A fully probabilistic data assimilation approach for range-limited observations

A. Shah[1], M. E. Gharamti[2], L. Bertino[1]

[1]Nansen Environmental and Remote Sensing Center, Norway

[2]National Center for Atmospheric Research, USA

7th International WMO Symposium on DA

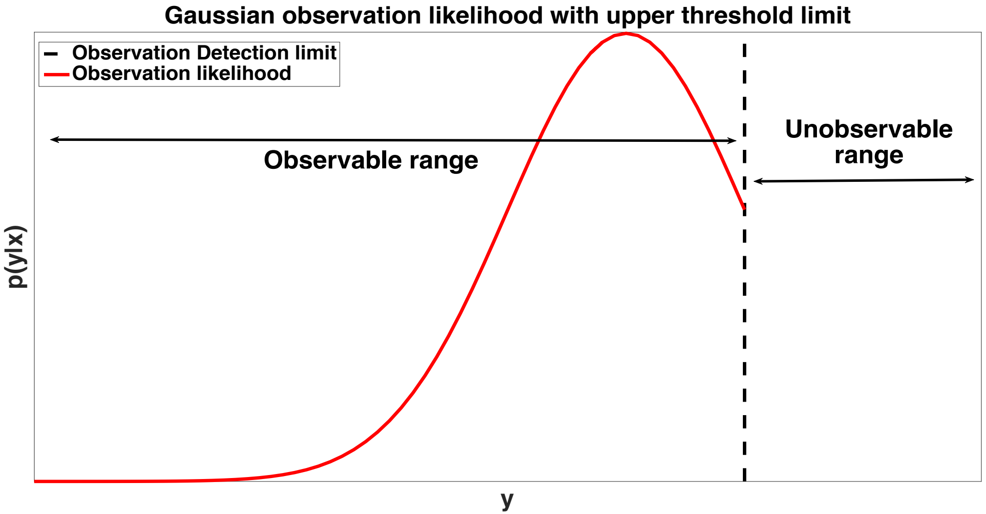
September 11th – 15th, 2017

Motivation

- Many observations in practice are available within a limited interval of actual variation due to the detection limit of the gauge or sensor.
- e.g. SMOS retrieved sea-ice thickness (upper detection limit 50 cm).
- To use the qualitative information available from the out-of-range observations (OR-observations), which were discarded as “not a number” otherwise.
- Very few studies carried out dealing OR-observations **Borup et. al. (2015)**.

Observation with detection limit

- It is a truncated observation likelihood function.



- It can be characterized as:
  - In-range observations(referred as hard data): $p(\mathbf{y_{ir}}|\mathbf{x})$
  - Out-range observations(referred as soft data): $p(\mathbf{y_{or}}|\mathbf{x})$

