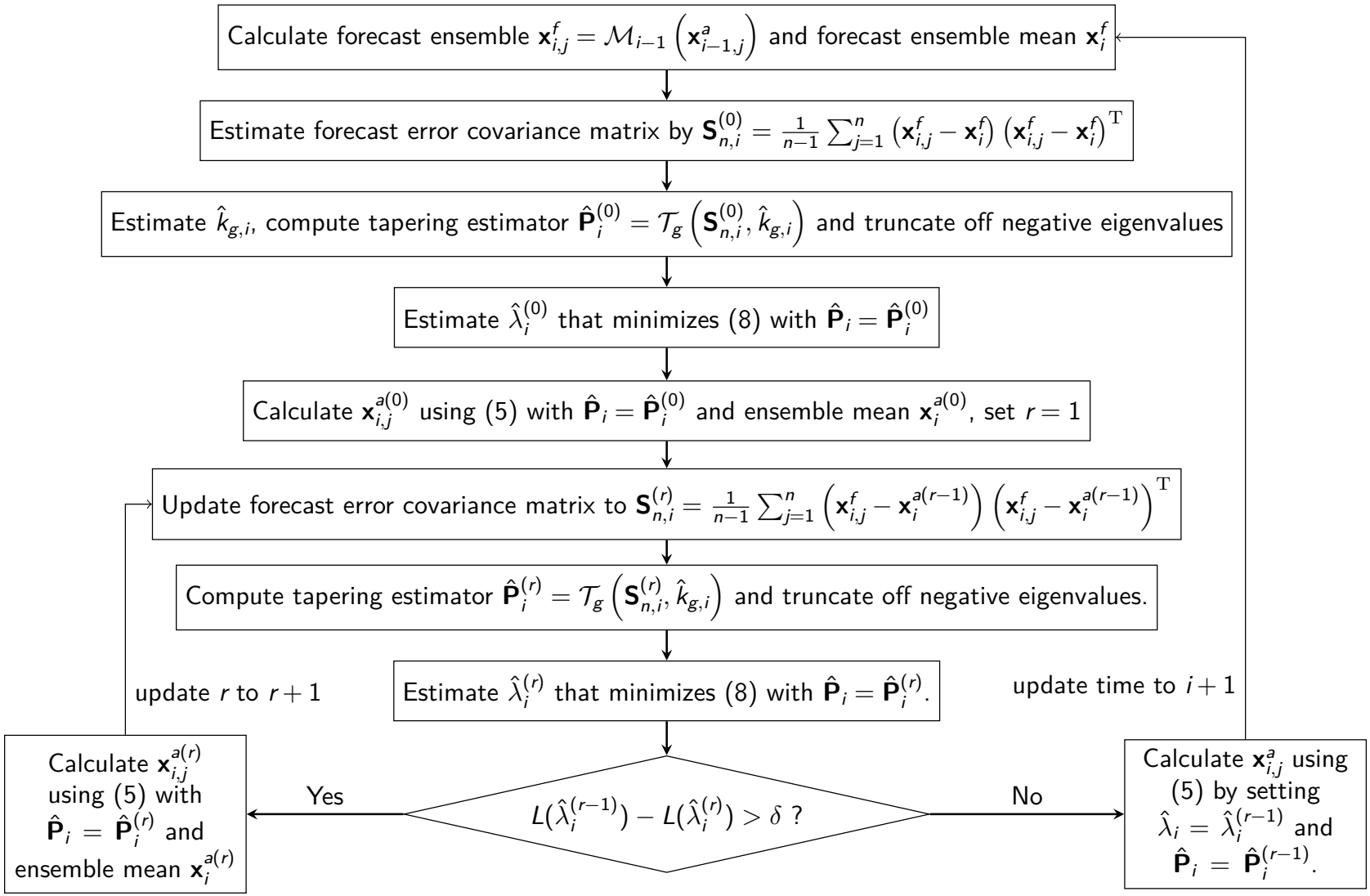


Figure 2: A flowchart of the proposed high-dimensional EnKF (HD-EnKF) assimilation scheme that conducts the localization, the inflation and the iterative updates.

Lorenz-96 model

Lorenz-96 model (L96, Lorenz 1996): a strongly nonlinear model:

$$\frac{d\mathbf{x}(j)}{dt} = \{\mathbf{x}(j+1) - \mathbf{x}(j-2)\}\,\mathbf{x}(j-1) - \mathbf{x}(j) + F, \tag{12}$$

for $j = 1, \ldots, p$, $\mathbf{x}_{-1} = \mathbf{x}_{p-1}, \mathbf{x}_0 = \mathbf{x}_p$ and $\mathbf{x}_{p+1} = \mathbf{x}_1$.

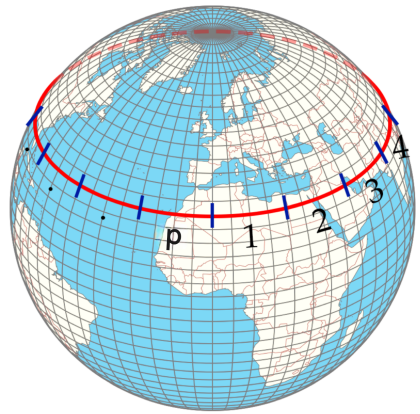


Figure 3: Latitude circle of the earth, divided into p equal-sized sectors (van Kekem, 2018).

- True state: solve (12) by the fourth-order Runge-Kutta integration (Butcher, 2016) with a time step 0.05 and 2000 steps in total.
- Initial condition $\mathbf{x}_1^t(j) = F$ for $j \neq \lfloor \frac{p}{2} \rfloor$ and $\mathbf{x}_1^t(\lfloor \frac{p}{2} \rfloor) = F + 0.001$.
- Observation: add a $\mathcal{N}_{q_i}(\mathbf{0}_{q_i}, \mathbf{R}_i)$ distributed noise to $\mathbf{x}_i^t$, with the $(\ell_1, \ell_2)-$element of $\mathbf{R}_i$ being $0.5^{\min\{|\ell_1-\ell_2|, p-|\ell_1-\ell_2|\}}$.
- $p = 40, 100, 200$ and $n = 20, 30, 40$, repsectively.
- Initial ensemble: add a $\mathcal{N}_p(\mathbf{0}_p, 0.1\boldsymbol{I}_p)$ distributed noise to $\mathbf{x}_1^t$.

