Noname manuscript No.
(will be inserted by the editor)

Efficient Parallel Implementation of DDDAS Inference using an Ensemble Kalman Filter with Shrinkage Covariance Matrix Estimation

Elias D. Nino-Ruiz · Adrian Sandu

Received: date / Accepted: date

Abstract This paper develops an efficient and parallel implementation of dynamically data-driven application systems (DDDAS) inference using an ensemble Kalman filter based on shrinkage covariance matrix estimation. The proposed implementation works as follows: each model component is surrounded by a local box of radius size r and then, local assimilation steps are carried out in parallel at the different local boxes. Once local analyses are obtained, they are mapped back onto the global domain from which the global analysis state is obtained. Local background error correlations are estimated using the Rao-Blackwell Ledoit and Wolf estimator in order to mitigate the impact of spurious correlations whenever the number of local model components is larger than the ensemble size. The numerical Atmospheric General Circulation Model (SPEEDY) is utilized for the numerical experiments with the T-63 resolution on the Blueridge cluster at Virginia Tech. The number of processors ranges from 96 to 2,048. The proposed implementation outperforms in terms of accuracy the well-known local ensemble transform Kalman filter (LETKF) for all the model variables. The computational time of the proposed implementation is similar to that of the parallel LETKF method (where no covariance estimation is performed) for the largest number of processors.

Elias D. Nino-Ruiz
Department of Computer Science
Universidad del Norte
Barranquilla, ATL, Colombia
Tel.: 57-5-3509268
Fax: 57-5-350-9268
E-mail: enino@uninorte.edu.co

Adrian Sandu
Computational Science Laboratory
Department of Computer Science
Virginia Tech, Blacksburg, VA 24060, USA
Tel.: 1-540-231-2193
Fax: 1-540-231-9218
E-mail: asandu7@vt.edu

Keywords Parallel EnKF · Shrinkage Covariance Matrix Estimation · Sampling Errors

1 Introduction

Dynamically data-driven application systems (DDDAS [**?** **?**]) is a paradigm whereby simulations and measurements become a symbiotic feedback control system. An important application of DDDAS is the solution of inference problems where information from physical measurements is combined with a mathematical model to obtain estimates of the state or parameters of a physical system. Algorithms to solve such DDDAS inference problems are known as "data assimilation" methodologies [**?** **?** **?** **?** **?** **?** **?** **?**].

In sequential data assimilation, the goal is to estimate the current state $\mathbf{x}^* \in \mathbb{R}^{n\times 1}$ [**?**] of a system which (approximately) evolves according to some numerical model operator,

$$\mathbf{x}^*_{\text{current}} = \mathcal{M}_{t_{\text{previous}}\rightarrow t_{\text{current}}}\left(\mathbf{x}_{\text{previous}}\right)$$

where, for instance, $\mathcal{M}$ mimics the behaviour of the ocean or the atmosphere, and n denotes the number of model components. The estimation is performed based on a prior estimate $\mathbf{x}^b \in \mathbb{R}^{n\times 1}$ of $\mathbf{x}^*$,

$$\mathbf{x} \sim \mathcal{N}\left(\mathbf{x}^b, \mathbf{B}\right) , \tag{1}$$

and the noisy observation,

$$\mathbf{y} = \mathbf{H}\cdot\mathbf{x}^* + \boldsymbol{\epsilon} \in \mathbb{R}^{m\times 1} ,$$

where m is the number of observed components from the vector state $\mathbf{x}^*$, $\mathbf{B} \in \mathbb{R}^{n\times n}$ is the unknown background error covariance matrix, and $\mathbf{H} \in \mathbb{R}^{m\times n}$ is a linear observation operator. Likewise, $\boldsymbol{\epsilon} \in \mathbb{R}^{m\times 1}$ follows a Normal distribution with moments,

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}_m, \mathbf{R}\right) ,$$

where $\mathbf{0}_m$ is the m-th dimensional vector whose components are all zeros, and $\mathbf{R} \in \mathbb{R}^{m\times m}$ is the estimated data error covariance matrix. Consider the three dimensional variational (3D-Var) cost function [**?**],

$$\mathcal{J}(\mathbf{x}) = \frac{1}{2}\cdot\left\|\mathbf{x}-\mathbf{x}^b\right\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\cdot\left\|\mathbf{y}-\mathbf{H}\cdot\mathbf{x}\right\|^2_{\mathbf{R}^{-1}} , \tag{2}$$

which is nothing but the negative log of a posterior error distribution when prior and observational errors are Normal distributed. A better estimate $\mathbf{x}^a \in \mathbb{R}^{n\times 1}$ of $\mathbf{x}^*$ can be sought via the minimization of (2),

$$\mathbf{x}^a = \arg\min_{\mathbf{x}} \mathcal{J}(\mathbf{x}) , \tag{3}$$

where $\mathbf{x}^a$ is well-known as the analysis state.

In ensemble based methods, the moments of the background error distribution are estimated via the empirical moments of an ensemble of model realizations. However, several challenges are present during this estimation process: model dimensions are in the order of millions which make impractical the direct minimization of the cost function (2) [**?**], since the dimension of the model state is several times the ensemble size, background error correlations are poorly captured by the ensemble members [**? ?**] and therefore, analysis corrections are impacted by spurious correlations, besides, high resolution numerical grids make mandatory the use of high performance computing in the context of operational data assimilation where, typically, domain decomposition is performed [**? ? ? ?**]. In general, ensemble based methods overcome these situations by considering local analyses during the assimilation process [**?**], however, for subdomain sizes larger than the ensemble size, sampling errors impact the local analysis increments, which is typical when sparse observational networks are involved during the assimilation process. As an alternative to local analyses and covariance matrix localization [**?**], ensemble methods based on shrinkage covariance matrix estimation [**? ?**] can be used in order to neglect the impact of spurious correlations at the same time that other ensemble capabilities are exploited. Since the main motivation of shrinkage covariance matrix estimation is to estimate covariance matrices of high-dimensional Normal distributions based on a few samples [**? ? ?**], this estimator can be exploited in the context of operational data assimilation where covariance matrix of high-dimensional background error distributions are estimated based on a few model realizations.