Methodology: Our Approach

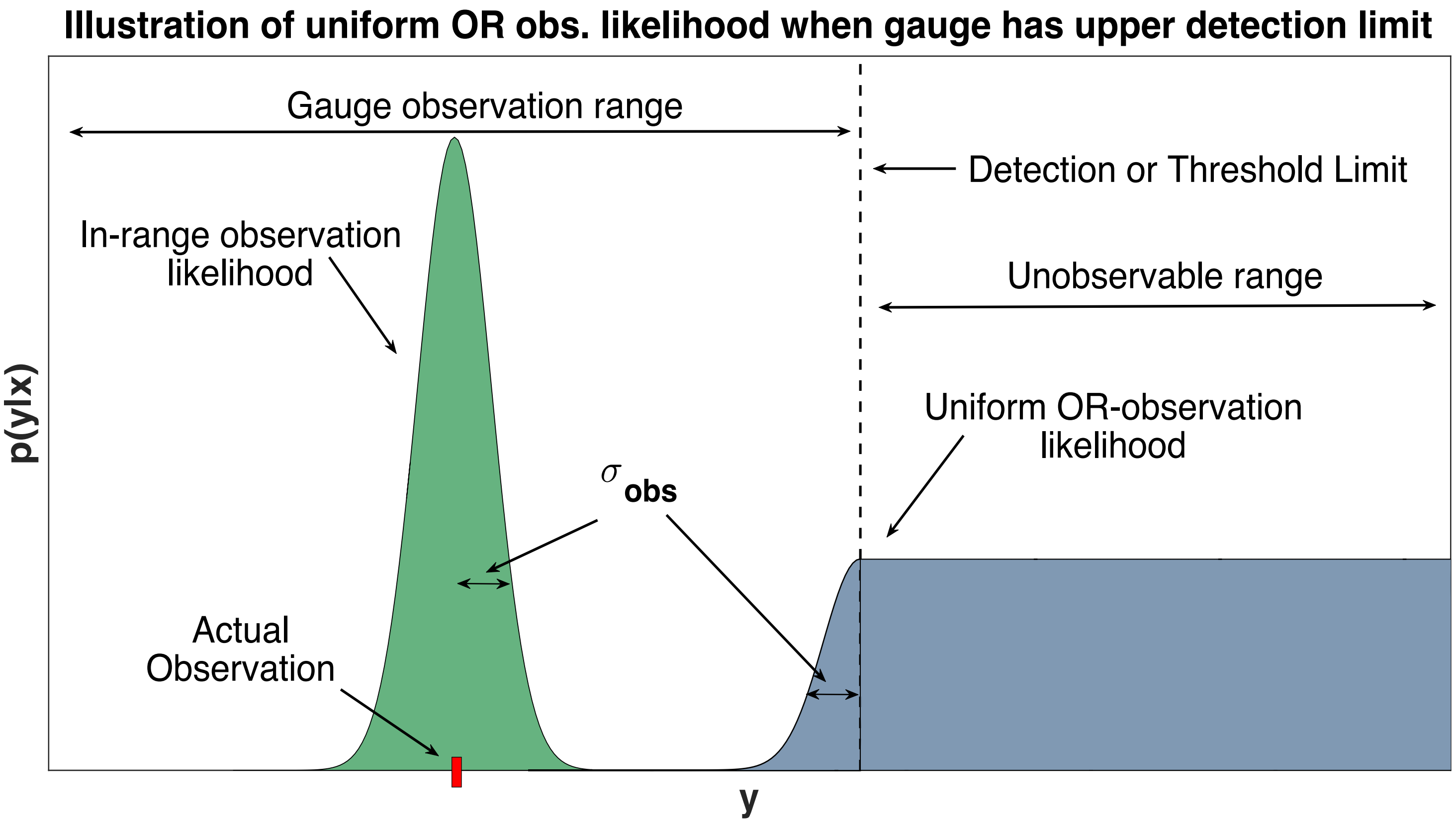
- Within the Bayesian framework, the goal of DA is to estimate the posterior distribution.
- Bayes' Rule: $$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}$$
- For an observation with a detection limit the update equation can be split into 2, depending on the nature of the observations; i.e.

$$p(\mathbf{x}|\mathbf{y}) \propto \begin{cases} p(\mathbf{y}_{ir}|\mathbf{x})p(\mathbf{x}) & \text{when } \mathbf{y}_{ir} \text{ is in range} \\ p(\mathbf{y}_{or}|\mathbf{x})p(\mathbf{x}) & \text{when } \mathbf{y}_{or} \text{ is out-of-range} \end{cases} \quad (1)$$

Ensemble members given OR-obs

- To obtain posterior estimate of individual ensemble member given OR-observations.
- To apply eq. (1) for individual member, we need an assumption about OR-observation likelihood.

