A deterministic, fully non Gaussian analysis scheme for ensemble filters: **Multivariate Rank Histogram Filter**

Sammy Metref[1], Emmanuel Cosme[2] et al.

[1] *CNRS, LEGI*; [2] *Université Joseph Fourier - Grenoble 1, LEGI*

Project supported by the region Rhône Alpes, NCAR, and the European Commission

In the context SANGOMA

WHY ?

Oceanic models are complexified
(e.g. ocean-ice coupling, ocean color [F.Garnier, MyOcean])
$\Downarrow$
Importance of non-Gaussian and nonlinear methods

$\Downarrow$
New methods must be truly understood to be well applied

$\Downarrow$
Small case benchmark is a first and crucial step

In the context SANGOMA

Small case benchmark

- **Model**

  Lorenz 96 : composed of 40 *equations and* 40 *variables.*
  Recursively defined by :

  $$\frac{dx_i}{dt} = x_{i-1}(x_{i+1} - x_{i-2}) - x_i + F,$$

  for all $i = 1, ..., n$ *and with* $x_{i-n} = x_i = x_{i+n}$
  F *being an external forcing term.*

- **Numerics**
- **Time settings**
- **Observations**

Multivariate Rank Histogram Filter

Benefits (in particular w.r.t. most particle filters):

- The analysis scheme is utterly deterministic;
- Localization is natural;
- Divergence is almost impossible for observed variables;

But:

- Is much more expensive. Though, it can take advantage of massively parallel computers;
- Remain to be thoroughly investigated and compared to other methods.

MRHF : On observed variables

Basic idea: Sequential realization method (Tarantola, 2005, Section 2.3.3):

$$p(x_1, ..., x_n|y_1) = p(x_1|y_1)p(x_2|x_1, y_1)p(x_3|x_1, x_2, y_1)p(x_4|x_1, x_2, x_3, y_1)...$$

- Correction on observed variables : RHF method (univariate) (Anderson, 2010)
  $\Rightarrow$ *Analysis on x_1 ($p(x_1|y_1)$) is performed with a non-Gaussian "Rank Histogram" scheme.*

- Correction on unobserved variables : MRHF method (multivariate)
  $\Rightarrow$ *Analysis on x_i ($p(x_i|x_1, y_1, x_2, x_3, ..., x_{i-1})$) is performed with a non-Gaussian "Rank Histogram" scheme as well.*

Remark : In (Anderson, 2010), Corrections on unobserved variables are determined with a linear regression to the corrections on observed variables.

MRHF : On observed variables

Univariate RHF

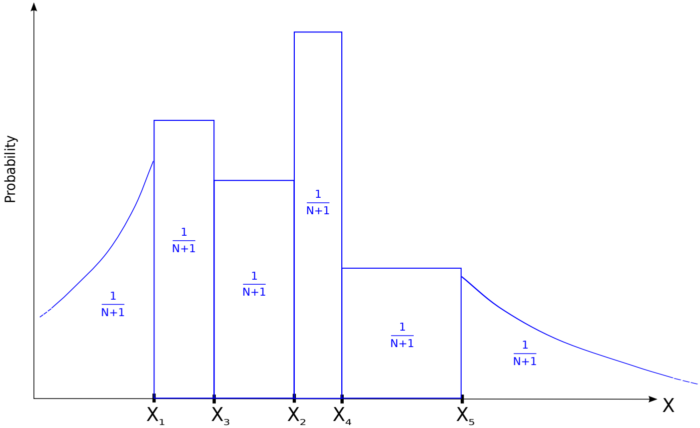
Let x_1 be an observed variable as y_1,
Bayes theorem then implies :

$$p(x_1|y_1) \propto p(y_1|x_1)p(x_1)$$

$\Rightarrow$ RHF principle : Strictly apply the latest formula with the sampled densities