Simulation Settings

Mis-specified Model: set $F' = 4, 5, \ldots, 12$ for L96 to generate biased forecasts while the true forcing term $F = 8$.

Measurement: averaged root mean square error (RMSE) over the last 1000 steps and 50 trials

$$\text{RMSE} = \frac{1}{50000}\sqrt{\sum_{b=1}^{50}\sum_{i=1001}^{2000} \|\mathbf{x}_i^{a,b} - \mathbf{x}_i^t\|_2^2}, \tag{13}$$

Comparing schemes: the standard EnKF (standard), the EnKF with inflation and iterative updates (inflation + iterative updates), the EnKF with localization (localization) and the Oracle.

Tapering functions:

- banding function $g(z) = \mathbb{I}\{0 \leq z \leq 1\}$ (Bickel and Levina, 2008);
- linearly banding function $g(z) = \mathbb{I}\{0 \leq z \leq \frac{1}{2}\} + (2 - 2z)\mathbb{I}\{\frac{1}{2} < z \leq 1\}$ (Cai et al., 2010);
- the 5th-order piece-wise rational function $g(z) = \phi(2z)$ (Gaspari and Cohn, 1999) where

$$\phi(z) = \begin{cases} 1 - \frac{5}{3}z^2 + \frac{5}{8}z^3 + \frac{1}{2}z^4 - \frac{1}{4}z^5, & 0 \leq z \leq 1; \\ -\frac{2}{3}z^{-1} + 4 - 5z + \frac{5}{3}z^2 + \frac{5}{8}z^3 - \frac{1}{2}z^4 + \frac{1}{12}z^5, & 1 < z \leq 2; \\ 0, & z \geq 2. \end{cases}$$

Assimilation Results on L96

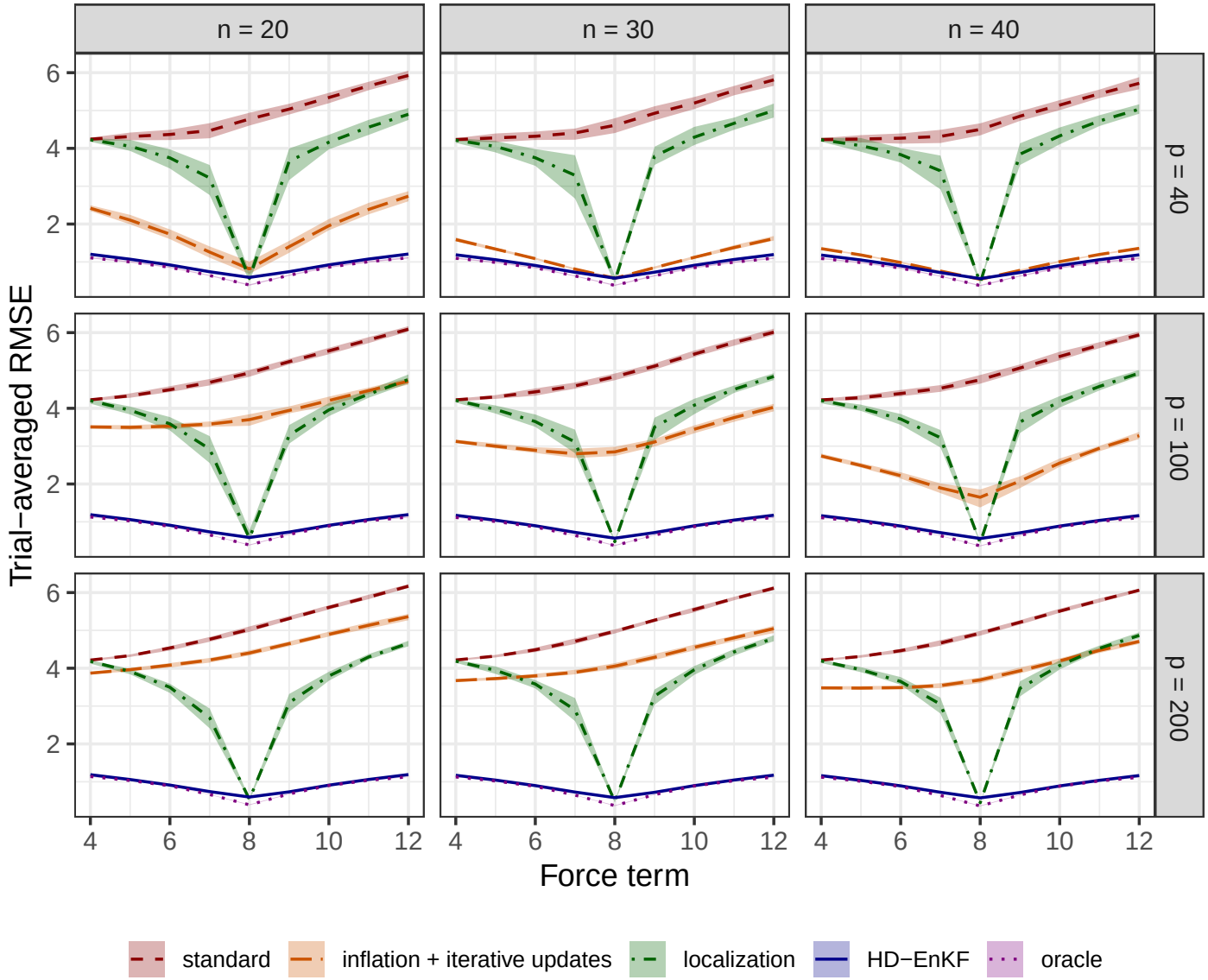


Figure 4: Trial-averaged RMSEs and their 5% − 95% quantile bands over the last 1000 steps as a function of forcing term F' under $p = q = 40, 100, 200$ and $n = 20, 30, 40$.