- Borup et al. (2015) assumes a uniform OR-obs likelihood

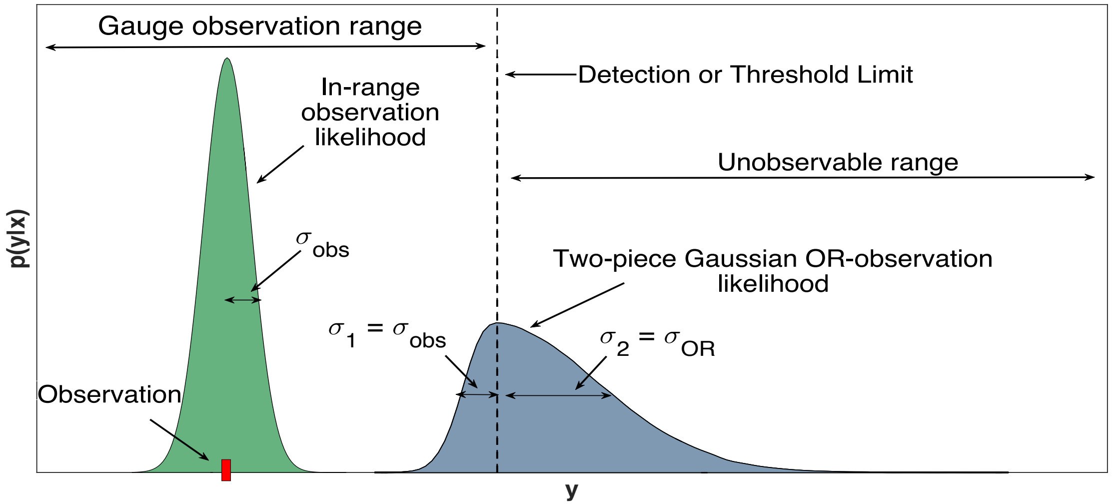


- Instead of a uniform likelihood, we assume a 2-piece Gaussian distribution (Gibbons, 1973) as a OR-obs likelihood

$$f(x) = \begin{cases} A \exp\left[-\frac{(x-\mu)^2}{2\sigma_1^2}\right] & x \leq \mu \\ A \exp\left[-\frac{(x-\mu)^2}{2\sigma_2^2}\right] & x > \mu \end{cases}$$

where $A = \sqrt{\frac{2}{\pi}}(\sigma_1 + \sigma_2)^{-1}$ is a normalizing constant

Illustration of 2-piece Gaussian OR-obs. likelihood when gauge has upper threshold limit



- Std. of the Gaussian half in the observable range is determined by the in-range observation uncertainty
- Std. of the half in the unobservable range is an arbitrary choice, which can be determined from the climatological data (if available) or by making an educated guess knowing that the extremely high values are less likely.

pdf of climatological data of observed quantity

$$\sigma_{or} = -\mu + \int_{\mu}^{\infty} y \overbrace{f_{\mathrm{CLIM}(y)}} dy$$

Why 2-piece Gaussian OR-obs likelihood?

- Closest choice to the assumption of Gaussianity in the EnKF.
- Uniform OR-obs likelihood gives equal weight to all values until its bounds, which is rarely a case in practice.

Ensemble Kalman Filter - Semi Qualitative (EnKF-SQ)

- For hard data the posterior in the Bayesian update is the product of two Gaussians: i.e., the prior and the observation likelihood.
- For the OR-obs it is the product of Gaussian prior and a 2-piece Gaussian OR-observation likelihood.