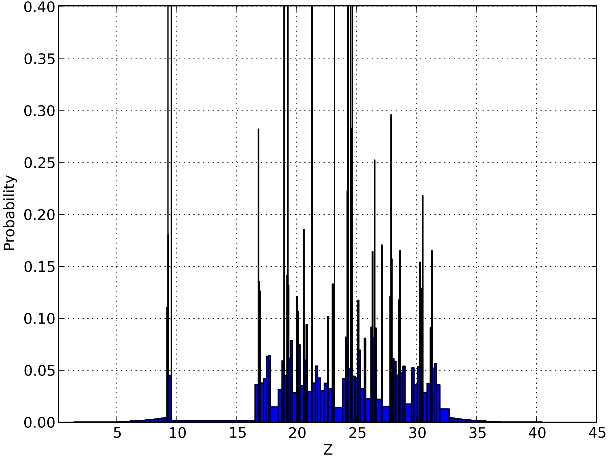
MRHF : On observed variables

Retrieving the p.d.f. $p(x_1)$ from the prior sample



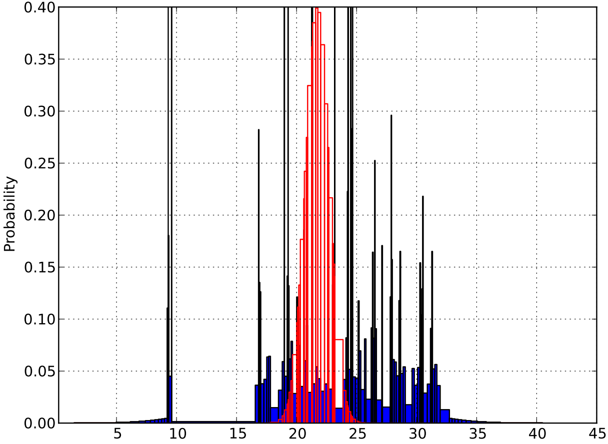
Areas in between particles sum up to 1

MRHF : On observed variables



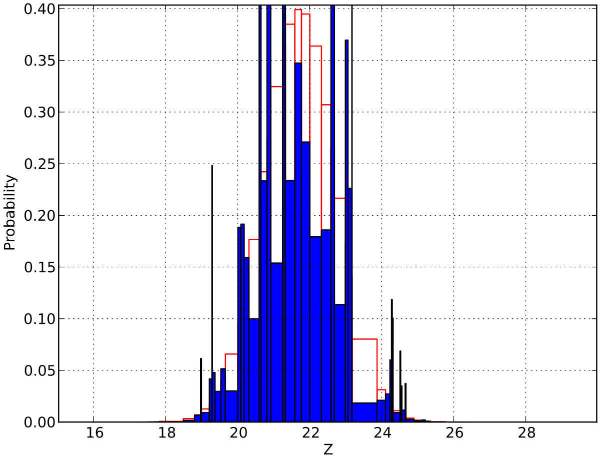
$p(x_1)$ is formed.

MRHF : On observed variables



$p(y_1|x_1)$ is discretized on the same grid.

MRHF : On observed variables



Shur product of the 2 : $p(x_1|y_1)$. Analysis particles are sampled by inversion of CDF.

MRHF : On unobserved variables

Same principle applied on unobserved variable

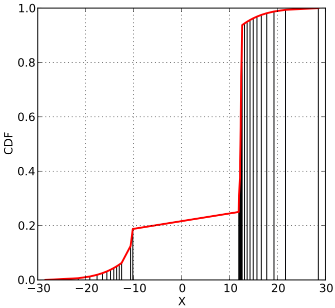
Random generation of x_i from $p(x_i | x_1 = \{x_1^a\}_{i_{part}}, x_{j \neq i})$

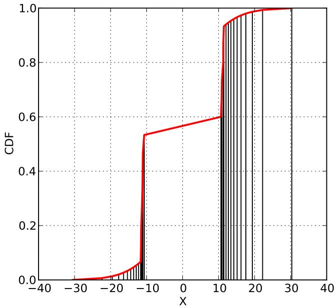
$\Rightarrow$ Pb: Leads to drastic corrections (e.g. multimodal systems)

Solution: [Other schemes in development]

