A Robust Ensemble-based Data Assimilation Method using Shrinkage Estimator and Adaptive Inflation

Santiago Lopez-Restrepo[1], Elias David Nino Ruiz[2], Andres Yarce Botero[3], Luis G Guzman-Reyes[2], O. Lucia Quintero Montoya[3], Nicolas Pinel[3], Arjo Segers[4], and Arnold Heemink[5]

[1]Delft University of Technologylogy
[2]Universidad del Norte
[3]Universidad EAFIT
[4]TNO
[5]Delft University of Technology

November 26, 2022

Abstract

This work proposes a robust and non-gaussian version of the shrinkage-based EnKF implementation, the EnKF-KA. The proposed method is based in the robust H filter and in its ensemble time-local version the EnTLHF, using an adaptive inflation factor depending on the shrinkage covariance estimated matrix. This implies a theoretical and solid background to construct robust filters from the well-known covariance inflation technique. The method is tested using the Lorenz-96 model to evaluate the robustness and performance under different scenarios as ensemble size, observation error, errors in the model specifications, and ensemble gaussianity. The results suggest good robustness of the proposed method in all the evaluated cases compared with the standard EnKF, the shrinkage-based EnKF-KA, and the robust EnTLHF.

A Robust Ensemble-based Data Assimilation Method using Shrinkage Estimator and Adaptive Inflation

Santiago Lopez-Restrepo[1,2,3], Elias D. Nino-Ruiz[4],Andres Yarce[1,2,3], Luis G. Guzman-Reyes[4], O. L. Quintero[1], Nicolás Pinel[5], Arjo Segers[6], A. W. Heemink[2]

[1]Universidad EAFIT, Mathematical Modelling Research Group, Medellín, Colombia
[2]Department of Applied Mathematics, TU Delft, The Netherlands
[3]SimpleSpace, Medellín, Colombia
[4]Universidad del Norte, Applied Math and Computer Science Laboratory, Department of Computer Science, Barranquilla, Colombia
[5]Universidad EAFIT, Grupo de Investigación en Biodiversidad Evolución y Conservación (BEC), Departamento de Ciencias Biológicas, Medellín, Colombia
[6]Department of Climate, Air and Sustainability, TNO, The Netherlands

Key Points:

- A robust ensemble based estimation is proposed
- Adaptive inflation for the proposed filter is derived
- Theoretical development of both the robust filter and adaptation scheme is presented

Corresponding author: Santiago Lopez-Restrepo, `s.lopezrestrepo@tudelft.nl`, `slopezr2@eafit.edu.co`

Abstract

This work proposes a robust and non-gaussian version of the shrinkage-based EnKF implementation, the EnKF-KA. The proposed method is based in the robust H_∞ filter and in its ensemble time-local version the EnTLHF, using an adaptive inflation factor depending on the shrinkage covariance estimated matrix. This implies a theoretical and solid background to construct robust filters from the well-known covariance inflation technique. The method is tested using the Lorenz-96 model to evaluate the robustness and performance under different scenarios as ensemble size, observation error, errors in the model specifications, and ensemble gaussianity. The results suggest good robustness of the proposed method in all the evaluated cases compared with the standard EnKF, the shrinkage-based EnKF-KA, and the robust EnTLHF.

Plain Language Summary

Data assimilation is a mathematical process that combines two sources of information (models and observations) in an optimal way. In this work, we propose a new robust ensemble-based data assimilation algorithm that allows the user to incorporate knowledge or dynamics that are not well represented by the model. Additionally, the proposed algorithm is suitable for non-linear and non-gaussian problems with a scarce error characterization. We evaluate the algorithm using the Lorenz-96 model and compare its performance against other well-known ensemble-based data assimilation algorithm. The results show that our propose can outperform the other methods reducing the estimation error and increasing the robustness.

1 Introduction

Data assimilation (DA) is a mathematical family of methods that allows the combination of observations and models. The model is used to fill observational gaps, and the observations constrain the model dynamics (Lahoz & Schneider, 2014; Bocquet et al., 2015). In most of the DA methods, the aim is to minimize the estimated error variance. For instance, Kalman Filter (KF) is an optimal method that minimizes the mean-squared-error in the estimation. The KF is optimal when the following assumptions are fulfilled: the dynamic system is linear, and the observation and model uncertainties follow a Gaussian distribution (Kalman, 1960). The Ensemble Kalman Filter (EnKF) is a KF-based Monte carlo approximation of the KF when the state space is large, and the model is non-linear (Evensen, 2003). The EnKF uses an ensemble of model realization to approximate the first and second background error moments, making it efficient for large-scale models and suitable in the presence of non-linearities. However, in real DA applications, the assumptions required to obtain the optimal solution may not be accurate, degrading the filter performance (Houtekamer et al., 2005; Evensen, 2003). Additionally, small ensemble sizes may produce a poor approximation of the model uncertainty, causing a reduction in the filter accuracy or even filter divergence.

When the system conditions do not satisfy the KF-based methods requirement, a different approach is a robust filter or robust estimator. The robust filters emphasize the robustness of the estimation to have better tolerances to high uncertainty sources. Since its purpose is not the optimality in the estimation, the robust estimator does not require a strictly statistical representation of the system and the observations (Luo & Hoteit, 2011), showing a better performance than the KF-based methods in scenarios with a poor statistical uncertainty representation (Han et al., 2009; Nan & Wu, 2017). There are several robust ensemble-based DA schemes based in different aspect such as H_∞ formulation (Han et al., 2009), replacing the traditional L_2 norm (Roh et al., 2013; Freitag et al., 2013; Rao et al., 2017), robust covariance estimation (Yang et al., 2001; E. Nino-Ruiz et al., 2018), and covariance inflation (Luo & Hoteit, 2011; Bai et al., 2016). The approach that we propose uses a shrinkage-based covariance estimator that improves the model robustness and performance when the ensemble size is small. Additionally, our method incorporates adaptive covariance inflation closely related to the H_∞ formulation.

2 Ensemble-Based Data Assimilation

In ensemble-based data assimilation, an ensemble of model realizations

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

is employed to estimate the first ($\mathbf{x}^b$) and second moments ($\mathbf{B}$) of the background error distributions, where $\mathbf{x}^{b[i]} \in \mathbb{R}^{n\times 1}$ is the i-th ensemble member, and N is the total number of ensemble members. Hence:

$$\mathbf{x}^b \approx \overline{\mathbf{x}}^b = \frac{1}{N-1}\cdot\sum_{e=1}^{N}\mathbf{x}^{b[e]} \in \mathbb{R}^{n\times 1}\,, \tag{2}$$

and

$$\mathbf{B} \approx \mathbf{P}^b = \frac{1}{N}\cdot\mathbf{\Delta X}\cdot\mathbf{\Delta X}^T \in \mathbb{R}^{n\times n}\,, \tag{3}$$

where

$$\mathbf{\Delta X} = \mathbf{X}^b - \overline{\mathbf{x}}^b\cdot\mathbf{1}^T \in \mathbb{R}^{n\times N}\,, \tag{4}$$

is the anomalies matrix, $\overline{\mathbf{x}}^b$ is the ensemble mean, $\mathbf{P}^b$ is the sample covariance matrix, and $\mathbf{1}$ is a vector with components all ones. Once an observation is available, the posterior state can be computed via an ensemble-based method as EnKF (Evensen, 2003) or its variants, EnKS (Evensen, 2003), EnHF (Liu et al., 2008), or 4DEnVAR (Liu et al., 2008) for instance.