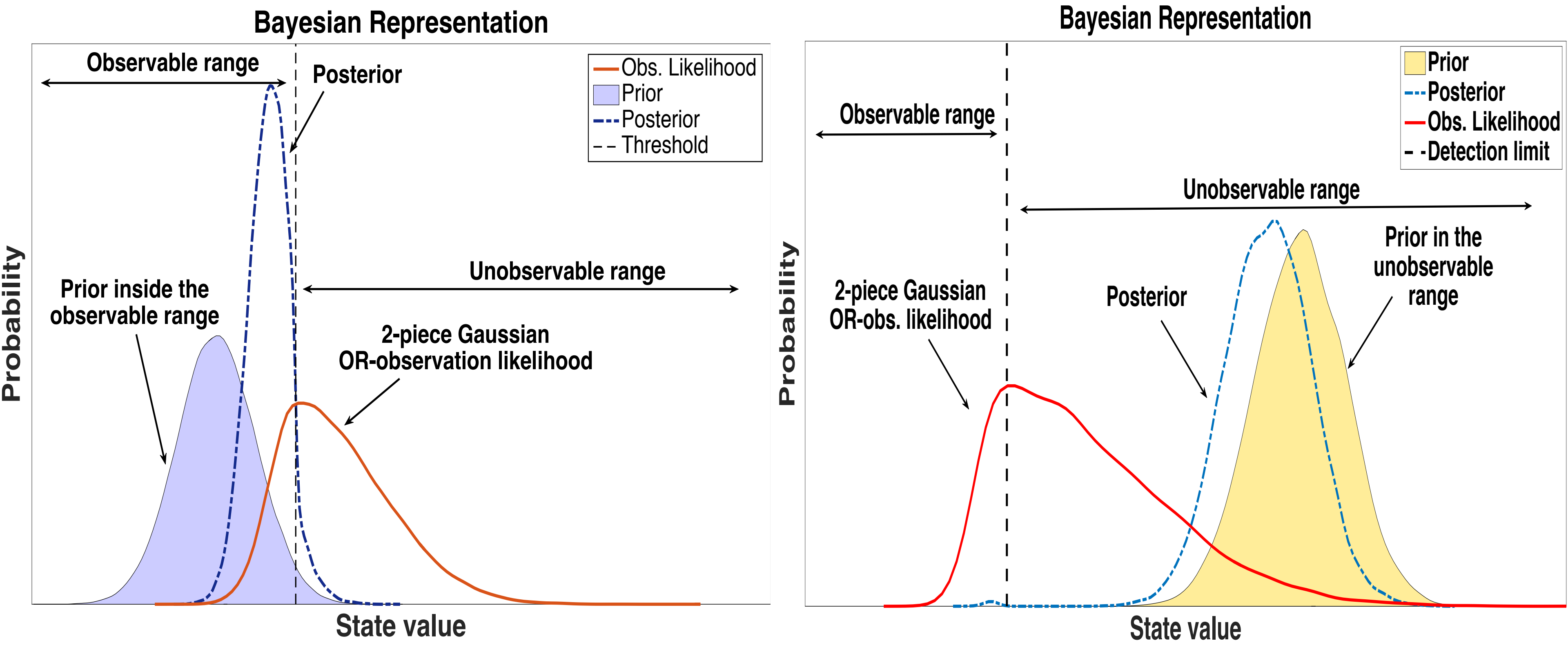
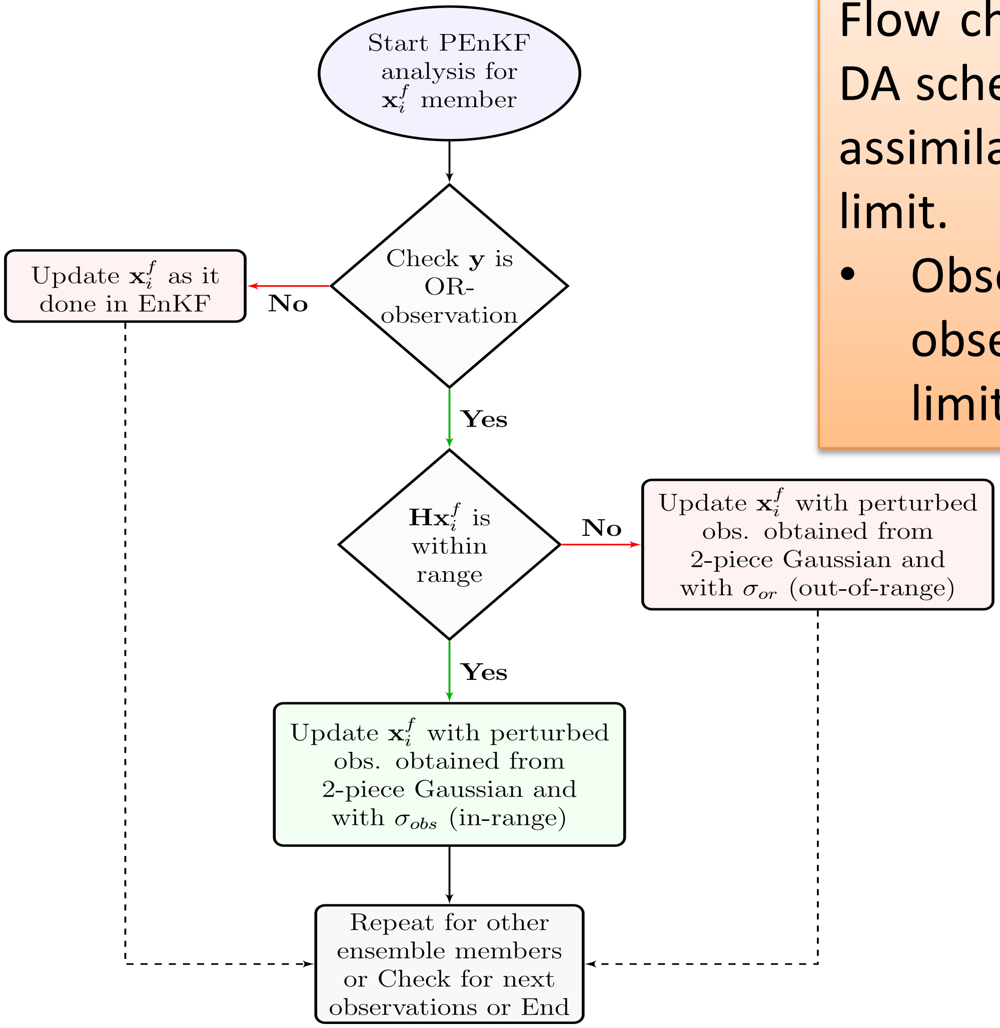