Matching CDF values of $x_i | x_1 = \{x_1^b\}_{i_{part}}$ to the CDF values of $x_i | x_1 = \{x_1^a\}_{i_{part}}$

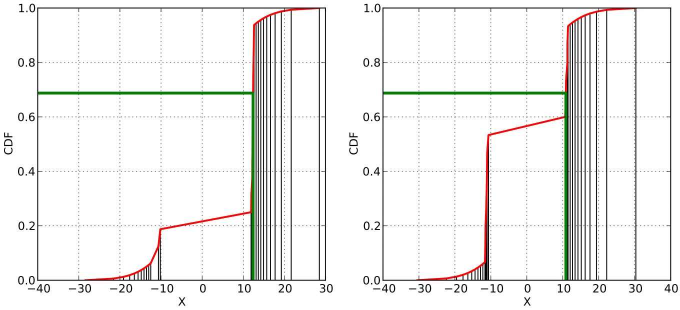
MRHF : On unobserved variables





Marginal CDF of $x_i | x_1 = \{x_1^b\}_{i_{part}}$ and $x_i | x_1 = \{x_1^a\}_{i_{part}}$ are formed.

MRHF : On unobserved variables



The analysis value for X_i^a is obtained by preserving the particle position in the marginal CDFs.

Multi-variate Rank Histogram Filter

Principle summary

- Sequential realisation method
- Deterministically based on the Bayes theory
- Use the rank histogram process to compute PDFs

  Remark: Process executed for each particle independently (parallel processing)

- Resample from the updated PDFs

Preliminary results on the small case benchmark

1. **Observations** every 2 grid points, 10 time steps, $\mathbf{R} = 2.25$ (Nakano et al., 2007);
2. **Localization**: 3 grid points, Eq. 4.10 of Gaspari and Cohn (1999). Applied to MRHF too; **Covariance inflation**, $\alpha = 1.005$;
3. The **joint PDF** decomposition,

$$p(x_1, ..., x_n|y_1) = p(x_1|y_1)p(x_2|x_1, y_1)$$

$$p(x_3|x_1, x_2, y1)p(x_4|x_1, x_2, x_3, y_1)...$$

is approximated by:

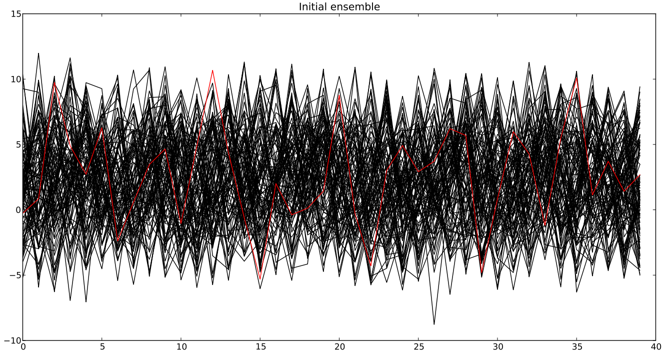
$$p(x_1, ..., x_n|y_1) \approx p(x_1|y_1)p(x_2|x_1, y_1)p(x_3|x_1, y1)p(x_4|x_1, y_1)....$$

Several reasons:

- Much less subject to sampling related problems;
- Can be parallelized (but it is not here).

Preliminary results on the small case benchmark

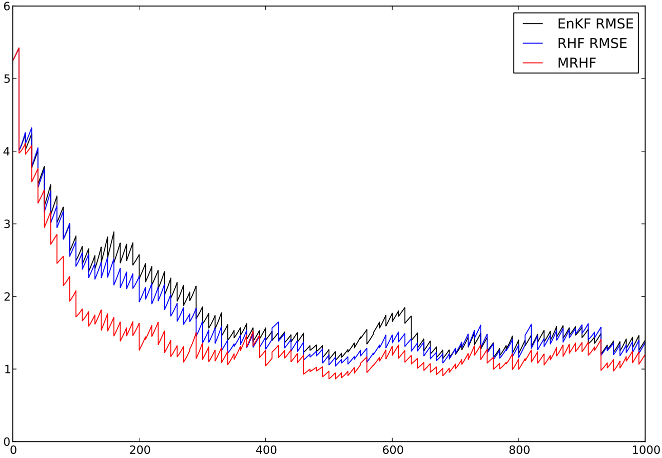
Ensemble states plot (100 particles, $TS = 0$)