This paper is organized as follows: Section 2 discusses efficient implementations of ensemble based methods and shrinkage covariance matrix estimation, in Section 3, the proposed implementation is formulated, Section 4 presents experimental results for the Atmospheric General Circulation Model (SPEEDY) and, in Section 5, the conclusions are stated.

2 Preliminaries

In this section, we briefly discuss some relevant concepts for the formulation of our proposed ensemble Kalman filter implementation.

2.1 Shrinkage covariance matrix estimation

Consider S samples from the distribution,

$$\mathbf{c}^{[s]} \sim \mathcal{N}\left(\mathbf{0}_p, \mathbf{M}\right) ,$$

for $1 \leq s \leq S$, where p is the dimension of the underlying probability distribution, and $\mathbf{M} \in \mathbb{R}^{p \times p}$ is a covariance matrix. These samples can be stored column-wise in a matrix as follows,

$$\mathbf{C} = \left[\mathbf{c}^{[1]}, \mathbf{c}^{[2]}, \ldots, \mathbf{c}^{[S]}\right] \in \mathbb{R}^{p \times S} . \tag{4}$$

The empirical covariance matrix,

$$\widetilde{\mathbf{M}} = \frac{1}{S-1} \cdot \widehat{\mathbf{C}} \cdot \widehat{\mathbf{C}}^T \in \mathbb{R}^{p \times p},$$

where

$$\widehat{\mathbf{C}} = \mathbf{C} - \bar{\mathbf{c}} \cdot \mathbf{1}_S^T \in \mathbb{R}^{p \times S},$$

with

$$\bar{\mathbf{c}} = \frac{1}{S} \cdot \sum_{s=1}^{S} \mathbf{c}^{[s]} \in \mathbb{R}^{p \times 1},$$

can be used in order to estimate $\mathbf{M} \approx \widetilde{\mathbf{M}}$. However, for cases in which $p \gg S$, $\widetilde{\mathbf{M}}$ is a low-rank estimate of $\mathbf{M}$ and therefore, spurious correlations owing to sampling errors are induced. A better estimate of $\mathbf{M}$ based on the samples (4) can be obtained making use of estimates based on shrinkage covariance matrices [**?** **?** **?**]. This estimator has the form,

$$\widehat{\mathbf{M}} = \lambda \cdot [\eta \cdot \mathbf{I}_p] + (1 - \lambda) \cdot \widetilde{\mathbf{M}}. \tag{5}$$

where $[\eta \cdot \mathbf{I}_p]$ is known as the target matrix, $\eta > 0$, and $\lambda \in (0, 1)$. Thus, shrinkage covariance matrix estimators are nothing but convex combinations of some target matrices and sample covariance matrices. The optimal choice of λ comes as the solution of the optimization problem,

$$\lambda^* = \arg\min_{\lambda} \mathbb{E}\left[\left\|\mathbf{M} - \widehat{\mathbf{M}}\right\|^2\right].$$

To the best of our knowledge, the best choice, under Gaussian assumptions on the samples, is proposed by the Rao-Blackwell Ledoit and Wolf estimator [**?**], it suggests a closed-form λ^* as follows,

$$\lambda^* = \min\left(1, \frac{\frac{S-2}{p} \cdot \mathrm{tr}\left(\widetilde{\mathbf{M}}^2\right) + \mathrm{tr}\left(\widetilde{\mathbf{M}}\right)^2}{(S+2) \cdot \left[\mathrm{tr}\left(\widetilde{\mathbf{M}}^2\right) - \frac{\mathrm{tr}\left(\widetilde{\mathbf{M}}\right)^2}{p}\right]}\right).$$

When λ^* is plugged in (5), the resulting estimator is well-conditioned [**?**] and even more,

$$\mathbb{E}\left[\left\|\mathbf{M} - \widehat{\mathbf{M}}\right\|^2\right] < \mathbb{E}\left[\left\|\mathbf{M} - \widetilde{\mathbf{M}}\right\|^2\right].$$

2.2 Ensemble Kalman filter

In the ensemble Kalman filter, an ensemble of model realizations,

$$\mathbf{X}^b = \left[\mathbf{x}^{b[1]}, \mathbf{x}^{b[2]}, \ldots, \mathbf{x}^{b[N]}\right] \in \mathbb{R}^{n\times N}\,, \tag{6}$$

is built in order to estimate the moments of the background error distribution via the empirical moments of the ensemble (6),

$$\mathbf{x}^b \approx \overline{\mathbf{x}}^b = \frac{1}{N}\cdot\sum_{i=1}^{N}\mathbf{x}^{b[i]} \in \mathbb{R}^{n\times 1}\,, \tag{7a}$$

and

$$\mathbf{B} \approx \mathbf{P}^b = \frac{1}{N-1}\cdot\mathbf{S}\cdot\mathbf{S}^T \in \mathbb{R}^{n\times n}\,, \tag{7b}$$

where N is the ensemble size, $\overline{\mathbf{x}}^b$ is the ensemble mean, $\mathbf{P}^b$ is the empirical background error covariance matrix, and the matrix of member deviations $\mathbf{S} \in \mathbb{R}^{n\times N}$ is given by,

$$\mathbf{S} = \mathbf{X}^b - \overline{\mathbf{x}}^b\cdot\mathbf{1}_N^T\,,$$

where $\mathbf{1}_N \in \mathbb{R}^{n\times 1}$ denotes the N dimensional vector whose components are all ones. When an observation $\mathbf{y} \in \mathbb{R}^{m\times 1}$ is available, the analysis ensemble can be computed as follows,

$$\mathbf{X}^a = \mathbf{X}^b + \mathbf{P}^b\cdot\mathbf{H}^T\cdot\left[\mathbf{R} + \mathbf{H}\cdot\mathbf{P}^b\cdot\mathbf{H}\right]^{-1}\cdot\mathbf{D} \in \mathbb{R}^{n\times N}\,, \tag{8}$$

where the matrix of innovations on the observations $\mathbf{D} \in \mathbb{R}^{m\times N}$ is obtained by,