Illustration of a scenario when the prior distribution is inside (left) and outside (right) the observable range for the 2-piece Gaussian OR-observations likelihood.



Flow chart for the implementation of EnKF-SQ DA scheme under the stochastic EnKF setup for assimilating the observations with detection limit.

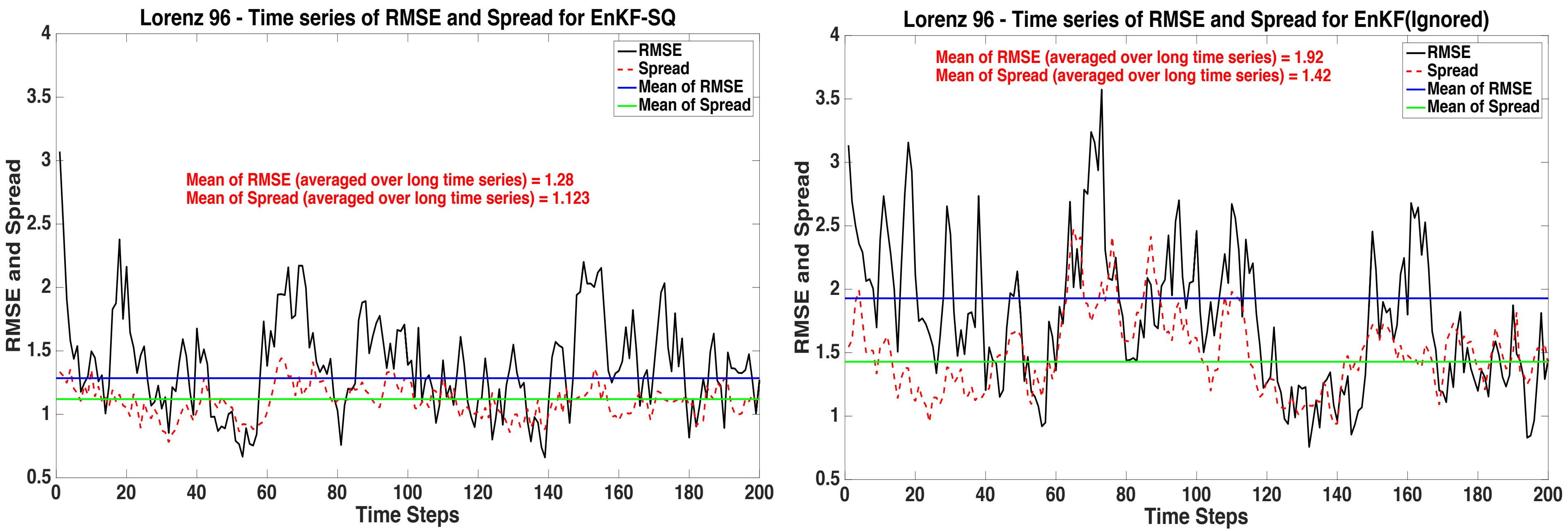
- Observations are assimilated serially as each observation is compared with detection limit, which is scalar.