— : Ensemble states — : True state

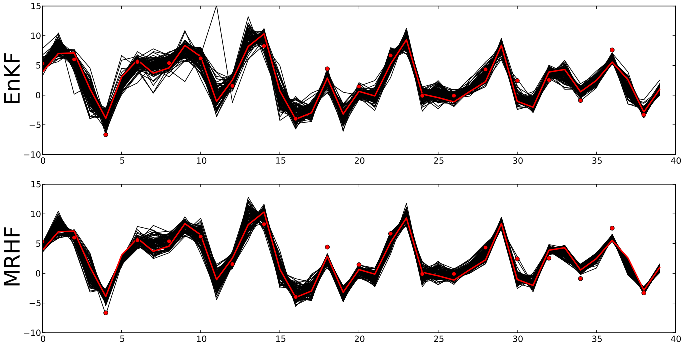
Preliminary results on the small case benchmark

RMSE plot (100 particles, $\alpha = 1.005$)



Preliminary results on the small case benchmark

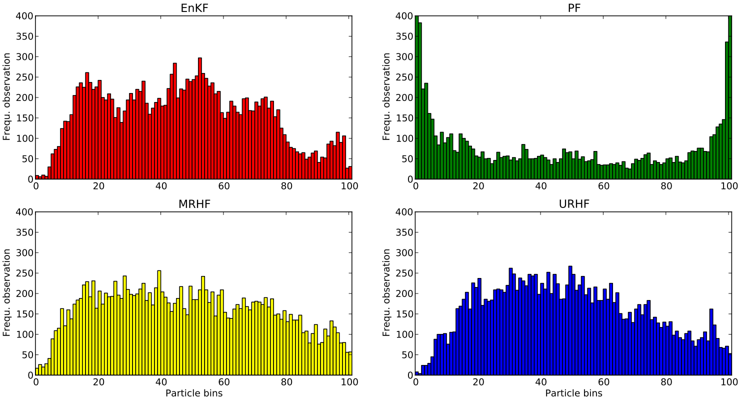
Ensemble states plot (100 particles, $\alpha = 1.005$, $TS = 1000$)



— : Ensemble states — : True state

Preliminary results on the small case benchmark

Talagrand diagram (100 particles, $\alpha = 1.005$)



Conclusions and Perspectives

What has been done ?

- This work is only starting... Conclusions are preliminary;
- The MRHF seems to perform well with L96;
- Emmanuel had promising results with L63 in a strongly non Gaussian setup

What is to be done ?

- Improve parallelization (communication between nodes);
- Implement evaluation tools (SANGOMA metrics);
- Compare with other methods (including implicit PF or PF with smart proposal).

A non exhaustive list of references

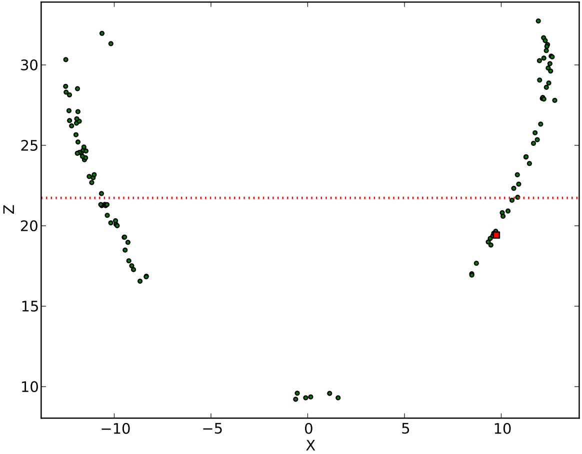
Anderson, J., 2010: A non-gaussian ensemble filter update for data assimilation. *Monthly Weather Review*, **138**, 4186–4198.