2.1 Shrinkage-based Ensemble Kalman Filter

A more robust family of covariance estimators for the case $n \gg N$ are the shrinkage based estimators (Touloumis, 2015; Couillet & McKay, 2014). This kind of estimators have the form (Ledoit & Wolf, 2018):

$$\mathbf{B} \approx \widehat{\mathbf{B}}(\alpha) = \alpha\cdot\mathbf{T} + (1-\alpha)\cdot\mathbf{P}^b \in \mathbb{R}^{n\times n}\,, \tag{5}$$

where $\alpha \in [0,\, 1]$, and $\mathbf{T} \in \mathbb{R}^{n\times n}$ is a user-defined matrix. The value of α is chosen to minimize

$$\alpha^* = \arg\min_{\alpha}\mathbb{E}\left[\left\|\mathbf{B} - \widehat{\mathbf{B}}(\alpha)\right\|_F^2\right]\,, \tag{6}$$

where $\|\bullet\|_F$ represents the Frobenius norm. A close formulation to calculate the weight value α using a general target matrix $\mathbf{T}_{KA}$ is proposed in (Stoica et al., 2008; Zhu et al., 2011) (hereafter KA estimator),

$$\hat{\mathbf{B}}_{KA} = \alpha_{KA}\cdot\mathbf{T}_{KA} + (1-\alpha_{KA})\cdot\mathbf{P}^b \in \mathbb{R}^{n\times n}, \tag{7a}$$

with

$$\alpha_{KA} = \min\left(\frac{\frac{1}{N^2}\cdot\sum_{i=1}^{N}\left\|\mathbf{\Delta x}^{[e]}\right\|^4 - \frac{1}{N}\cdot\left\|\mathbf{P}^{\mathbf{b}}\right\|^2}{\left\|\mathbf{P}^b - \mathbf{T}_{KA}\right\|^2}, 1\right). \tag{7b}$$

This general target matrix enables the incorporation of *prior* information about the system into the error covariance matrix. Additionally, the KA estimator does not make any distributional assumptions, thus can also be used for non-gaussian covariance matrix estimation (Zhu et al., 2011). An implementation of the EnKF can be obtained using the KA estimator, known as EnKF-KA (Lopez-Restrepo et al., 2021):

$$\mathbf{X}^a = \mathbf{X}^b + \hat{\mathbf{B}}_{KA}\cdot\mathbf{H}^T\cdot[\mathbf{R} + \mathbf{H}\cdot\hat{\mathbf{B}}_{KA}\cdot\mathbf{H}^T]\cdot\mathbf{D},$$

where $\mathbf{X}^a$ is the analysis ensemble, $\mathbf{H}$ is the linear (or linearizated) output operator, and the e-th column of the innovation matrix on the synthetic observations $\mathbf{D} \in \mathbb{R}^{n\times N}$ reads $\mathbf{d}^{[e]} = \mathbf{y}+\boldsymbol{\epsilon}^{[e]}-\mathcal{H}\left(\mathbf{x}^{b[e]}\right) \in \mathbb{R}^{m\times 1}$, with $\boldsymbol{\epsilon}^{[e]} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{R}\right)$. In Lopez-Restrepo et al.(2021), it is shown than incorporating *prior* information of the system in the data assimilation process can outperforms the EnKF when $n \gg N$, and when there are errors in the model specifications.

2.2 Ensemble time-local H$_\infty$ filter

One of the most widely used robust filter is the H_∞ Filter (HF) (Hassibi et al., 2000). The HF is based on the criterion of minimizing the supremum of the L_2 norm of the uncertainty sources (Han et al., 2009). The HF ensures that the total energy of the estimation errors, is not larger than the uncertainty energy times a factor $1/\gamma$:

$$\sum_{t=0}^{M} ||\mathbf{x}_t^t - \mathbf{x}_t^a||^2_{\mathbf{S}_t} \leq \frac{1}{\gamma}\left(||\mathbf{x}_0^t - \mathbf{x}_0^a||^2_{\mathbf{\Delta_0}^{-1}} + \sum_{t=0}^{M} ||\mathbf{u}_t||^2_{\mathbf{Q}_t^{-1}} + \sum_{t=0}^{M} ||\mathbf{v}_t||^2_{\mathbf{R}_t^{-1}}\right), \tag{8}$$

where $\mathbf{x}^t$ is the true state, $\mathbf{x}^a$ is the analysis state, $\mathbf{S}$ is a user-chosen matrix of weights, $\mathbf{u}$ and $\mathbf{v}$ are the model and observation uncertainty respectively, $\mathbf{\Delta_0}, \mathbf{Q}$ and $\mathbf{R}$ are the uncertainty weighting matrices with respect to the initial conditions, model error and observations error, and M is the data assimilation windows length (Luo & Hoteit, 2011). To solve (8), the cost function $\mathcal{J}^{\mathrm{HF}}$ is defined as:

$$\mathcal{J}^{\mathrm{HF}} = \frac{\sum_{t=0}^{M} ||\mathbf{x}_t^t - \mathbf{x}_t^a||^2_{\mathbf{S}_t}}{||\mathbf{x}_0^t - \mathbf{x}_0||^2_{\mathbf{\Delta_0}^{-1}} + \sum_{t=0}^{M} ||\mathbf{u}_t||^2_{\mathbf{Q}_t^{-1}} + \sum_{t=0}^{M} ||\mathbf{v}_t||^2_{\mathbf{R}_t^{-1}}}. \tag{9}$$

Then inequality (8) is equivalent to $\mathcal{J}^{\mathrm{HF}} \leq \frac{1}{\gamma}$. Let γ^* be the value such that:

$$\frac{1}{\gamma^*} = \underset{\{\mathbf{x}_t^a\}}{inf} \quad \underset{\mathbf{x}_0, \{\mathbf{u}_t\}, \{\mathbf{v}_t\}}{sup} \quad \mathcal{J}^{\mathrm{HF}}, t \leq M, \tag{10}$$

the optimal HF is then achieved when $\gamma = \gamma^*$. In this formulation, the evaluation of γ^* is an application of the minimax rule (Berger, 1985), a strategy that aims to provide robust estimates and is different from its Bayesian counterpart (Luo & Hoteit, 2011). An Ensemble-based HF implementation for a nonlinear DA problem is the Ensemble time-local H$_\infty$ filter (EnLTHF) proposed by (Luo & Hoteit, 2011). In the EnLTHF a local cost function is proposed:

$$\mathcal{J}_t^{\mathrm{HF}} = \frac{||\mathbf{x}_t^t - \mathbf{x}_t^a||^2_{\mathbf{S}_t}}{||\mathbf{x}_0^t - \mathbf{x}_0||^2_{\mathbf{\Delta_0}^{-1}} + ||\mathbf{u}_t||^2_{\mathbf{Q}_t^{-1}} + ||\mathbf{v}_t||^2_{\mathbf{R}_t^{-1}}}. \tag{11}$$

The local performance level γ_t satisfies:

$$\frac{1}{\gamma_t} \geq \frac{1}{\gamma_t^*} = \underset{\{\mathbf{x}_t^a\}}{inf} \quad \underset{\mathbf{x}_0, \{\mathbf{u}_t\}, \{\mathbf{v}_t\}}{sup} \quad \mathcal{J}_t^{\mathrm{HF}}, \tag{12}$$