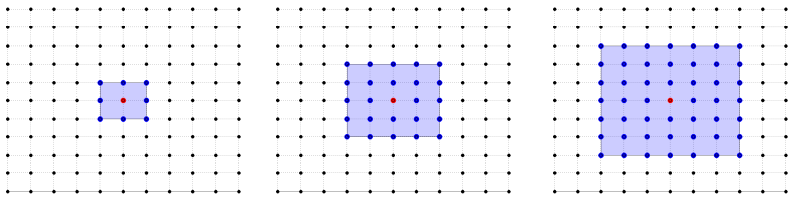
$$\mathbf{D} = \mathbf{Y}^s - \mathbf{H}\cdot\mathbf{X}^b\,,$$

and the i-th column of the matrix of perturbed observations $\mathbf{Y}^s \in \mathbb{R}^{m\times N}$ reads,

$$\mathbf{y}^{[i]} \sim \mathcal{N}\left(\mathbf{y},\,\mathbf{R}\right)\,,\ \text{for } 1 \le i \le N\,.$$

However, since ensemble sizes are much lower than model resolutions, $\mathbf{P}^b$ is a low-rank estimator of $\mathbf{B}$ and therefore, it is sensitive to sampling noise. Efficient EnKF implementations such as the local ensemble transform Kalman filter (LETKF) [**?**] overcome this situation by making use of domain localization, this is, for each model component, a radius r is considered and a local box is built, then, local information (observed components, local background error correlations, etc) is utilized in order to perform the local assimilation. After this, all local analysis components are mapped back onto the global domain from which the global analysis is obtained. Figure 1 shows some local boxes for

different radii of influence. Notice, local analyses can be computed simultaneously at different processors which makes practical the parallel implementation of the analysis step.



(a) $r = 1$ (b) $r = 2$ (c) $r = 3$

Fig. 1: Local boxes for different radii of influence r about the red model component. Blue model components denote components within the scope of r.

The global LETKF analysis equations are detailed below:

1. Compute the analysis covariance matrix in the ensemble space,
$$\widehat{\mathbf{A}} = \left[(N-1)\cdot \mathbf{I}_N + \mathbf{Q}^T \cdot \mathbf{R}^{-1} \cdot \mathbf{Q}\right]^{-1} \in \mathbb{R}^{N\times N},$$
where $\mathbf{I}_N$ is the identity matrix in the ensemble space, $\mathbf{Q} = \mathbf{H}\cdot\mathbf{U} \in \mathbb{R}^{m\times N}$, and $\mathbf{U} = \sqrt{N-1}\cdot\mathbf{S} \in \mathbb{R}^{n\times N}$.
2. Calculate the analysis mean by the optimal increments from the ensemble space,
$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^b + \mathbf{U}\cdot\left[\widehat{\mathbf{A}}\cdot\mathbf{Q}^T\cdot\mathbf{R}^{-1}\cdot\left(\mathbf{y}-\mathbf{H}\cdot\overline{\mathbf{x}}^b\right)\right].$$
3. Generate the analysis ensemble,
$$\mathbf{X}^a = \overline{\mathbf{x}}^a\cdot\mathbf{1}_N^T + \mathbf{U}\cdot\left[(N-1)\cdot\widehat{\mathbf{A}}\right]^{1/2}.$$

Note that, in the context of LETKF, background error correlations are estimated based on the ensemble covariance matrix, whenever the number of local components is larger than the ensemble size, local background error covariance matrices are obtained and therefore, local analysis can be impacted by spurious correlations. We think, there is an opportunity in order to avoid this situation by estimating local background error correlations via shrinkage covariance matrix estimators, which have proven to work under realistic model scenarios [**?** **?**]. For instance, the ensemble Kalman filter based on the Rao-Blackwell Ledoit and Wolf estimator [**?** **?**] performs the global assimilation as follows,

1. Compute the ensemble covariance matrix (7b).
2. Compute the traces,
$$\alpha = \mathrm{tr}\left(\left[\mathbf{P}^b\right]^2\right), \text{ and, } \beta = \left[\mathrm{tr}\left(\mathbf{P}^b\right)\right]^2. \tag{9a}$$

3. Set

$$\mu = \frac{\alpha}{n}, \text{ and, } \gamma = \min\left(1, \frac{\frac{N-2}{n}\cdot\alpha + \beta}{(N+2)\cdot\left[\alpha - \frac{\beta}{n}\right]}\right)$$

4. Build the shrinkage covariance matrix,

$$\widehat{\mathbf{B}} = \gamma\cdot\mathbf{I}_n + (1-\gamma)\cdot\mathbf{P}^b \in \mathbb{R}^{n\times n}. \tag{9b}$$

5. Compute the analysis ensemble,

$$\mathbf{X}^a = \mathbf{X}^b + \widehat{\mathbf{B}}\cdot\mathbf{H}^T\cdot\left[\mathbf{R} + \mathbf{H}\cdot\widehat{\mathbf{B}}\cdot\mathbf{H}^T\right]^{-1}\cdot\mathbf{D}\,. \tag{9c}$$

Efficient implementations of this filter wherein the explicit computation of the equations (9) are avoided can be seen in [**?**]. For instance, it is enough to note that, for (9a),

$$\mathrm{tr}\left(\left[\mathbf{P}^b\right]^2\right) = \sum_{i=1}^{N-1}\sigma_i^4\,, \text{ and } ,\left[\mathrm{tr}\left(\mathbf{P}^b\right)\right]^2 = \left[\sum_{i=1}^{N-1}\sigma_i^2\right]^2$$

where σ_i is the i-th singular value of $\mathbf{S}$, for $1 \leq i \leq N$. Recall that, $\mathbf{P}^b = \mathbf{S}\cdot\mathbf{S}^T$.

Now, we are ready to present an efficient implementation of the ensemble Kalman filter based on shrinkage covariance matrix estimation.