Numerical Experiments

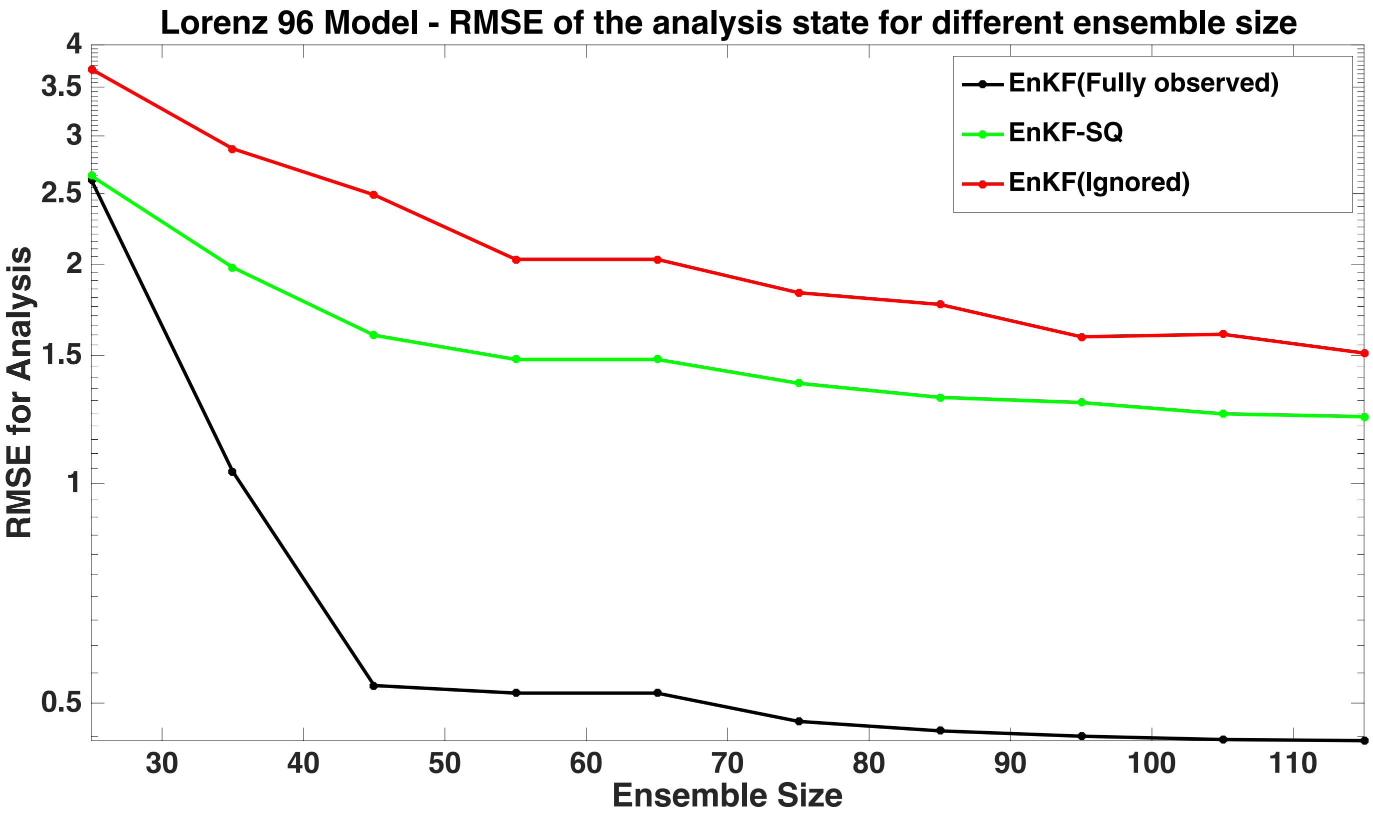
- Newly developed DA scheme is tested under the framework of twin experiments with a toy model:
  - Lorenz 96 (40 variable chaotic non-linear model)
- Sensitivity Experiments for:
  - Ensemble size
  - Observation frequency (linear to non-linear regimes)

- Detection limit for observations (more to less observable)
- Number of observation (densely to sparsely obs. network)
- Model forcing (via forcing parameter)