The EnLTHF can be expressed in terms of the EnKF algorithm using the notation of (Luo & Hoteit, 2011):

$$\begin{aligned}
[\mathbf{P}_t^a, \mathbf{K}_t] &= EnKF(\mathbf{x}_t^a, \mathbf{Q}_t, \mathbf{H}), && (13a)\\
\mathbf{G_t} &= \left[\boldsymbol{I}_m - \gamma_t \cdot \mathbf{P}_t^a \cdot \mathbf{S}_t\right]^{-1} \cdot \mathbf{K}_t, && (13b)\\
\mathbf{x}_t^{a(i)} &= \mathbf{x}_t^{b(i)} + \mathbf{G_t} \cdot [\boldsymbol{y}_\mathbf{t} - \mathbf{H_t} \cdot \mathbf{x}_\mathbf{t}^{\mathbf{b(i)}} + \mathbf{v}_\mathbf{t}^\mathbf{i}], && (13c)\\
\mathbf{x}_t^a &= \left(\sum_{i=1}^{N} \mathbf{x}_t^{a(i)}\right)/N, && (13d)\\
(\mathbf{\Delta}_\mathbf{t}^\mathbf{a})^{-1} &= (\mathbf{P}_t^a)^{-1} - \gamma_t \cdot \mathbf{S}_t, && (13e)
\end{aligned}$$

subject to the constraint

$$(\mathbf{\Delta}_\mathbf{t}^\mathbf{a})^{\mathbf{-1}} = (\mathbf{P}_\mathbf{t}^\mathbf{a})^{\mathbf{-1}} - \gamma_\mathbf{t} \cdot \mathbf{S_t} \geq \mathbf{0}, \tag{13f}$$

where the operator $EnKF(\cdot,\cdot,\cdot)$ means that $\mathbf{P}_t^a$ and $\mathbf{K}_t$ are obtained through the EnKF.

3 Robust Shrinkage-based Ensemble Kalman Filter

3.1 Adaptive inflation

A particular issue with ensemble-based DA algorithms is the covariance undersampling. Undersampling leads to further problems such as the ensemble collapse to an overconfident, but incorrect state, or even filter divergence (Anderson, 2001). The covariance inflation artificially increases uncertainties in the background covariance avoiding the underestimation of uncertainties, and undersampling (Bellsky & Mitchell, 2018). The magnitude of the inflation depends to a large degree on each system and application (Houtekamer & Zhang, 2016).

In (13e), the presence of the extra term $-\gamma_t{\cdot}\mathbf{S}_t$ inflates the EnKF covariance matrix. In this way, it is possible to interpret the EnTLHF as an EnKF formulation with a specific value of inflation. This implies a theoretical and solid background to construct robust filters. Consider the case where $\mathbf{S} = \mathbf{I}_n$, that corresponds with an inflation of the analysis covariance matrix eigenvalues. To satisfy the constraint (13f), or what is equivalent, to make $(\mathbf{\Delta^a_t})^{-\mathbf{1}}$ semi-definite positive, consider the SVD decomposition of $\mathbf{P}^a_t$

$$\mathbf{P}^a_t = \mathbf{V_t} \cdot \mathbf{\Sigma_t} \cdot \mathbf{U_t}, \tag{14}$$

where $\mathbf{\Sigma_t} = \mathbf{diag}(\sigma_{\mathbf{t,1}}, ..., \sigma_{\mathbf{t,n}})$ is a diagonal matrix with all the eigenvalues of $\mathbf{P}^a_t$ in descending order, that is, $\sigma_{t,1} \geq \sigma_{t,2} \geq .... \geq \sigma_{t,n}$ and γ_t is a variable that satisfies

$$\sigma_{t,1}{}^{-1} - \gamma_t \geq 0,$$

that corresponds with

$$\gamma_t \leq \frac{1}{\sigma_{t,1}},$$

guaranteeing that $(\mathbf{\Delta^a_t})^{-\mathbf{1}}$ is semi-definite positive. It is convenient to introduce a performance level coefficient (PLC) c by defining

$$\gamma_t \leq \frac{c}{\sigma_{t,1}}. \tag{15}$$

In contrast to conventional inflation schemes, γ_t is adaptive in time even for a fixed c value, and it is directly related with the analysis covariance matrix.

3.2 EnTLHF-KA

According to sections 2.2 and 3.1, with a specific structure and inflation value, it is possible to obtain a robust version of the EnKF. Although the EnTLHF has shown to have a better performance than the EnKF in scenarios with high uncertainty (Luo & Hoteit, 2011; Altaf et al., 2013; Triantafyllou et al., 2013), the limitations of the EnKF with respect to the ensemble size and the ensemble normality distribution are inherited in its robust version. When the ensemble size is small $N << n$, sampling errors can have impact on the quality of covariances matrix estimation causing problems such as filter divergence and spurious correlations (Evensen, 2003; Houtekamer & Zhang, 2016) . Even though many localization techniques have been developed to mitigate those problems, it usually prohibits its implementation in high dimensional applications (Sakov & Bertino, 2011). The shrinkage-covariance estimator methods have shown a better performance than the classical sampling covariance matrix in scenarios with small ensemble size and non-gaussianities (Chen et al., 2009; E. D. Nino-Ruiz & Sandu, 2015, 2017; Ledoit & Wolf, 2018). We propose a robust implementation of the EnKF-KA shrinkage-based method following the principles of the EnTLHF and the adaptive inflation denoted EnTLHF-KA. The EnTLHF-KA can be obtained similarly to the EnLTHF by taking as base the EnKF-KA:

$$\begin{aligned}
\left[\hat{\mathbf{B}}^a_{KA}, \mathbf{K}_t\right] &= \text{EnKF-KA}(\mathbf{x}^a_t, \mathbf{T}_{KA}, \mathbf{H}), & \text{(16a)}\\
\mathbf{G_t} &= \left[\boldsymbol{I}_m - \gamma_t \cdot \hat{\mathbf{B}}^a_{KA} \cdot \mathbf{S}_t\right]^{-1} \cdot \mathbf{K}_t, & \text{(16b)}\\
\mathbf{x}^{a(i)}_t &= \mathbf{x}^{b(i)}_t + \mathbf{G_t} \cdot [\boldsymbol{y}_\mathbf{t} - \mathbf{H_t} \cdot \mathbf{x^{b(i)}_t} + \mathbf{v^i_t}], & \text{(16c)}\\
\mathbf{x}^a_t &= \left(\sum_{i=1}^{N} \mathbf{x}^{a(i)}_t\right) / N, & \text{(16d)}
\end{aligned}$$

where the operator EnKF-KA$(\cdot, \cdot, \cdot)$ represents the EnKF-KA shrinkage-based method (see Section 2.1). For an specific PLC, the inflation value is obtained using (15).