Comparison of L96 Trials

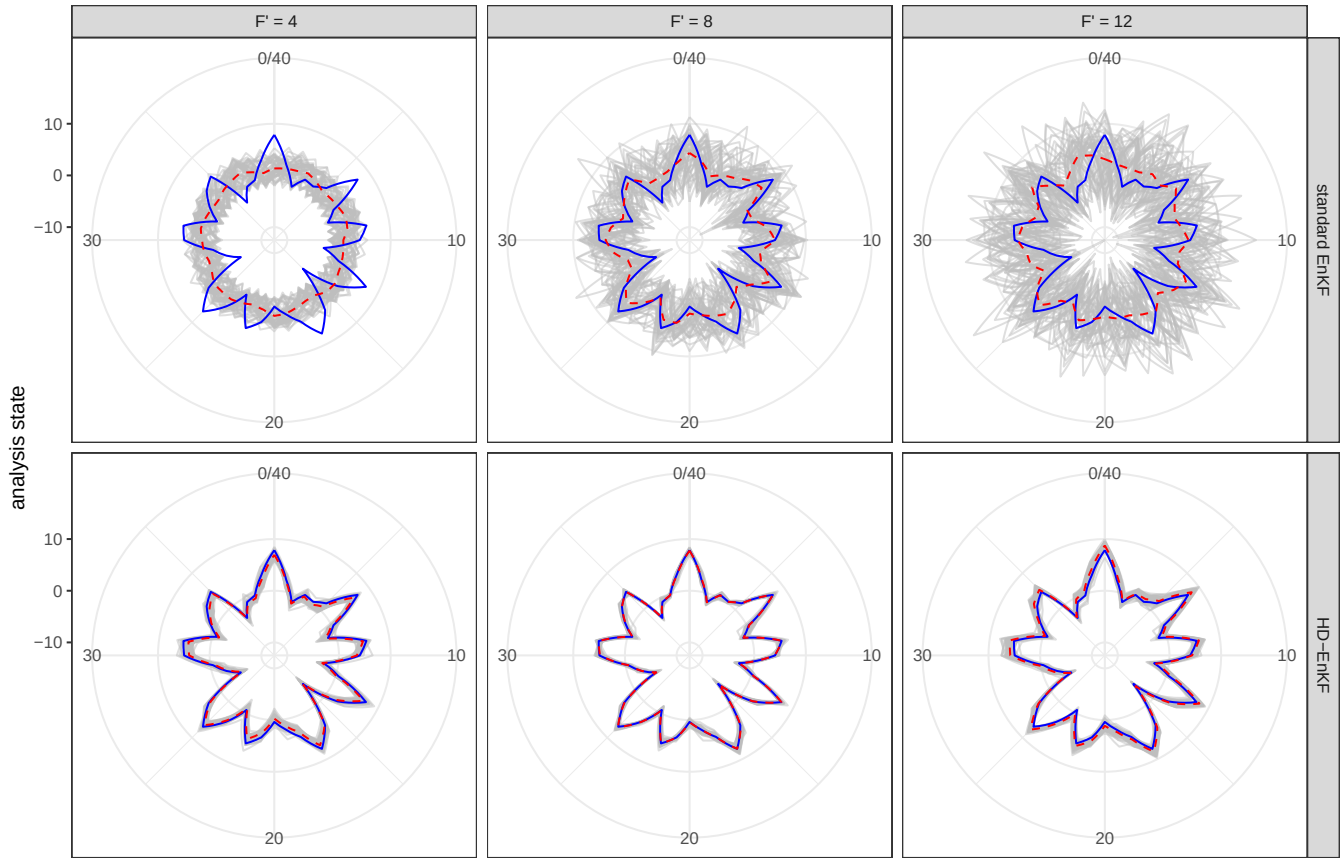
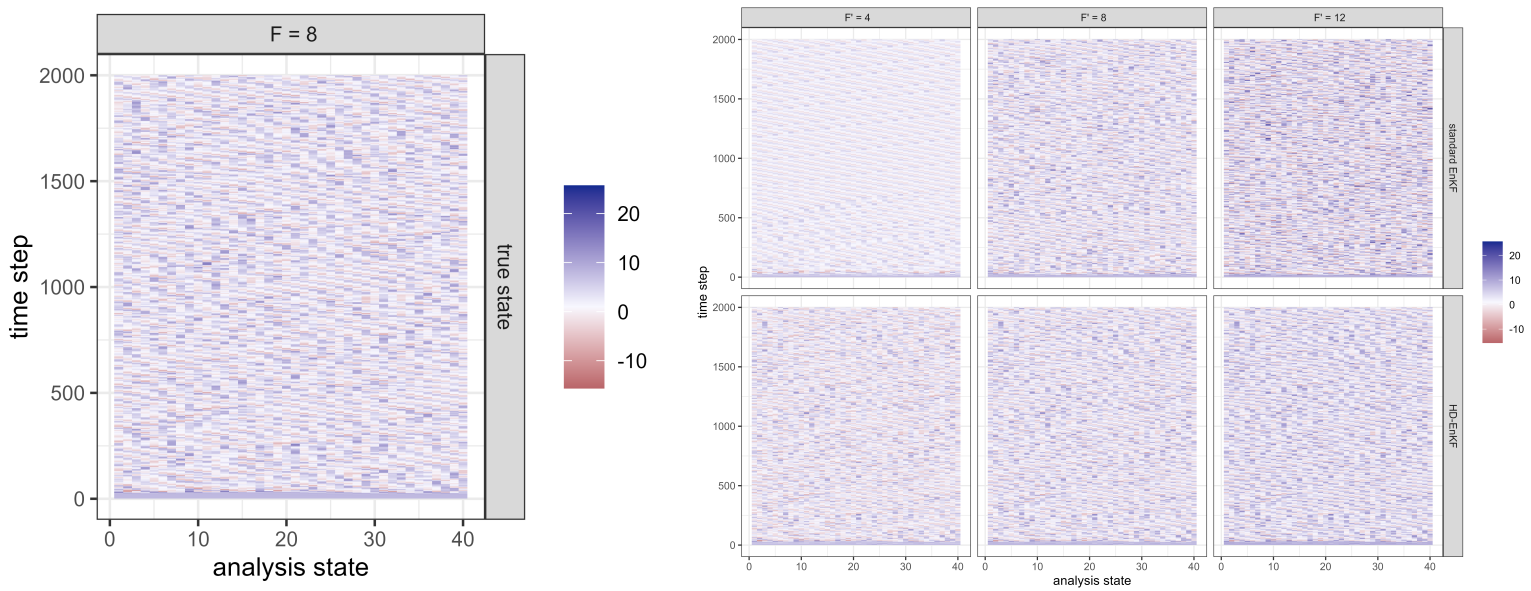


Figure 5: The background error (red solid line), the background spread (red dashed line), the analysis error (blue solid line) and the analysis spread (blue dashed line) in a simulation trial of the standard EnKF scheme and the HD-EnKF scheme under $p = q = 40$, $n = 30$ and $F' = 8$.