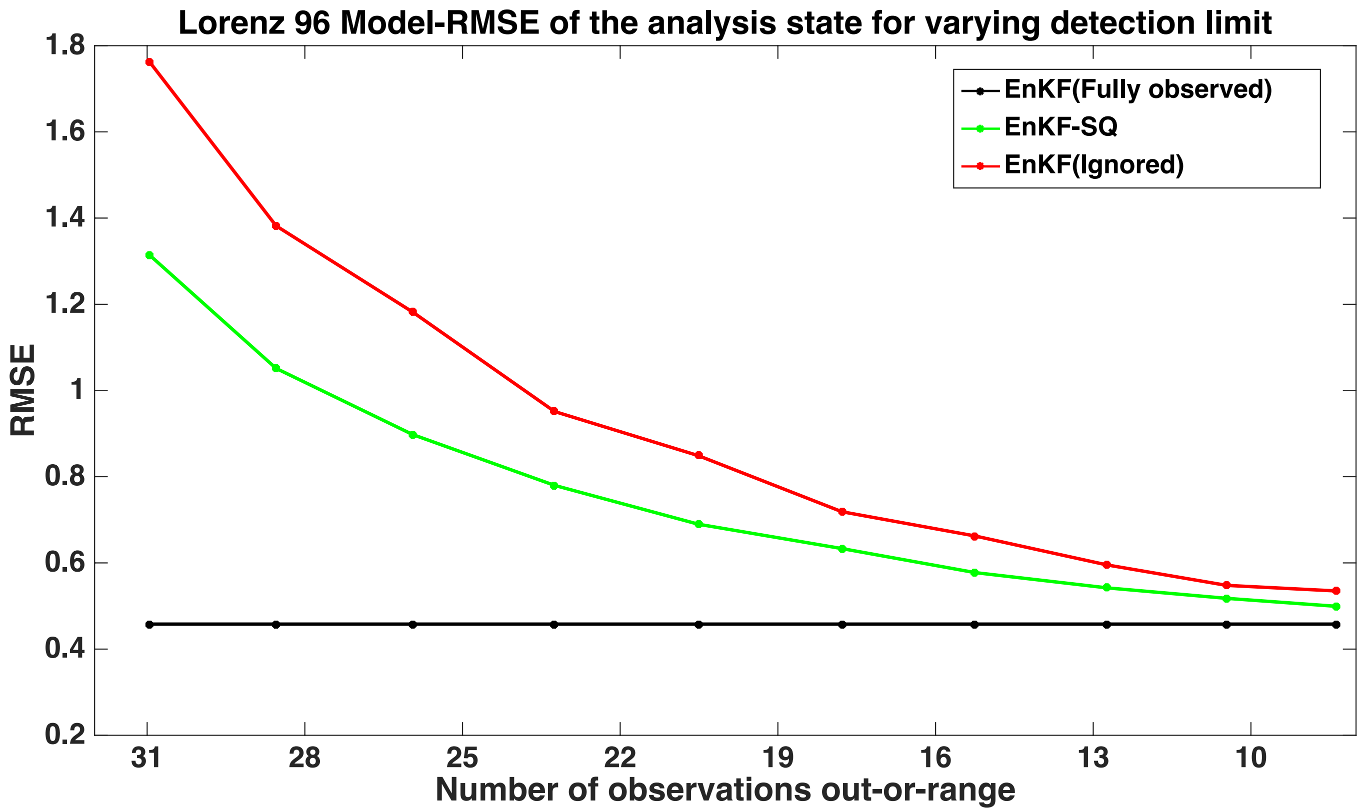
- The performance of the EnKF-SQ DA scheme is inspected with following diagnostic tools:
  - Root mean square error
  - Average Ensemble spread