4 Results and discussion

4.1 Numerical experiments

The Lorenz-96 is one of the most used benchmarks for testing data assimilation algorithms. The model is highly non-linear and with a strong relationship between the states. The Lorenz-96 dynamics are described by: (Lorenz & Emanuel, 1998; Gottwald & Melbourne, 2005):

$$\frac{dx_j}{dt} = \begin{cases} (x_2 - x_{n-1}) \cdot x_n - x_1 + F & \text{for } j = 1, \\ (x_{j+1} - x_{j-2}) \cdot x_{j-1} - x_j + F & \text{for } 2 \leq j \leq n-1, \\ (x_1 - x_{n-2}) \cdot x_{n-1} - x_n + F & \text{for } j = n, \end{cases} \tag{17}$$

where n is the state number chosen as 40, and F is the external force. For consistency, periodic boundary conditions are assumed. We take the next considerations for the numerical experiments:

- The assimilation window consist of $M = 500$ observations.
- The number of observed components is $m = 20$, representing and 50% of the model components.
- The observation statistics are associated with the Gaussian distribution,

  $$\mathbf{y}_t \sim \mathcal{N}(\mathbf{H} \cdot \mathbf{x}^a_t, \rho_o^2 \cdot \mathbf{I}), \text{ for } 1 \leq t \leq M, \tag{18}$$

  where $\rho_o = 0.001$, and $\mathbf{H}$ is a linear operator that randomly chooses the m observed components.
- To avoid random fluctuations, each experiment is repeated 20 times ($L = 20$).
- We compare the performance and robustness of the EnTLHF-KA against the non-robust methods EnKF and EnKF-KA, and the robust method EnTLHF.
- We take the Root-Mean-Square-Error (RMSE) of L experiments as a measure of performance,

  $$\text{RMSE} = \frac{1}{L} \cdot \sum_{l=1}^{L} \left(\sqrt{\frac{1}{M} \cdot \sum_{t=1}^{M} \left([\mathbf{x}^*_t - \mathbf{x}^a_t]^T \cdot [\mathbf{x}^*_t - \mathbf{x}^a_t] \right)^2} \right). \tag{19}$$

- We choice a PLC value $c = 0.5$ for all the experiments, following Luo and Hoteit(2011). Other c values have been tested (not reported here), but no performance improvements were obtained.

4.2 Robustness against Ensemble members

When the state dimension is large, it is important to test the performance with relative small ensemble sizes. We evaluate both the accuracy and the robustness of the EnTLHF-KA with respect to the ensemble size. For this case we set the observation error $\delta = 1\times\ 10^{-3}$,

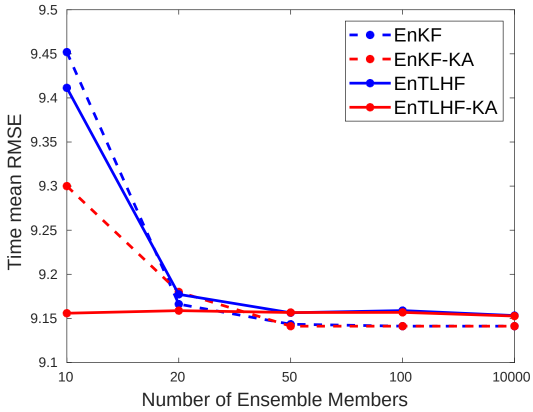


Figure 1. Error evaluation of the robust and non-robust methods respect to the ensemble member number.

the observation frequency $f = 1$, and the external force $F = 8$. The ensemble size $N \in [10, 20, 50, 100, 1000]$. Figure 1 presents the RMSE value for those values of N.

The EnTLHF-KA has more constant RMSE values for different N. The other methods present variation in its performance when the ensemble size changes. In general, the RMSE values decrease for larger N values for all the methods. For $N = 10$, the EnTLHF-KA presents a superior performance compared to the others, followed by the EnKF-KA. This behavior is attributed to the shrinkage-based estimator used in both methods, that have shown a better covariance estimation when $N << n$ (E. D. Nino-Ruiz & Sandu, 2017; Lopez-Restrepo et al., 2021). However, the adaptive inflation factor of the EnTLHF, and the ENTLHF-KA improves these methods' performance against its non-robust counterpart. For larger ensemble size, both EnTLHF-KA and EnKF-KA tend to converge to the EnTLHF and EnKF respectively, since the sampling ensemble matrix represents a good estimator for the covariance matrix and $\hat{\mathbf{B}}_{KA}$ converge to $\mathbf{P}^a$. Due to the good estimation of $\mathbf{B}$ by $\mathbf{P}^a$, and all the EnKF assumptions are satisfied, the non-robust methods present lower RMSE value for large ensemble size. This example clarifies the different advantages and disadvantages of the robust approach compared to the optimal approach. Although the EnTLHF-KA performance is not the best in all the scenarios, its robustness allows it to have low RMSE values in all the scenarios.

4.3 Robustness against observation error

Figure 2 shows the RMSE value when $\delta \in [1\times10^{-4}, 1\times10^{-3}, 1\times10^{-2}, 1\times10^{-1}]$. The other model parameters are: $N = 20$, $f = 1$, and $F = 8$. The idea now is to evaluate the impact of the observation error in the new robust EnTLHF-KA. It can be seen that the performance of the non-robust methods is affected by the increase of the observation error, causing divergence of the EnKF-KA. This kind of behavior is one of the main reasons for the development of the new robust techniques (Rao et al., 2017). The observation error's impact is much lower in the robust methods, and the performance is almost constant, especially in the EnTLHF-KA. When $\delta = 1\times10^{-4}$, the EnKF and the EnKF-KA perform better than its robust counterpart, but the robust filters hold a good performance even for large observation errors.

4.4 Robustness against model errors

To evaluate the EnTLHF-KA robustness respect to model errors, we compare the method's performance when $F \in [6, 7, 8, 9, 10]$. $F = 8$ corresponds with the assump-

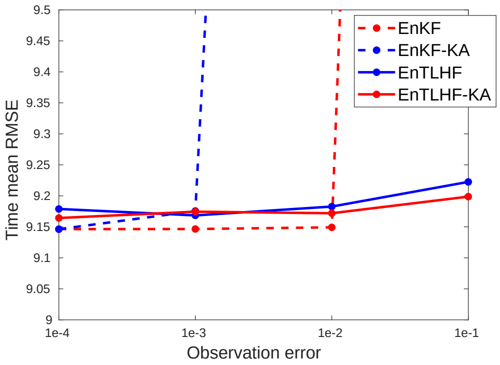


Figure 2. Error evaluation of the robust and non-robust methods respect to the observation error.

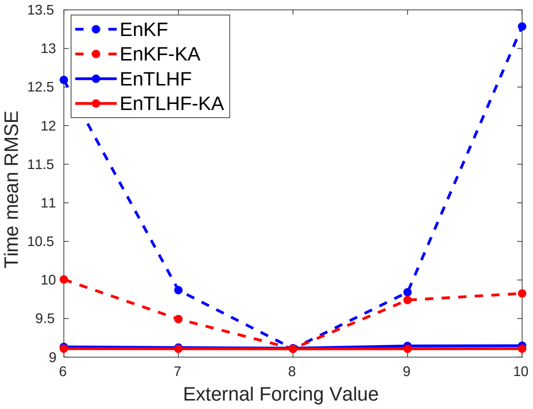


Figure 3. Error evaluation of the robust and non-robust methods respect errors in the model.

tion of a perfect model. Figure 3 presents the RMSE value for each F value and the comparison among the four filters. The RMSE values remain almost constant for both robust filters, with smaller values for the EnTLHF-KA. The adaptive inflation makes the analysis covariance matrix larger in the robust filters that in its non-robust counterpart, given the same background covariance. Consequently, the EnTLHF and the EnTLHF-KA put more weight in the observations, convenient when there are larger model errors.