3 Proposed Implementation

We consider the use of shrinkage covariance matrix estimation in order to estimate the background error correlations during the assimilation step of the EnKF and even more, the use of domain decomposition in order to propose a parallel implementation of the EnKF based on shrinkage covariance matrix estimation. The covariance is estimated via the Rao-Blackwell Ledoit and Wolf estimator [**?**] which, under Gaussian assumptions on background errors, provides better asymptotic properties than the Ledoit and Wolf estimator presented in [**?**]. The analysis step of the parallel ensemble Kalman filter based on shrinkage covariance matrix estimation (EnKF-SC) proceed as follows,

1. Following the idea of LETKF, local boxes of radius size r are built for each model component. The k-th local box, for $1 \leq k \leq n$, is formed by n_k model components and m_k observed components. In such box, $\mathbf{X}_k^b \in \mathbb{R}^{n_k\times 1}$ denotes the background ensemble, $\mathbf{y}_k \in \mathbb{R}^{m_k\times 1}$ is the local observation with data error covariance matrix $\mathbf{R}_k \in \mathbb{R}^{m_k\times m_k}$, and $\mathbf{D}_k \in \mathbb{R}^{m_k\times N}$ is the local innovation matrix. Each local box is potentially mapped to an unique processor.

2. Compute the local ensemble perturbation matrix,

$$\mathbf{S}_k = \frac{1}{\sqrt{N-1}} \cdot \left[\mathbf{X}_k^b - \overline{\mathbf{x}}_k^b \cdot \mathbf{1}_N^T\right] \in \mathbb{R}^{n_k \cdot N}$$

where $\overline{\mathbf{x}}_k^b \in \mathbb{R}^{n_k \times 1}$ is the local ensemble mean,

$$\overline{\mathbf{x}}_k^b = \frac{1}{N} \cdot \mathbf{X}_k^b \cdot \mathbf{1}_N \, .$$

3. Compute the $N-1$ singular values $\sigma_j^{(k)}$ of $\mathbf{S}_k$, for $1 \leq j \leq N-1$.
4. Compute,

$$\alpha_k = \sum_{i=1}^{N-1} \left[\sigma_i^{(k)}\right]^4 \, , \text{ and } \, , \beta_k = \left[\sum_{i=1}^{N-1} \left[\sigma_i^{(k)}\right]^2\right]^2$$

5. Set,

$$\mu_k = \frac{\alpha_k}{n}, \text{ and, } \gamma_k = \min\left(1, \frac{\frac{N-2}{n} \cdot \alpha_k + \beta_k}{(N+2) \cdot \left[\alpha_k - \frac{\beta_k}{n}\right]}\right)$$

6. Set $\varphi_k = \mu_k \cdot \gamma_k$ and $\delta_k = 1 - \gamma_k$ and compute the local analysis ensemble,

$$\mathbf{X}_k^a = \mathbf{X}_k^b + \mathbf{E}_k \cdot \boldsymbol{\Pi}_k \cdot \mathbf{Z}_k + \varphi \cdot \mathbf{H}_k^T \cdot \mathbf{Z}_k$$

where

$$\mathbf{E}_k = \sqrt{\delta_k} \cdot \mathbf{S}_k \in \mathbb{R}^{n_k \times N},$$
$$\boldsymbol{\Pi}_k = \mathbf{H}_k \cdot \mathbf{E}_k \in \mathbb{R}^{m_k \times N} \, ,$$

$\mathbf{Z}_k$ is given by the solution of the next linear system,

$$\left[\boldsymbol{\Gamma}_k + \boldsymbol{\Pi}_k \cdot \boldsymbol{\Pi}_k^T\right] \cdot \mathbf{Z}_k = \mathbf{D}_k \, , \tag{10}$$

with

$$\boldsymbol{\Gamma}_k = \mathbf{R}_k + \varphi_k \cdot \mathbf{H}_k \cdot \mathbf{H}_k^T \in \mathbb{R}^{m_k \times m_k} \, .$$

The linear system (10) can be solved making use of the iterative Sherman Morrison formula [**?** **?**] and therefore, direct matrix inversion can be avoided.

It can be easily shown that, when covariance inflation [**?**] is utilized in the context of EnKF-SC, for the k-th local box, the following effect on $\widehat{\mathbf{B}}$ is noted,

$$\widehat{\mathbf{B}}_k = \varphi \cdot \mathbf{I}_{n_k} + \left[\rho^2 \cdot \delta_k\right] \cdot \mathbf{S}_k \cdot \mathbf{S}_k^T \in \mathbb{R}^{n_k \cdot n_k} \, .$$

where $\rho > 1$ is the inflation factor.

4 Experimental Results

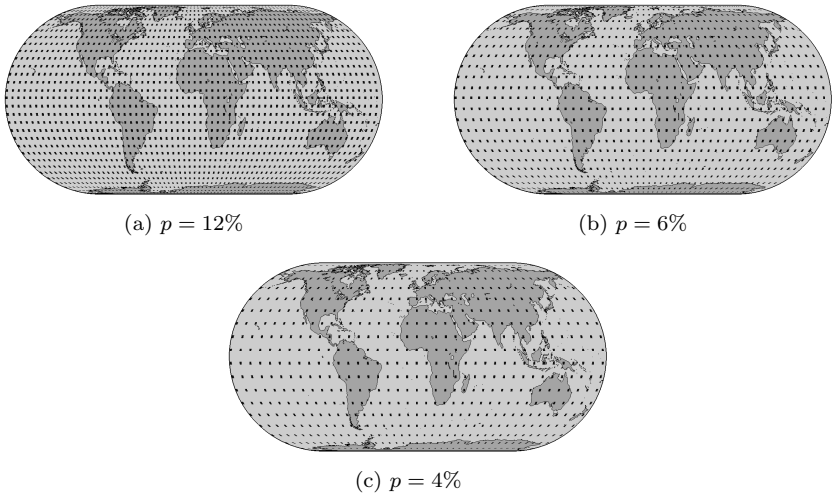
The proposed EnKF implementation is compared against the LETKF formulation in terms of accuracy and parallel performance. The tests are performed on the super computer Blueridge cluster at the university of Virginia Tech. BlueRidge is a 408-node Cray CS-300 cluster. Each node is outfitted with two octa-core Intel Sandy Bridge CPUs and 64 GB of memory, for a total of 6,528 cores and 27.3 TB of memory systemwide. Eighteen nodes have 128 GB of memory. In addition, 130 nodes are outfitted with two Intel MIC (Xeon Phi) coprocessors. Each of the 260 coprocessors on BlueRidge has 60 1.05 GHz cores and a theoretical peak performance of approximately 1 TeraFlop (double-precision) per second. Each computing node has a total of 16 processors.