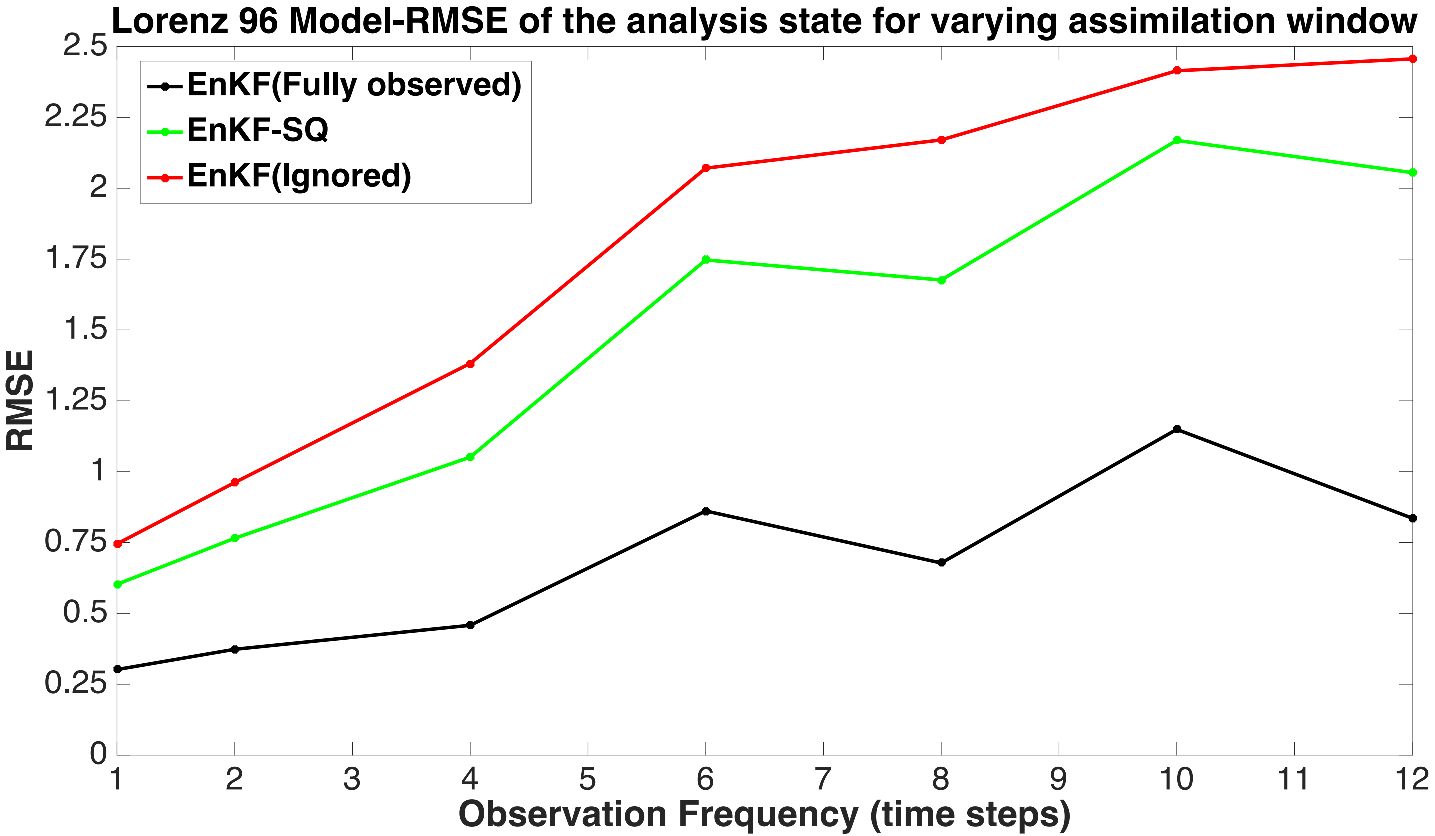
- RMSE and ensemble spread of the EnKF-SQ and EnKF (Ignored). Assimilating observations at every day and 80% of observations are out-of-range
- The RMSE and ensemble spread are more consistent in EnKF-SQ compare to the EnKF(Ignored)



- Detection limit on observations is such that 80% of observations falls out-of-range on average for total integration time. Assimilating observations at every day



- RMSE of the analysis for the varying detection limit leading to change in the number of observations out-of-range.



- RMSE of the analysis for the varying observation frequency i.e., increasing the assimilation time window

Conclusion and Future work

- Assimilating OR-observations i.e., qualitative data:
  - Improves the quality of forecast
  - Reduces uncertainty
  - Produces reliable forecasts
- Adding strong model error deteriorates the performance of the proposed DA scheme.
- Implementing it with complex and higher dimensional model.


THANK YOU FOR LISTENING
ANY QUESTIONS?