Gaspari, G. and S. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quaterly Journal of the Royal Metorological Society*, **125**, 723–757.

Nakano, S., G. Ueno, and T. Higuchi, 2007: Merging particle filter for sequential data assimilation. *Nonlin. Processes Geophys.*, **14**, 395–408.

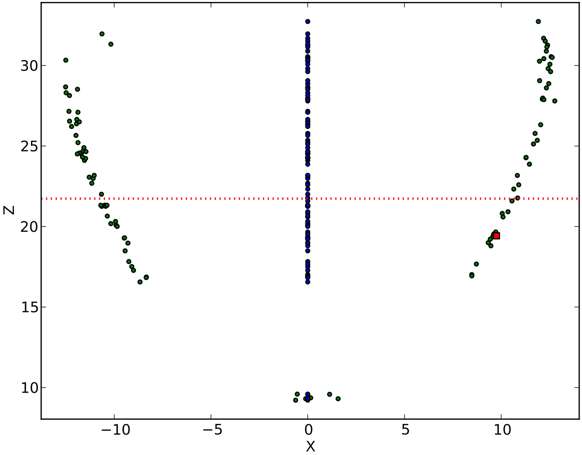
Tarantola, A., 2005: *Inverse problem theory and methods for model parameter estimation*. SIAM.

MRHF : On observed variables



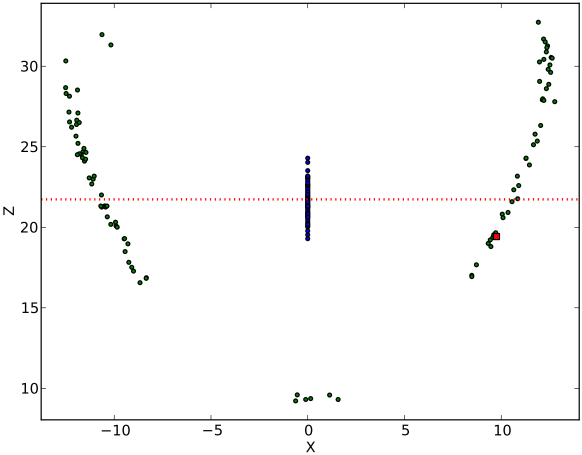
Background ensemble in $X - Z$ plane.
Red dotted line: Z obs. Red square: truth.

MRHF : On observed variables



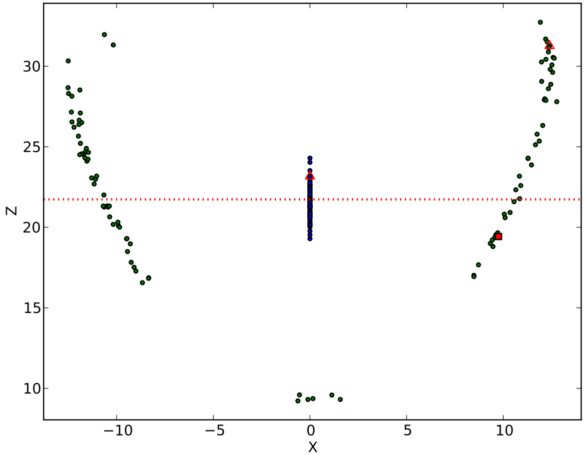
Background Z ensemble for RHF analysis.

MRHF : On observed variables



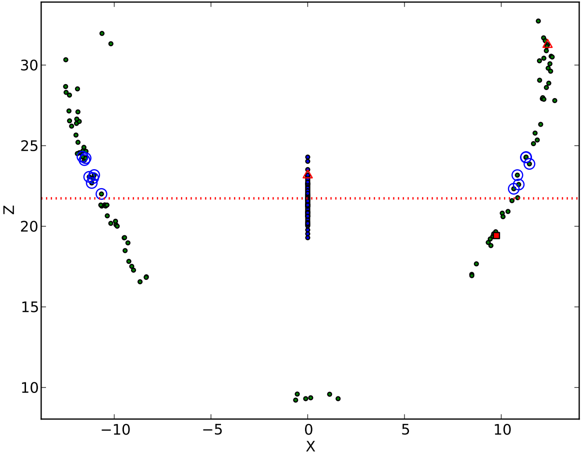
RHF analysis ensemble on Z line.