4.5 Robustness against ensemble distribution

The standard EnKF assumes that the ensemble state has a Gaussian distribution. This assumption is especially essential because the state covariance $\mathbf{B}$ is approximated by the ensemble sample covariance $\mathbf{P}^b$. Although the ensemble at t_0 is Gaussian, nonlinearities in the model dynamics can modify the ensemble distribution, causing the approximation of $\mathbf{B}$ by $\mathbf{P}^b$ to lose accuracy. Figure 4 presents an evaluation of the ensemble distribution for different times steps using the Lorenz-96 model. We use the Shapiro-Wilk to evaluate the gaussianity of each state variable (Shapiro & Wilk, 1965). We take an initial Gaussian ensemble of 100 members as reference. After 15-time steps, some variables begin to change its initial distribution, and after 30-time steps, the Gaussian assumption is not valid anymore for the ensemble.

We perform different experiments varying the observation frequency or the number of time steps between two available observations. Figure 5 shows the time averaged

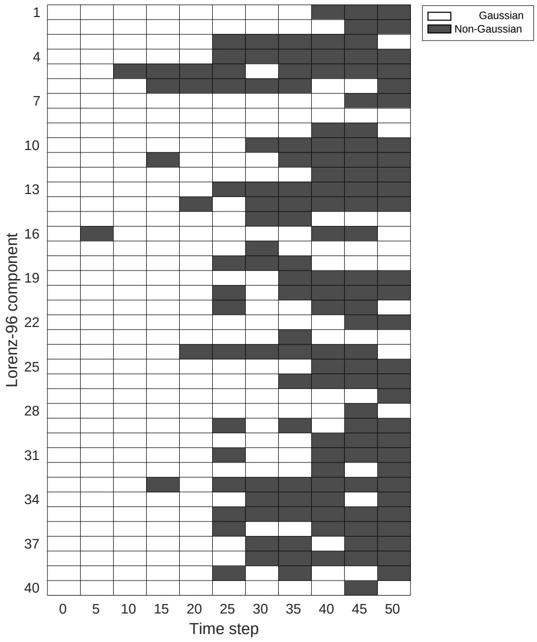


Figure 4. Shapiro-Wilk test for each Lorenz component at different time step. The ensemble size is 100. The white color represents that the null-hypothesis is not rejected (the ensemble for that specific variable is Gaussian). The grey color represents that the null-hypothesis is rejected (the ensemble for that specific variable is non-gaussian).

RMSE for the EnKF, EnKF-KA, EnTLHF and the EnTLHF-KA using a observation frequency $f \in [1, 5, 10, 20, 30, 50]$ times steps. We set an ensemble size of $N = 20$, an observation error of $\delta = 1 \times 10^{-3}$, and the external force $F = 8$. The EnKF performance decreases considerably when f increases, and after the value of $f = 30$ the method diverges. This result illustrates the importance of the Gaussian distribution for obtaining a good representation of $\mathbf{B}$ throw $\mathbf{P}^b$. The adaptive inflation increases EnTLHF robustness and performance, even when both EnKF and EnTLHF are using the same approximation of $\mathbf{B}$. Nevertheless, the EnTLHF performance decrease considerably when $f = 50$. In contrast, EnKF-KA and EnTLHF-KA use a shrinkage-based estimator for $\mathbf{B}$. The KA estimator does not assume a Gaussian distribution, as other shrinkage-based estimators do (Ledoit & Wolf, 2018; E. D. Nino-Ruiz et al., 2021). Thus, the EnKF-KA presents better performance than EnKF for large f values, and similar error levels than EnTLH without incorporating adaptive inflation. In the case of the EnTLHF-KA, the combination of both the shrinkage-based estimator and the adaptive inflation produces high robustness and performance even when the ensemble distribution is non-gaussian.

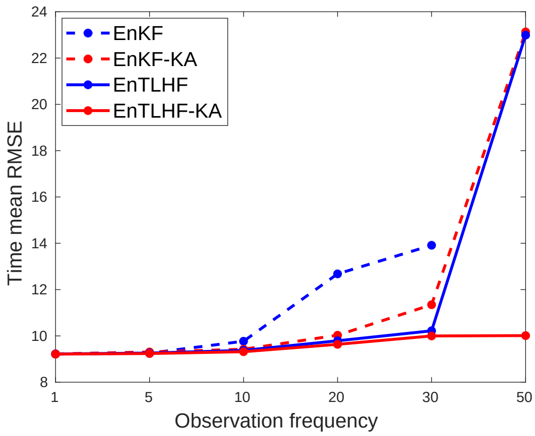


Figure 5. Error evaluation of the robust and non-robust methods respect to the observation frequency.

5 Conclusions

We propose a robust version of the shrinkage-based EnKF-KA algorithm using adaptive inflation derived from the concept of H_∞ filter (EnTLHF-KA). The EnTLHF-KA uses a covariance estimator that allows the incorporation of prior information and does not assume a Gaussian distribution in the background. Using numerical experiments, we compared the proposed method's robustness and performance against the standard EnKF, the shrinkage-based EnKF-KA, and the robust filter EnTLHF. The EnTLHF-KA has lower RMSE values in conditions with high observation error and model errors than the other methods. When the number of ensembles is small, the shrinkage estimator gives a better approximation of the background covariance matrix than the sample covariance matrix, generating lower errors in both shrinkage-based algorithm, especially in the EnTLHF-KA. The combination of the non-gaussian shrinkage estimator and the adaptive inflation grant a higher robustness to the EnTLHF-KA when the ensemble distribution is non-gaussian. All these characteristics make the EnTLHF-KA a suitable option in applications with highly non-linear models, high observation frequency, and computational restrictions in the number of ensembles.

Acknowledgments

Data availability is not applicable to this article as no new data were created or analysed in this study.

References

Altaf, M. U., Butler, T., Luo, X., Dawson, C., Mayo, T., & Hoteit, I. (2013). Improving short-range ensemble kalman storm surge forecasting using robust adaptive inflation. *Monthly Weather Review*, *141*(8), 2705–2720. doi: 10.1175/MWR-D-12-00310.1

Anderson, J. L. (2001). An ensemble adjustment kalman filter for data assimilation. *Monthly Weather Review*, *129*(12), 2884-2903. doi: 10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2

Bai, Y., Zhang, Z., Zhang, Y., & Wang, L. (2016). Inflating transform matrices to mitigate assimilation errors with robust filtering based ensemble Kalman filters. *Atmospheric Science Letters*, *17*(8), 470–478. doi: 10.1002/asl.681

Bellsky, T., & Mitchell, L. (2018). A shadowing-based inflation scheme for ensemble data assimilation. *Physica D: Nonlinear Phenomena*, *380-381*, 1–7. Retrieved from `https://doi.org/10.1016/j.physd.2018.05.002` doi: 10.1016/j.physd.2018.05.002

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* New York: Springer.

Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Aabkar, R., Carmichael, G. R., ... Seigneur, C. (2015). Data assimilation in atmospheric chemistry models: Current status and future prospects for coupled chemistry meteorology models. *Atmospheric Chemistry and Physics*, *15*(10), 5325–5358. doi: 10.5194/acp-15-5325-2015

Chen, Y., Wiesel, A., & Hero, A. O. (2009). Shrinkage estimation of high dimensional covariance matrices. In *2009 ieee international conference on acoustics, speech and signal processing* (pp. 2937–2940).