Comparison of L96 Analysis States



(a) true state (b) analysis state

Figure 6: The true state (panel (a)) and the analysis states (panel (b)) of the standard EnKF scheme and the HD-EnKF scheme in a trial under $p = q = 40$, $n = 30$, $F' = 4, 8, 12$ and at time step $i = 1$ to 2000, with the colors from red to blue denoting the analysis states from small to large.

Results for Mis-specified Models

EnKF scheme	RMSE	$L(\hat{\lambda}_i)$	$\hat{k}_{g,i}$	RMSE	$L(\hat{\lambda}_i)$	$\hat{k}_{g,i}$	RMSE	$L(\hat{\lambda}_i)$	$\hat{k}_{g,i}$
	$p = 40$ $q = 40$ $n = 20$			$p = 40$ $q = 40$ $n = 30$			$p = 40$ $q = 40$ $n = 40$		
standard	5.93(0.069)	2173.91	-	5.81(0.087)	2076.58	-	5.72(0.09)	2005.57	-
inflation + iterative updates	2.74(0.085)	287.22	-	1.62(0.038)	78.15	-	1.36(0.019)	48.80	-
localization	4.9(0.102)	1436.41	11.6	5(0.106)	1498.34	14.2	5.04(0.086)	1519.89	16.4
HD-EnKF(BL)	1.36(0.016)*	67.26	5.2	1.31(0.014)*	61.63	5.9	1.3(0.011)*	58.58	6.4
HD-EnKF(CZZ)	1.33(0.013)*	63.73	6.7	1.29(0.012)*	58.84	7.6	1.27(0.012)*	55.89	8.3
HD-EnKF(GC)	1.21(0.01)*	50.53	15	1.19(0.009)*	47.36	17.1	1.19(0.011)*	45.58	18.7
	$p = 100$ $q = 100$ $n = 20$			$p = 100$ $q = 100$ $n = 30$			$p = 100$ $q = 100$ $n = 40$		
standard	6.09(0.043)	5823.40		6.01(0.052)	5635.99		5.94(0.05)	5480.51	
inflation + iterative updates	4.72(0.059)	2980.31		4.03(0.062)	1929.16		3.28(0.055)	1111.42	
localization	4.76(0.075)	3370.92	10.4	4.84(0.056)	3490.36	12.8	4.93(0.057)	3637.73	14.7
HD-EnKF(BL)	1.34(0.009)*	165.74	4.6	1.3(0.011)*	150.92	5.7	1.28(0.009)*	145.04	6
HD-EnKF(CZZ)	1.3(0.009)*	154.75	6.6	1.27(0.009)*	142.98	7.3	1.25(0.007)*	135.82	7.8
HD-EnKF(GC)	1.19(0.007)*	120.17	14.7	1.17(0.007)*	112.59	16.6	1.16(0.007)*	107.97	17.9
	$p = 200$ $q = 200$ $n = 20$			$p = 200$ $q = 200$ $n = 30$			$p = 200$ $q = 200$ $n = 40$		
standard	6.17(0.027)	11991.93		6.12(0.033)	11750.98		6.06(0.029)	11499.24	
inflation + iterative updates	5.36(0.05)	8465.45		5.05(0.06)	7135.95		4.71(0.043)	5872.27	
localization	4.66(0.047)	6437.26	9.7	4.79(0.045)	6814.06	11.9	4.87(0.045)	7058.22	13.7
HD-EnKF(BL)	1.34(0.007)*	329.13	4.7	1.3(0.006)*	304.08	5.6	1.29(0.007)*	292.07	5.9
HD-EnKF(CZZ)	1.31(0.006)*	309.09	6.4	1.27(0.006)*	286.33	7.1	1.25(0.006)*	271.84	7.7
HD-EnKF(GC)	1.18(0.005)*	236.22	14.5	1.17(0.005)*	221.35	16.2	1.16(0.005)*	212.43	17.5

Table 1: Trial-averaged RMSEs with the standard deviation in the parentheses, minimized trial-averaged objective function values and average selected localization length scales over the last 1000 steps under $p = q = 40, 100, 200$, $n = 20, 30, 40$ and $F' = 12$. The subscripts * mark the trial-averaged RMSEs of the proposed HD-EnKF which were significantly smaller at 99% confidence than those of the three existing EnKF methods (standard, inflation + iterative updates and localization) under the same (p, n) setting.