For the model operator, the Atmospheric General Circulation Model (SPEEDY) [**?** **?**] is utilized, and the T-63 model resolution (96×192 grid components per layer) is considered. SPEEDY is based on a spectral dynamical core. It is a hydrostatic, s-coordinate, spectral-transform model in the vorticity-divergence form, with semi-implicit treatment of gravity waves. The model variables are the Specific Humidity $sph(g/Kg)$, the Temperature (T), the Zonal Wind Component u (m/sg), the Meridional Wind Component v (m/sg). The total number of model components is 589,824.

- Starting in rest, the SPEEDY model is run for about three months from which an initial state consistent with the physics and dynamics of the model is obtained. We let this state to be the reference solution.
- The reference solution is perturbed with a random vector $\boldsymbol{\nu} \in \mathbb{R}^{n\times 1}$ with statistics,

$$\boldsymbol{\nu} \sim \mathcal{N}\left(\mathbf{0}_n,\, 0.05 \cdot \mathbf{I}_n\right) , \tag{11}$$

  from which an initial perturbed background state is obtained. This state is propagated for about three months from which the initial background state is obtained.
- Making use of the initial background state, an initial perturbed ensemble is built. Samples from the distribution (11) are taken in order to create the synthetic members. Of course, these members are not consistent with the physics and the dynamics of the numerical model and therefore, they are propagated for about three days in order to make them consistent. From here, the initial ensemble is obtained. The reference solution and the background state are then propagated in time until the initial time.
- Three different observational networks are utilized during the experiments, they are shown in Figure 2 with their respective percentage of observed components. The percentage of observed components are 12%, 6%, and 4%.
- The number of ensemble members for all experiments is $N = 96$. Notice, the model resolution is approximately 6,144 times larger than the ensemble size.



(a) $p = 12\%$ (b) $p = 6\%$

(c) $p = 4\%$

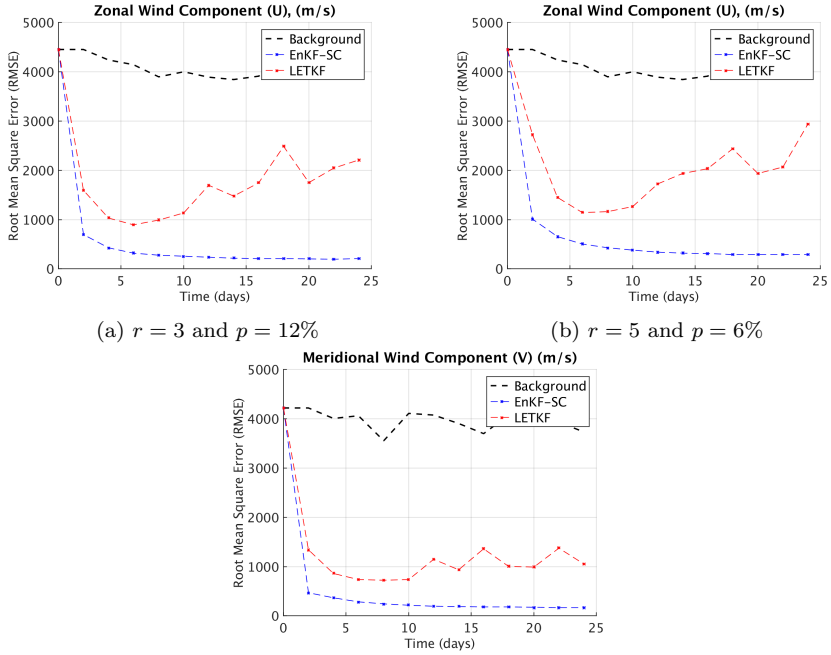
Fig. 2: Observational networks for the assimilation steps. p denotes the percentage of observed components from the model state.

- Three radius sizes are utilized during the experiments: $r = 3$, 4, and 5.
- Data errors are assumed uncorrelated and to follow a Normal distribution with zero mean and data error covariance matrix $\mathbf{R} = 0.01^2 \cdot \mathbf{I}_m$.
- Observations are taken every three days. The assimilation window contain twelve observations distributed uniformly.
- To assess the accuracy of the proposed implementation, the results are compared against those obtained by the LETKF [**?** **?**] in terms of Root-Mean-Square-Error,

$$RMSE = \sqrt{\frac{1}{12} \cdot \sum_{t=1}^{12} \left\| \mathbf{x}^*_{[t]} - \overline{\mathbf{x}}^a_{[t]} \right\|^2},$$

  where t denotes time index, for $1 \leq t \leq 12$.
- Numerical computations are performed making use of the BLAS [**?**] and the LAPACK [**?**] libraries.
- The parallel implementations were carried out making use of FORTRAN and the MPI framework.
- The number of computing nodes is ranged from 16 (96 processors) to 128 (2,048)

The results for all model configurations and variables are shown in the Table 1. As can be seen, all RMSE values of the proposed implementation are below to those obtained by the LETKF. This is expected since for large radius sizes, spurious correlations impact the estimation of local analysis corrections in the context of LETKF at the different local boxes. We can analyse Figure 3 where the RMSE of some model variables are shown for the LETKF and