Couillet, R., & McKay, M. (2014). Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, *131*, 99–120.

Evensen, G. (2003). The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. doi: 10.1007/s10236-003-0036-9

Freitag, M. A., Nichols, N. K., & Budd, C. J. (2013). Resolution of sharp fronts in the presence of model error in variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *139*(672), 742-757. doi: 10.1002/qj.2002

Gottwald, G. A., & Melbourne, I. (2005). Testing for chaos in deterministic systems with noise. *Physica D: Nonlinear Phenomena*, *212*(1), 100 - 110. doi: https://doi.org/10.1016/j.physd.2005.09.011

Han, Y., Zhang, Y., Wang, Y., Ye, S., & Fang, H. (2009). A new sequential data assimilation method. *Science in China, Series E: Technological Sciences*, *52*(4), 1027–1038. doi: 10.1007/s11431-008-0189-3

Hassibi, B., Kailath, T., & Sayed, A. (2000). Array algorithms for h∞ estimation. *Automatic Control, IEEE*, *45*(4), 702–706. doi: 10.1109/9.847105

Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., & Hansen, B. (2005). Atmospheric data assimilation with an ensemble kalman filter: Results with real observations. *Monthly weather review*, *133*(3), 604–620.

Houtekamer, P. L., & Zhang, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *144*(12), 4489–4532. doi: 10.1175/MWR-D-15-0440.1

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, *82*(Series D), 35–45.

Lahoz, W. A., & Schneider, P. (2014). Data assimilation: Making sense of Earth Observation. *Frontiers in Environmental Science*, *2*(MAY), 1–28. doi: 10.3389/fenvs.2014.00016

Ledoit, O., & Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under stein's loss. *Bernoulli*, *24*(4B), 3791–3832.

Liu, C., Xiao, Q., & Wang, B. (2008). An ensemble-based four-dimensional variational data assimilation scheme. part i: Technical formulation and preliminary test. *Monthly Weather Review*, *136*(9), 3363–3373.

Lopez-Restrepo, S., Nino-Ruis, E. D., Yarce, A., Quintero, O. L., Pinel, N., Segers, A., & Heemink, A. W. (2021). An Efficient Ensemble Kalman Filter Implementation Via Shrinkage Covariance Matrix Estimation: Exploiting Prior Knowledge. *Computational Geosciences*, *25*, 985–1003. Retrieved from `https://doi.org/10.1007/s10596-021-10035-4`

Lorenz, E. N., & Emanuel, K. A. (1998, 02). Optimal Sites for Supplementary Weather Observations: Simulation with a Small Model. *Journal* *of the Atmospheric Sciences*, *55*(3), 399-414. Retrieved from `https://doi.org/10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2` doi: 10.1175/1520-0469(1998)055⟨0399:OSFSWO⟩2.0.CO;2

Luo, X., & Hoteit, I. (2011). Robust Ensemble Filtering and Its Relation to Covariance Inflation in the Ensemble Kalman Filter. *Monthly Weather Review*, *139*(12), 3938–3953. doi: 10.1175/MWR-D-10-05068.1

Nan, T.-c., & Wu, J.-c. (2017). Application of ensemble H-infinity filter in aquifer characterization and comparison to ensemble Kalman filter. *Water Science and Engineering*, *10*(1), 25–35. doi: 10.1016/j.wse.2017.03.009

Nino-Ruiz, E., Cheng, H., Beltran, R., Nino-Ruiz, E. D., Cheng, H., & Beltran, R. (2018). A Robust Non-Gaussian Data Assimilation Method for Highly Non-Linear Models. *Atmosphere*, *9*(4), 126. doi: 10.3390/atmos9040126

Nino-Ruiz, E. D., Guzman, L., & Jabba, D. (2021). An ensemble Kalman filter implementation based on the Ledoit and Wolf covariance matrix estimator. *Journal of Computational and Applied Mathematics*, *384*. Retrieved from `https://doi.org/10.1016/j.cam.2020.113163` doi: 10.1016/j.cam.2020.113163

Nino-Ruiz, E. D., & Sandu, A. (2015). Ensemble kalman filter implementations based on shrinkage covariance matrix estimation. *Ocean Dynamics*, *65*(11), 1423–1439.

Nino-Ruiz, E. D., & Sandu, A. (2017). Efficient parallel implementation of dddas inference using an ensemble kalman filter with shrinkage covariance matrix estimation. *Cluster Computing*, 1–11.

Rao, V., Sandu, A., Ng, M., & Nino-Ruiz, E. D. (2017). Robust data assimilation using l1 and huber norms. *SIAM Journal on Scientific Computing*, *39*(3), B548–B570. doi: 10.1137/15M1045910

Roh, S., Genton, M. G., Jun, M., Szunyogh, I., & Hoteit, I. (2013, 11). Observation Quality Control with a Robust Ensemble Kalman Filter. *Monthly Weather Review*, *141*(12), 4414-4428. doi: 10.1175/MWR-D-13-00091.1

Sakov, P., & Bertino, L. (2011). Relation between two common localisation methods for the EnKF. *Computational Geosciences*, *15*(2), 225–237. doi: 10.1007/s10596-010-9202-6

Shapiro, S., & Wilk, M. (1965, 12). An analysis of variance test for normality (complete samples)†. *Biometrika*, *52*(3-4), 591-611. Retrieved from `https://doi.org/10.1093/biomet/52.3-4.591` doi: 10.1093/biomet/52.3-4.591

Stoica, P., Li, J., Zhu, X., & Guerci, J. R. (2008). On using a priori knowledge in space-time adaptive processing. *IEEE Transactions on Signal Processing*, *56*(6), 2598–2602. doi: 10.1109/TSP.2007.914347

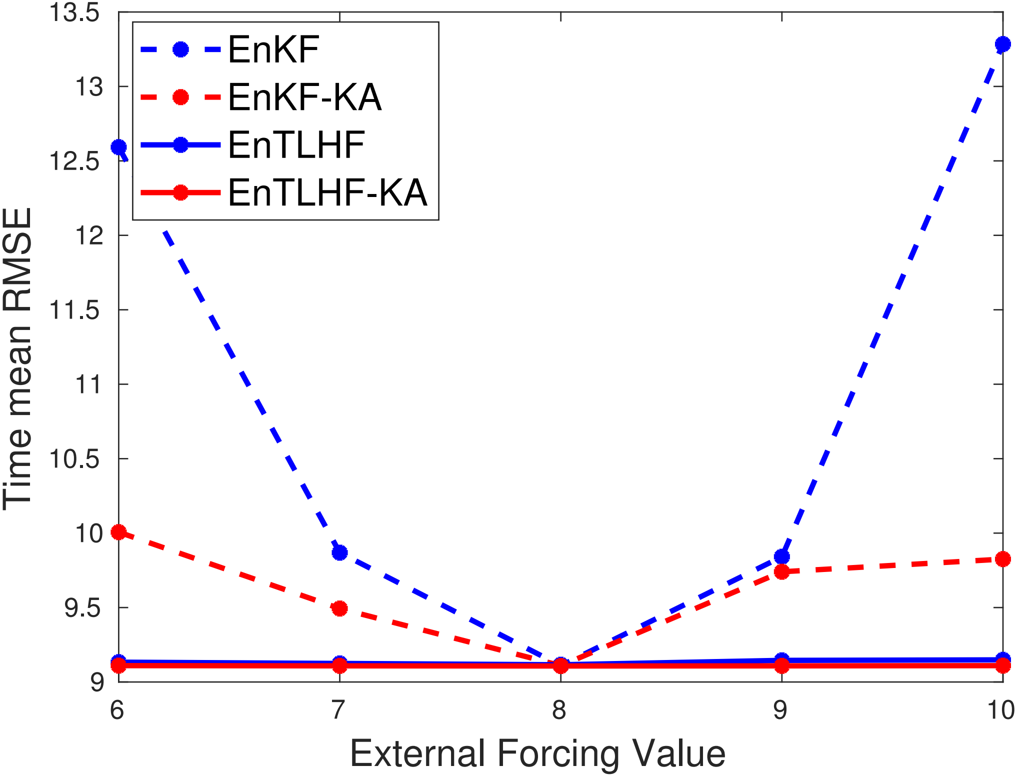
Touloumis, A. (2015). Nonparametric stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Computational Statistics & Data Analysis*, *83*, 251–261.