Results for Sparse Observations

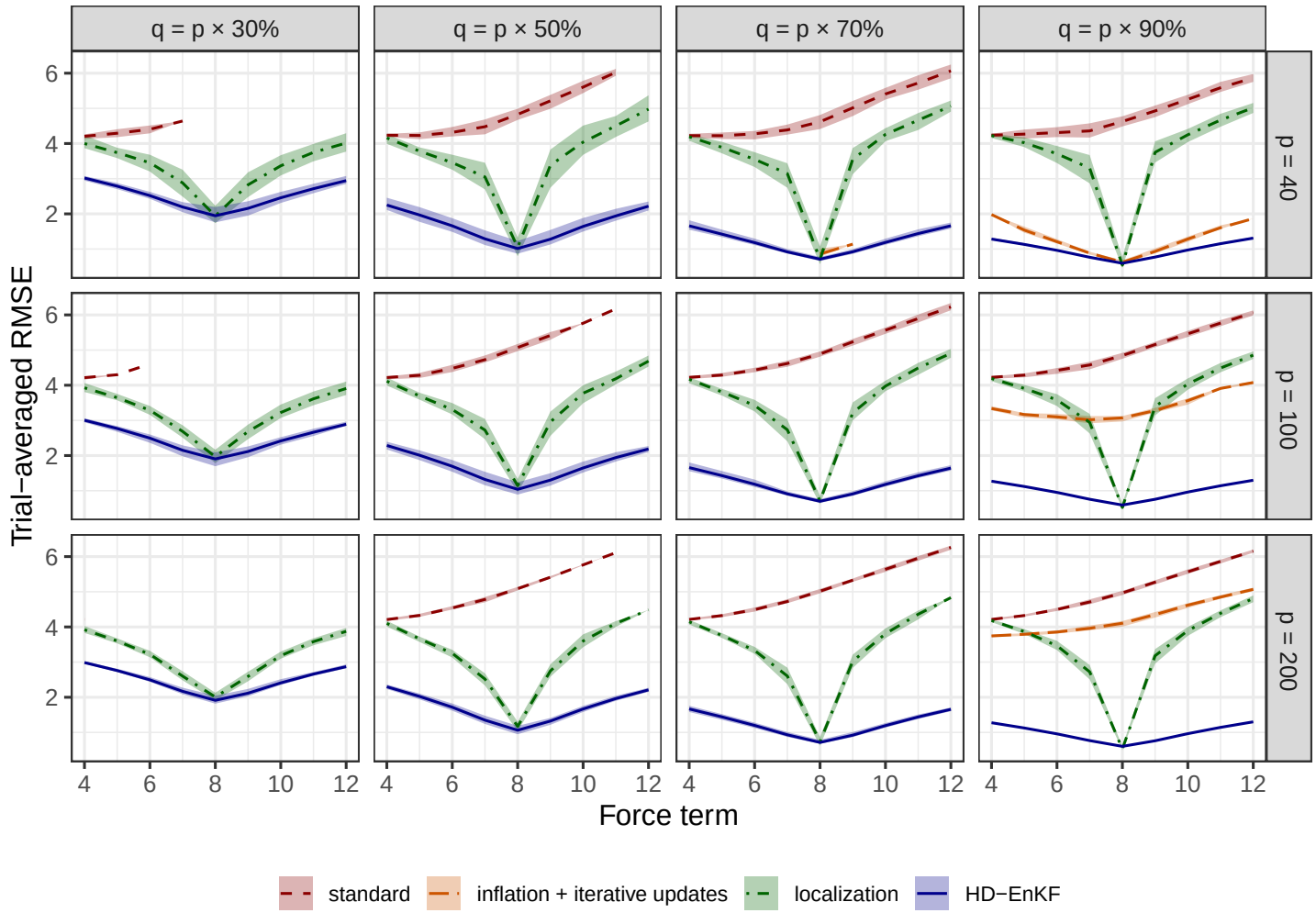


Figure 7: Trial-averaged RMSEs and their 5% − 95% quantile bands over the last 1000 steps as a function of the forcing term F' under $p = 40, 100, 200$, $q = p \times 30\%, 50\%, 70\%, 90\%$ and $n = 30$.

Convergence Rate under Sparse Observations

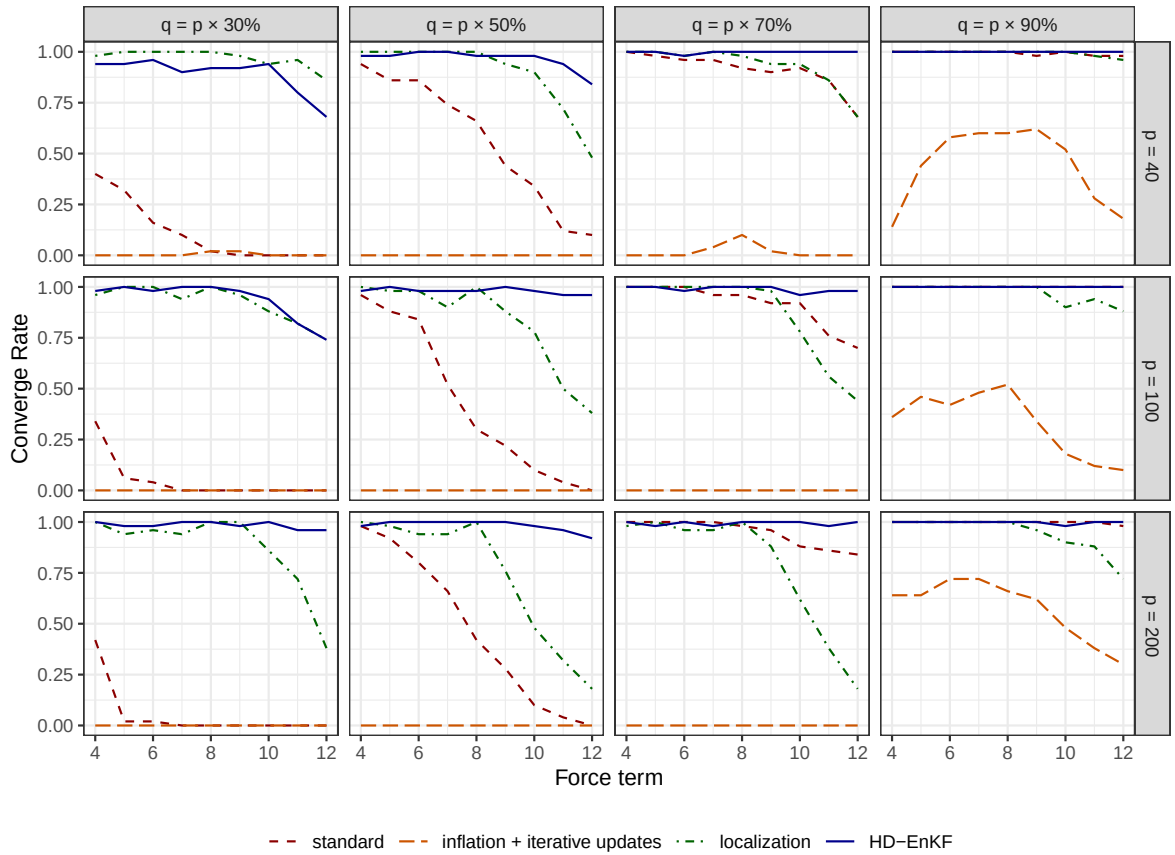


Figure 8: The converge rate (1 − diverge rate) over the 50 repetitions as a function of forcing term F' under $p = 40, 100, 200$, $q = p \times 30\%, 50\%, 70\%, 90\%$ and $n = 30$ at each time step for the four EnKF schemes: the standard EnKF (red dashed line); the EnKF with inflation and iterative updates (orange long-dashed line); the EnKF with localization (green dot-dashed line) and the HD-EnKF (blue solid line).