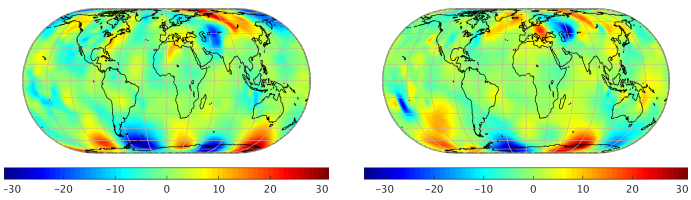
(a) $r = 3$ and $p = 12\%$

(b) $r = 5$ and $p = 6\%$

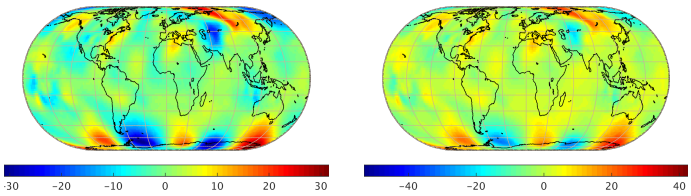
(c) $r = 4$ and $p = 12\%$

Fig. 3: RMSE of the LETKF and EnKF-MC implementations for different model variables, radii of influence and observational networks.

the EnKF-SC within the assimilation window. As can be seen, the parallel EnKF-SC is able to obtain good estimates of the background error correlations among different assimilation steps while the opposite case is evident for the LETKF, which obeys to spurious correlations in the local analysis corrections. Note that, in Figure 3, the behaviour of errors for different values of p and r in the EnKF-SC are similar, which implies the importance of estimating background error correlations based on shrinkage covariance matrix estimation. For instance, consider Figure 4 where snapshots of the first assimilation step are taken. Notice, LETKF is able to recover the structure of the meridional wind components regarding the reference state of the system. However, contour levels are far from those observed in the reference solution. On the other hand, the parallel EnKF-SC is able to recover the contour fields of the reference solution and even more, the values of the meridional wind components are close to those observed in the actual solution. Notice, spurious waves near the poles are quickly dissipated by the proposed method while this is not the case for the LETKF implementation. A similar case can be observed for the zonal wind component in Figure 5. Spurious zonal wind components near the poles are dissipated by EnKF-SC while those are held by the LETKF after the first assimilation step.



(a) Reference (b) Background

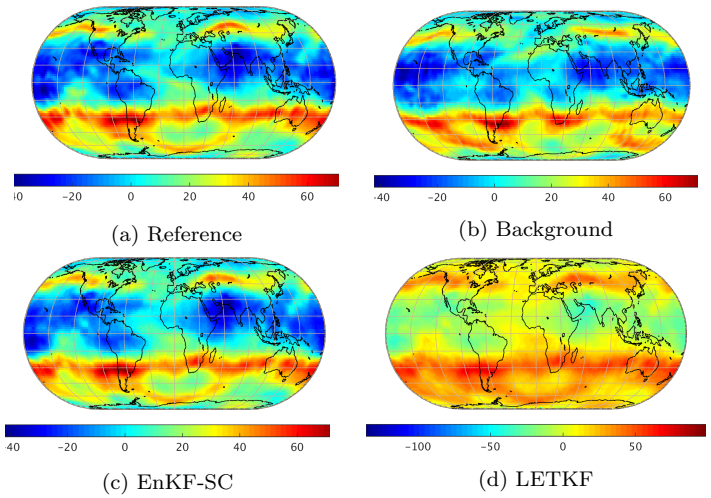


(c) EnKF-SC (d) LETKF

Fig. 4: Snapshots of the reference solution, background state and analyses fields from the EnKF-SC and LETKF for the 5th layer of the meridional wind component (v).

A very important concern is how the accuracy of the EnKF-SC is impacted when the number of processors is changed. In Figure 6, this issue is addressed. As can be seen, for each model variable, all solutions obtained by the EnKF-SC are almost identical when different number of processors are utilized. The small differences obey to the synthetic data generated at each local box. The random seed for each local box is fed by the processor id and therefore, when more processors are added, new random numbers are generated at some local boxes and therefore, the synthetic data is not replicated. Hence, small differences in the RMSE values can be shown. On the other hand, since the LETKF is a deterministic filter, all results are equal.

Lastly, Figure 7 shows the averaged time consumed for the compared implementations in order to perform a single assimilation step. As expected, since no covariance matrix estimation is performed in the LETKF context, this method is faster than the parallel EnKF-SC but, as the number of processors is increased the gap between the two curves is decreased.



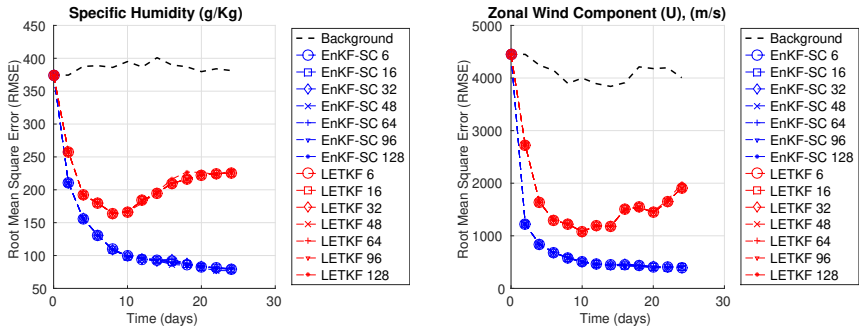
(a) Reference (b) Background

(c) EnKF-SC (d) LETKF

Fig. 5: Snapshots of the reference solution, background state and analyses fields from the EnKF-SC and LETKF for the second layer of the zonal wind component (u).