Triantafyllou, G., Hoteit, I., Luo, X., Tsiaras, K., & Petihakis, G. (2013). Assessing a robust ensemble-based Kalman filter for efficient ecosystem data assimilation of the Cretan Sea. *Journal of Marine Systems*, *125*, 90–100. doi: 10.1016/j.jmarsys.2012.12.006

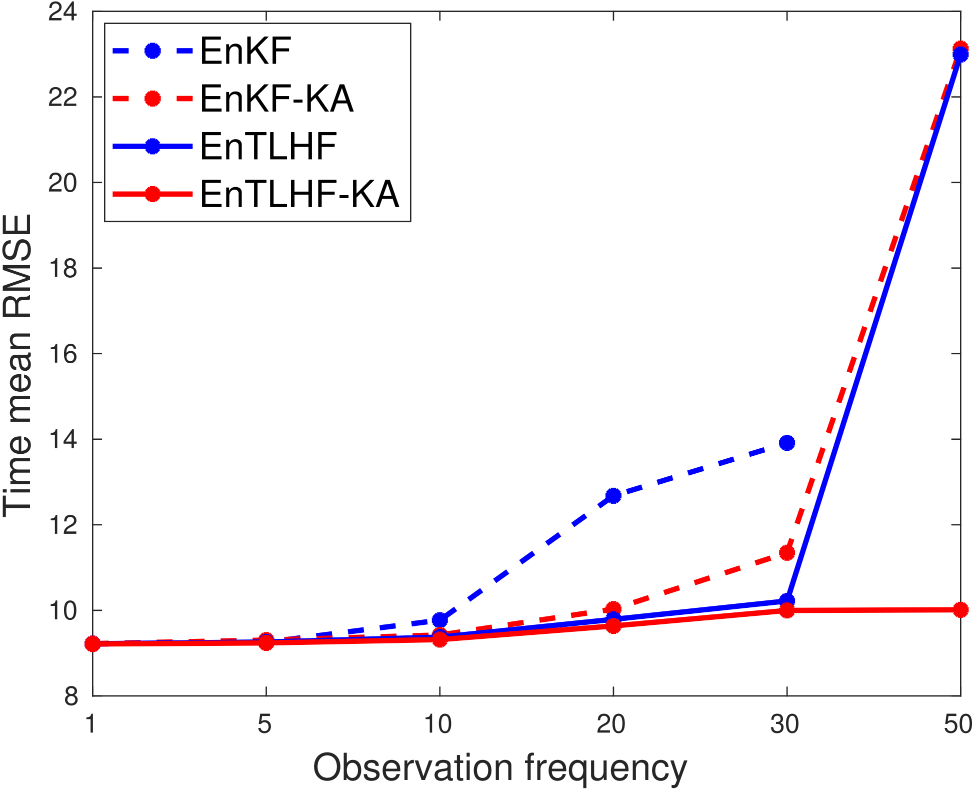
Yang, Y., He, H., & Xu, G. (2001). Adaptively robust filtering for kinematic geodetic positioning. *Journal of Geodesy*, *75*(75), 109–116.

Zhu, X., Li, J., & Stoica, P. (2011). Knowledge-Aided Space-Time Adaptive Processing. *IEEE Transaction on Aerospace And Electronic Systems*, *47*(2), 1325–1336.

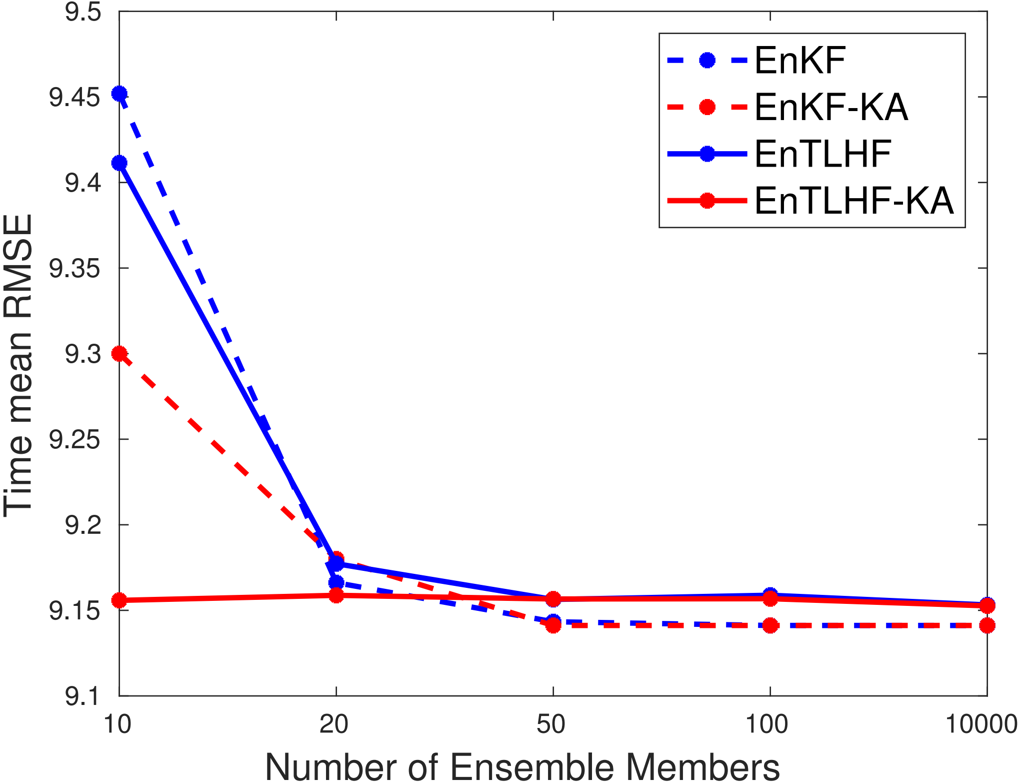
Robust_Comparison_F.