HD-EnKF under Skewed Distributed Observation Error

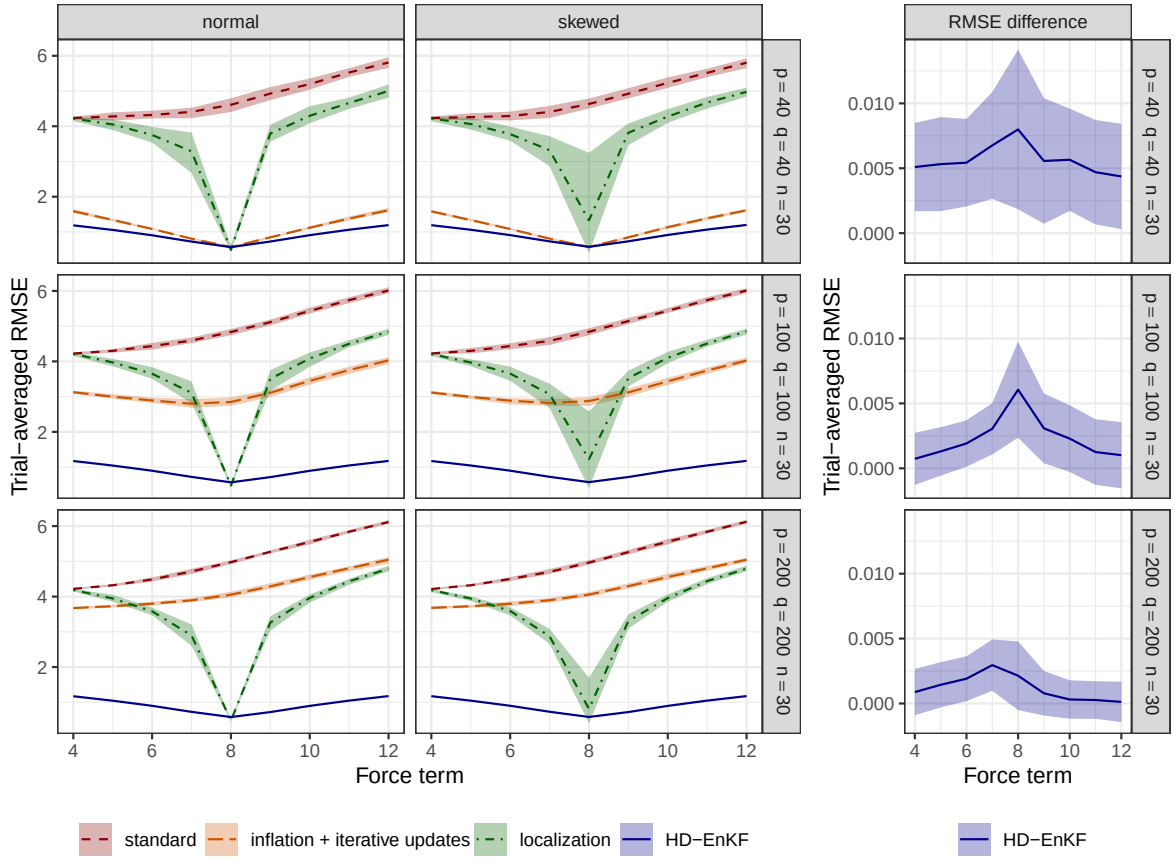


Figure 9: Left panel: the trial-averaged RMSEs and their 5% − 95% quantile bands over the last 1000 steps as a function of forcing term F' under $p = q = 40, 100, 200$, $n = 30$ and normal or skewed distributed observation errors for the four EnKF schemes: the standard EnKF (red dashed line); the EnKF with inflation and iterative updates (orange long-dashed line); the EnKF with localization (green dot-dashed line) and the HD-EnKF (blue solid line). Right panel: the differences of the HD-EnKF trial-averaged RMSEs by assimilating observations with the skewed and the normal distributed errors (skewed-minus-normal differences), with the bands denoting the 95% confidence intervals.