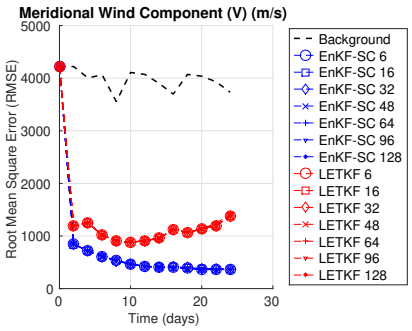
Variable (units)	r	p	EnKF-SC	LETKF
$u\ (m/sg)$	3	12%	3.185×10^2	1.661×10^3
		6%	4.854×10^2	1.237×10^3
		4%	6.750×10^2	9.997×10^2
	4	12%	3.161×10^2	1.752×10^3
		6%	4.729×10^2	1.608×10^3
		4%	6.261×10^2	1.258×10^3
	5	12%	3.334×10^2	1.861×10^3
		6%	4.729×10^2	1.983×10^3
		4%	6.148×10^2	1.601×10^3
$v\ (m/sg)$	3	12%	2.463×10^2	9.510×10^2
		6%	3.844×10^2	8.334×10^2
		4%	5.615×10^2	7.455×10^2
	4	12%	2.513×10^2	1.048×10^3
		6%	3.786×10^2	1.146×10^3
		4%	5.189×10^2	9.026×10^2
	5	12%	2.729×10^2	1.001×10^3
		6%	3.871×10^2	1.574×10^3
		4%	5.139×10^2	1.102×10^3
$T\ (K)$	3	12%	2.728×10^2	1.216×10^3
		6%	3.308×10^2	6.458×10^2
		4%	3.966×10^2	6.073×10^2
	4	12%	2.576×10^2	1.816×10^3
		6%	3.097×10^2	1.030×10^3
		4%	3.664×10^2	7.464×10^2
	5	12%	2.561×10^2	1.600×10^3
		6%	3.015×10^2	1.472×10^3
		4%	3.511×10^2	1.171×10^3
$sh\ (g/kg)$	3	12%	7.750×10	1.340×10^2
		6%	9.813×10	1.417×10^2
		4%	1.222×10^2	1.457×10^2
	4	12%	7.762×10	1.639×10^2
		6%	9.692×10	1.652×10^2
		4%	1.167×10^2	1.739×10^2
	5	12%	8.094×10	2.077×10^2
		6%	9.775×10	1.949×10^2
		4%	1.158×10^2	2.068×10^2

Table 1: RMSE values for the EnKF-SC and LETKF making use of the SPEEDY model.



(a) $p = 12\%$

(b) $p = 6\%$



(c) $p = 4\%$

Fig. 6: RMSE values for different number of computing nodes (x 16 processors) for the compared implementations. There is no significant difference between the RMSE values obtained by the parallel EnKF-SC for different number of processors.

5 Conclusions

This paper proposes an efficient parallel implementation of the ensemble Kalman filter based on shrinkage covariance matrix estimation. The proposed implementation exploits the use of shrinkage covariance matrix estimation in order to mitigate the impact of spurious correlations. Even more, well-known capabilities of ensemble based methods are exploited in order to perform in parallel the assimilation step. Numerical experiments are performed making use of the Atmospheric General Circulation Model (SPEEDY) with resolution T-63 for a total number of 589,824 model components while the number of ensemble members is set to 96 for all the experiments. The number of processors for the parallel assimilation step is ranged from 96 to 2,048. Experimental results reveal that the use of shrinkage covariance matrix estimation can mitigate the impact of spurious correlation when sparse observational networks are utilized and large local boxes are considered. Even more, in terms of accuracy, the proposed method outperforms the well-known LETKF implementation for all

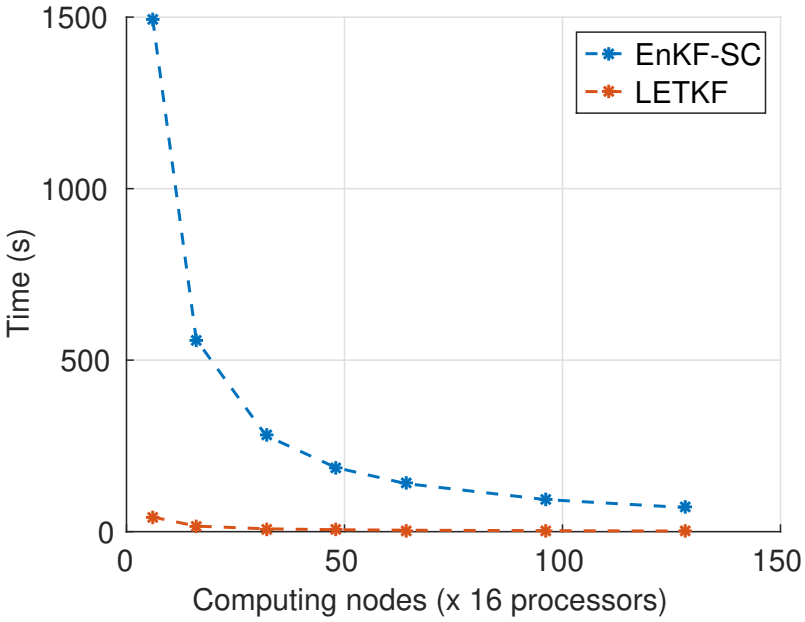


Fig. 7: Average time of the parallel EnKF-SC and the LETKF implementations for the assimilation step using different number of processors. As the number of processors increases, the differences between elapsed times of the compared implementations decreases.

model variables. Lastly, when the number of processors increases, the elapsed time of EnKF-SC and LETKF (where no covariance estimation is performed) get closer.

Acknowledgements

This work was supported in part by awards NSF CCF–1218454, AFOSR FA9550–12–1–0293–DEF, and by the Computational Science Laboratory at Virginia Tech.

References

. Jeffrey L. Anderson and Stephen L. Anderson. A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. Monthly Weather Review, 127(12):2741–2758, 1999.
. E. Anderson, Z. Bai, J. Dongarra, A. Greenbaum, A. McKenney, J. Du Croz, S. Hammerling, J. Demmel, C. Bischof,

and D. Sorensen. LAPACK: A Portable Linear Algebra Library for High-performance Computers. In Proceedings of the 1990 ACM/IEEE Conference on Supercomputing, Supercomputing '90, pages 2–11, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

. A. Aved, F. Darema, and E. Blasch. Dynamic data driven application systems. `www.1dddas.org`, 2014.

. Jeffrey L. Anderson. Localization and Sampling Error Correction in Ensemble Kalman Filter Data Assimilation. Monthly Weather Review, 140(7):2359–2371, 2012.