EnKF
EnKF-KA
EnTLHF
EnTLHF-KA
Time mean RMSE
External Forcing Value
13.5
13
12.5
12
11.5
11
10.5
10
9.5
9
6
7
8
9
10

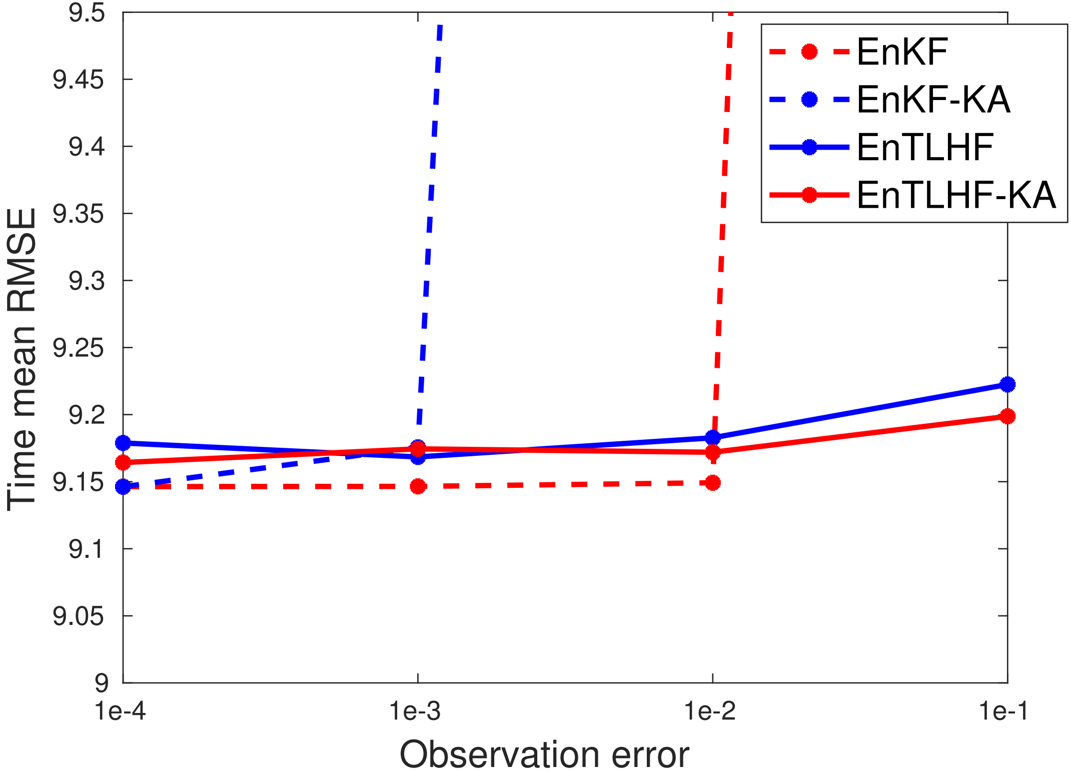
Robust_Comparison_Freq.


EnKF
EnKF-KA
EnTLHF
EnTLHF-KA
Time mean RMSE
8
10
12
14
16
18
20
22
24
1
5
10
20
30
50
Observation frequency

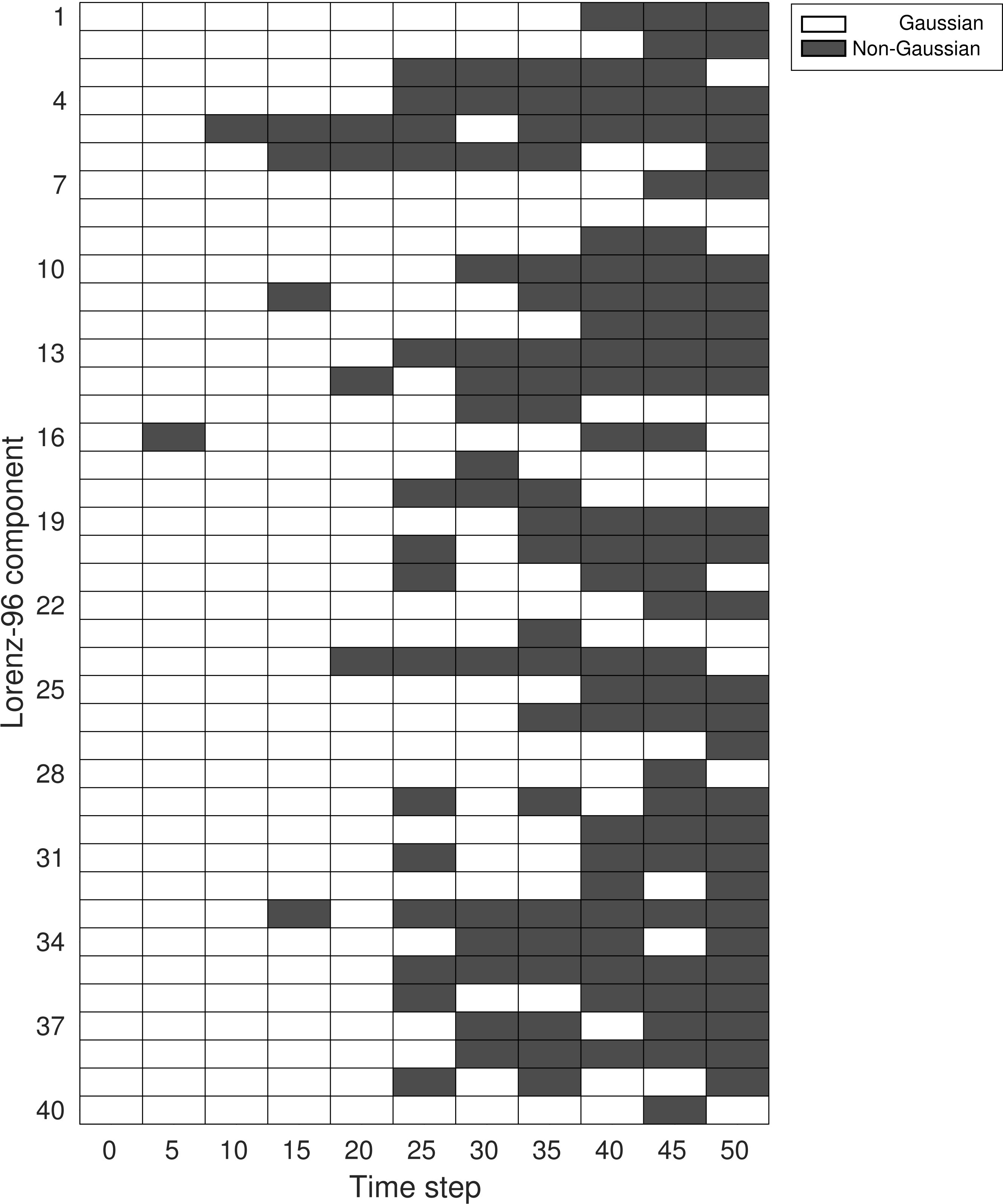
Robust_Comparison_N.


EnKF
EnKF-KA
EnTLHF
EnTLHF-KA
Time mean RMSE
9.5
9.45
9.4
9.35
9.3
9.25
9.2
9.15
9.1
10
20
50
100
10000
Number of Ensemble Members

Robust_Comparison_sigma.


9.5
9.45
9.4
9.35
9.3
9.25
9.2
9.15
9.1
9.05
9
Time mean RMSE
1e-4
1e-3
1e-2
1e-1
Observation error
EnKF
EnKF-KA
EnTLHF
EnTLHF-KA

Shapiro_Matrix.



Gaussian
Non-Gaussian
Lorenz-96 component
1
4
7
10
13
16
19
22
25
28
31
34
37
40
Time step
0
5
10
15
20
25
30
35
40
45
50