HD-EnKF with Constant Localization Length Scale

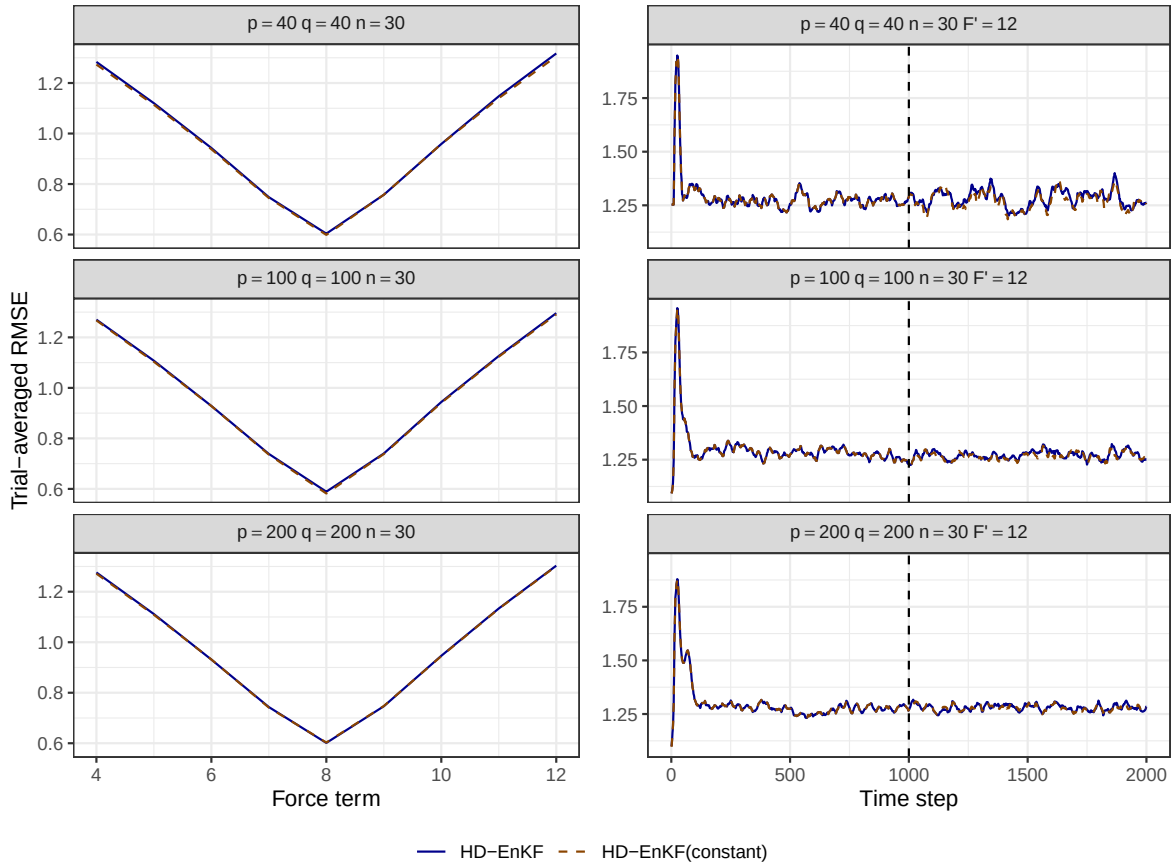


Figure 10: The trial-averaged RMSEs over the last 1000 steps as a function of forcing term F' under $p = q = 40, 100, 200$ and $n = 30$ (left) and at each time step under $p = q = 40, 100, 200$, $n = 30$ and $F' = 12$ (right) for the HD-EnKF assimilation scheme with the time-varying localization length scales selected at each assimilation step (blue solid line) or the constant localization length scale which is equal to the average localization length scales of the first 1000 steps (denoted by 'constant', brown dashed line).

HD-EnKF with Eigenvalue Truancation

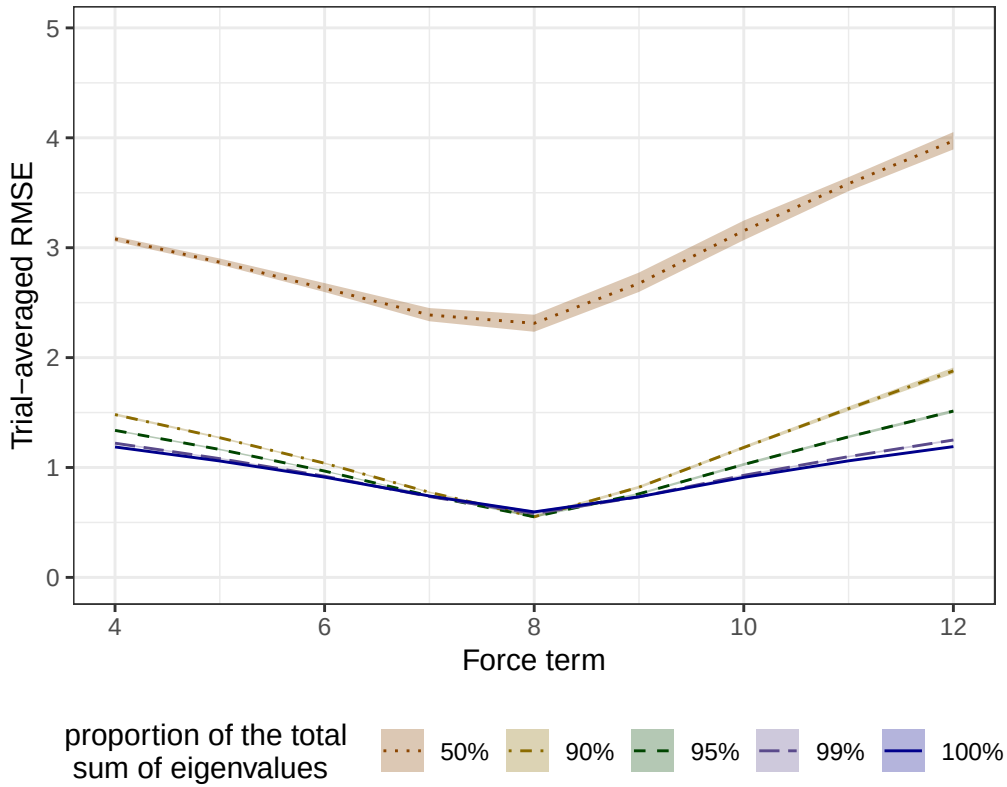


Figure 11: The trial-averaged RMSEs and their 5% − 95% quantile bands over the last 1000 steps as a function of forcing term F' under $p = q = 200$ and $n = 20$ for the HD-EnKF scheme reconstructed by the eigenvalues that account for 50% (brown dotted line), 90% (light brown dot-dashed line), 95% (green dashed line), 99% (light blue long-dashed line) and 100% (blue solid line) of the total sum of all the eigenvalues, respectively.