. L. S. Blackford, J. Demmel, J. Dongarra, I. Duff, S. Hammarling, G. Henry, M. Heroux, L. Kaufman, A. Lumsdaine, A. Petitet, R. Pozo, K. Remington, and R. C. Whaley. An Updated Set of Basic Linear Algebra Subprograms (BLAS). ACM Transactions on Mathematical Software, 28:135–151, 2001.

. E. Blasch, G. Seetharaman, and K. Reinhardt. Dynamic data driven applications system concept for information fusion. Procedia Computer Science, 18(0):1999 – 2007, 2013. 2013 International Conference on Computational Science.

. Haiyan Cheng, Mohamed Jardak, Mihai Alexe, and Adrian Sandu. A Hybrid Approach to Estimating Error Covariances in Variational Data Assimilation. Tellus A, 62(3):288–297, 2010.

. Haiyan Cheng, Mohamed Jardak, Mihai Alexe, and Adrian Sandu. A Hybrid Approach to Estimating Error Covariances in Variational Data Assimilation. Tellus A, 62(3):288–297, March 2010.

. Romain Couillet and Matthew McKay. Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. Journal of Multivariate Analysis, 131:99–120, 2014.

. Yilun Chen, A Wiesel, Y.C. Eldar, and AO. Hero. Shrinkage Algorithms for MMSE Covariance Estimation. Signal Processing, IEEE Transactions on, 58(10):5016–5029, Oct 2010.

. Xiaohui Chen, Z.J. Wang, and M.J. McKeown. Shrinkage-to-Tapering Estimation of Large Covariance Matrices. Signal Processing, IEEE Transactions on, 60(11):5640–5656, Nov 2012.

. Michael J Daniels and Robert E Kass. Shrinkage estimators for covariance matrices. Biometrics, 57(4):1173–1184, 2001.

. Geir Evensen. EnKF-The Ensemble Kalman Filter. http://enkf.nersc.no/. Accessed: 04-24-2015.

. Geir Evensen. Data Assimilation: The Ensemble Kalman Filter. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

. HumbertoC. Godinez and J.David Moulton. An Efficient Matrix-free Algorithm for the Ensemble Kalman Filter. Computational Geosciences, 16(3):565–575, 2012.

. Poterjoy Jonathan, Zhang Fuqing, and Yonghui Weng. The Effects of Sampling Errors on the EnKF Assimilation of Inner-Core Hurricane Observations. Monthly Weather Review, 142(4):1609–

1630, 2014.

. Christian L. Keppenne. Data Assimilation into a Primitive-Equation Model with a Parallel Ensemble Kalman Filter. Monthly Weather Review, 128(6):1971–1981, 2000.

. Fred Kucharski, Franco Molteni, and Annalisa Bracco. Decadal interactions between the western tropical pacific and the north atlantic oscillation. Climate Dynamics, 26(1):79–91, 2006.

. A. C. Lorenc. Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society, 112(474):1177–1194, 1986.

. Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. UPF economics and business working paper, (691), 2003.

. Olivier Ledoit and Michael Wolf. A Well-conditioned Estimator for Large-dimensional Covariance Matrices. Journal of Multivariate Analysis, 88(2):365 – 411, 2004.

. F. Molteni. Atmospheric simulations using a gcm with simplified physical parametrizations. i: model climatology and variability in multi-decadal experiments. Climate Dynamics, 20(2-3):175–191, 2003.

. Elias D Nino-Ruiz and Adrian Sandu. An efficient parallel implementation of the ensemble kalman filter based on shrinkage covariance matrix estimation. In Proceedings of the 2015 IEEE 22nd International Conference on High Performance Computing Workshops (HiPCW), pages 54–54. IEEE Computer Society, 2015.

. Elias D Nino-Ruiz and Adrian Sandu. Ensemble kalman filter implementations based on shrinkage covariance matrix estimation. Ocean Dynamics, 65(11):1423–1439, 2015.

. EliasD. Nino Ruiz, Adrian Sandu, and Jeffrey Anderson. An Efficient Implementation of the Ensemble Kalman Filter Based on an Iterative Sherman–Morrison Formula. Statistics and Computing, pages 1–17, 2014.

. Edward Ott, Brian R. Hunt, Istvan Szunyogh, Aleksey V. Zimin, Eric J. Kostelich, Matteo Corazza, Eugenia Kalnay, D. J. Patil, and James A. Yorke. A local ensemble kalman filter for atmospheric data assimilation. Tellus A, 56(5):415–428, 2004.

. Edward Ott, Brian Hunt, Istvan Szunyogh, Aleksey V Zimin, Eic J. Kostelich, Matteo Corazza, Eugenia Kalnay, D. J. Patil, and James A. Yorke. A Local Ensemble Transform Kalman Filter Data Assimilation System for the NCEP Global Model. Tellus A, 60(1):113–130, 2008.

. Cosmin G Petra, Victor M Zavala, Elias D Nino-Ruiz, and Mihai Anitescu. A high-performance computing framework for analyzing the economic impacts of wind correlation. Electric Power Systems Research, 141:372–380, 2016.

. V. Rao and A. Sandu. A posteriori error estimates for DDDAS inference problems. In International Conference on Computational Science (ICCS-2014), volume 29, pages 1256–1265, 2014.
. V. Rao and A. Sandu. A posteriori error estimates for inverse problems. SIAM/ASA Journal on Uncertainty Quantification, 3(1):737–761, 2015.
. Pavel Sakov and Laurent Bertino. Relation between two common localisation methods for the enkf. Computational Geosciences, 15(2):225–237, 2011.
. A. Sandu, E.M. Constantinescu, G.R. Carmichael, T. Chai, D. Daescu, and J.H. Seinfeld. Ensemble methods for dynamic data assimilation of chemical observations in atmospheric models. Journal of Algorithms and Computational Technology, 5(4):667–692, 2011.
. Juliane Schäfer, Korbinian Strimmer, et al. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology, 4(1):32, 2005.
. Milija Zupanski. Theoretical and Practical Issues of Ensemble Data Assimilation in Weather and Climate. In SeonK. Park and Liang Xu, editors, Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications, pages 67–84. Springer Berlin Heidelberg, 2009.