Blockwise HD-EnKF

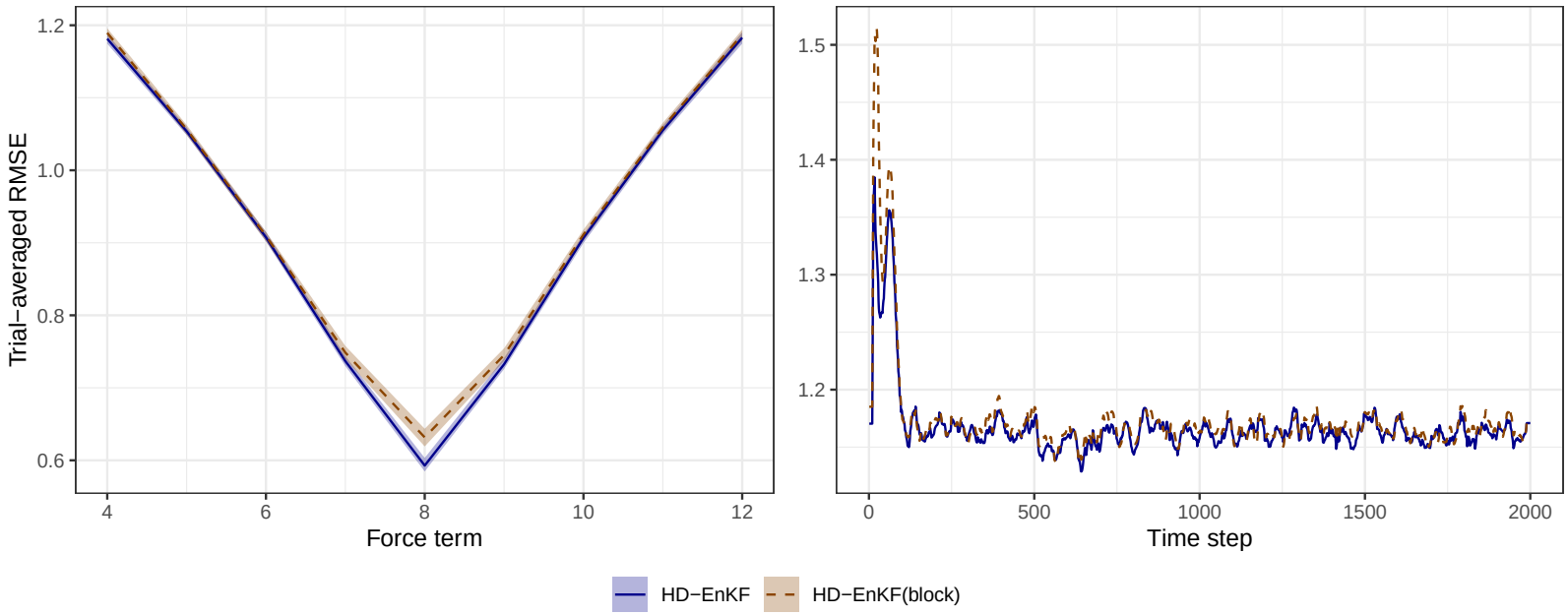


Figure 12: Left panel: The trial-averaged RMSEs and their 5% − 95% quantile bands over the last 1000 steps as a function of forcing term F' under $p = q = 200$ and $n = 20$. Right panel: the trial-averaged RMSEs at each time step under $p = q = 200$, $n = 20$ and $F' = 12$ for the original HD-EnKF assimilation scheme (blue solid line) and the blockwise HD-EnKF assimilation scheme (brown dashed line).

Summary

- The standard EnKF may filter to assimilate the observations due to the limited ensemble size and the model mis-specification.
- We introduced the high-dimensional covariance tapering estimator to the EnKF assimilation scheme and provided the selection method for the localization length scale.
- We proposed the HD-EnKF algorithm, which has three ingredients to the standard EnKF, that is, the localization, the inflation and the iterative updates.
- The proposed assimilation scheme is tested on the L96 model and can achieved assimilation performances in very close proximity to the best-tuned versions.

Reference I

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Butcher, J. C. (2016). *Numerical methods for ordinary differential equations*. John Wiley & Sons.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5).

Constantinescu, E. M., Sandu, A., Chai, T., and Carmichael, G. R. (2007). Ensemble-based chemical data assimilation. I: General approach. *Quarterly Journal of the Royal Meteorological Society*, 133(626):1229–1243.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162.

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

Hamill, T. M. and Whitaker, J. S. (2005). Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Monthly weather review*, 133(11):3132–3147.

Hamill, T. M., Whitaker, J. S., and Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, 129(11):2776–2790.

Houtekamer, P. L. and Mitchell, H. L. (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126(3):796–811.

Houtekamer, P. L. and Mitchell, H. L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129(1):123–137.

Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C. (1997). Unified notation for data assimilation: Operational, sequential and variational. *Journal of the Meteorological Society of Japan. Ser. II*, 75(1B):181–189.

Liang, X., Zheng, X., Zhang, S., Wu, G., Dai, Y., and Li, Y. (2012). Maximum likelihood estimation of inflation factors on error covariance matrices for ensemble Kalman filter assimilation. *Quarterly Journal of the Royal Meteorological Society*, 138(662):263–273.

Lorenz, E. N. (1996). Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1. Reading.

Qiu, Y. and Chen, S. X. (2015). Bandwidth selection for high-dimensional covariance matrix estimation. *Journal of the American Statistical Association*, 110(511):1160–1174.

Talagrand, O. (1997). Assimilation of observations, an introduction. *Journal of the Meteorological Society of Japan. Ser. II*, 75(1B):191–209.

van Kekem, D. L. (2018). Dynamics of the lorenz-96 model: Bifurcations, symmetries and waves. *Rijksuniversiteit Groningen*.

Reference II

Wang, X. and Bishop, C. H. (2003). A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *Journal of the atmospheric sciences*, 60(9):1140–1158.

Wu, G., Zheng, X., Wang, L., Zhang, S., Liang, X., and Li, Y. (2013). A new structure for error covariance matrices and their adaptive estimation in enkf assimilation. *Quarterly Journal of the Royal Meteorological Society*, 139(672):795–804.

Zheng, X. (2009). An adaptive estimation of forecast error covariance parameters for Kalman filtering data assimilation. *Advances in Atmospheric Sciences*, 26:154–160.

Zheng, X., Wu, G., Zhang, S., Liang, X., Dai, Y., and Li, Y. (2013). Using analysis state to construct a forecast error covariance matrix in ensemble Kalman filter assimilation. *Advances in Atmospheric Sciences*, 30:1303–1312.

Thanks!

Paper: Hao-Xuan Sun, Shouxia Wang, Xiaogu Zheng and Song Xi Chen. High-dimensional Ensemble Kalman Filter with Localization, Inflation and Iterative Updates. Under Minor revision at *Quarterly Journal of the Royal Meteorological Society (QJRMS)*

Homepage: `sun-haoxuan.github.io`

Contact: Hao-Xuan Sun (`hxsun@pku.edu.cn`)

Feel free to contact me!