MRHF : On unobserved variables



For each particle i, an analyzed value for X must be calculated.

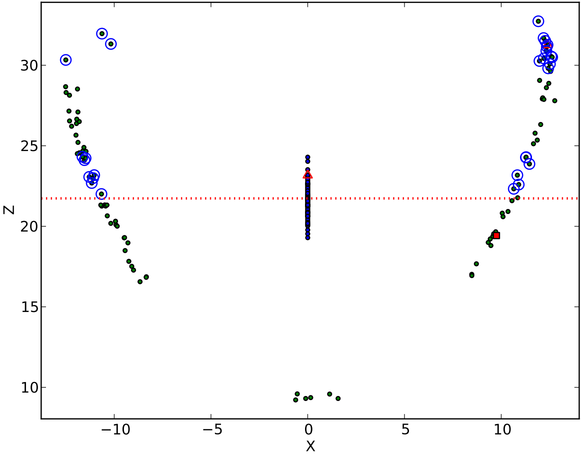
MRHF : On unobserved variables



To form $p(X|Z = Z_i^a)$, select particles in the analysis ensemble.

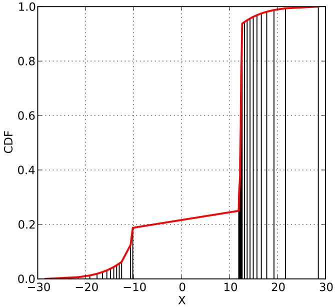
Analysis could be randomly drawn from it. (Not optimal)

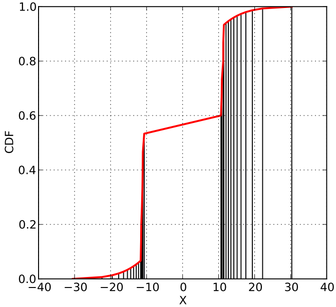
MRHF : On unobserved variables



Instead, we select particles to estimate $p(X|Z = Z_i^b)$ too.

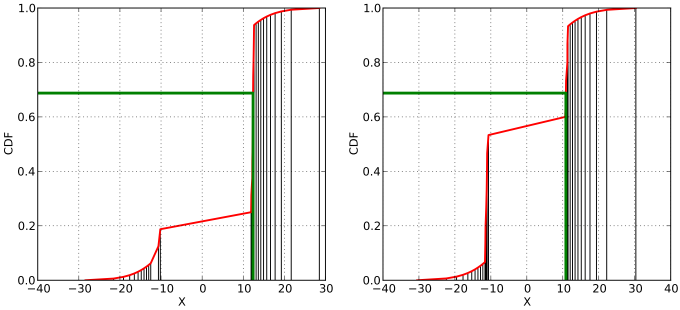
MRHF : On unobserved variables





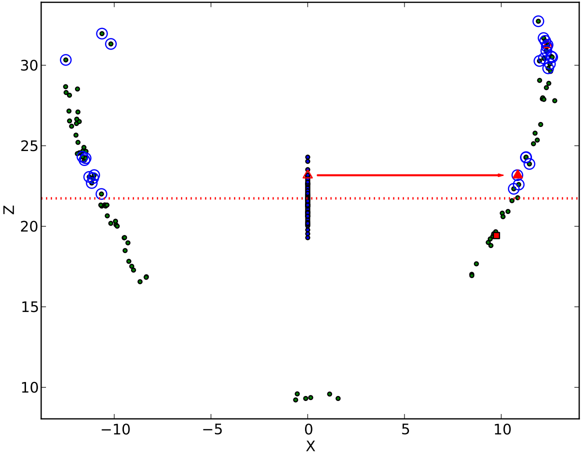
Marginal CDF of $x_i|x_1 = \{x_1^b\}_{i_{part}}$ and $x_i|x_1 = \{x_1^a\}_{i_{part}}$ are formed.

MRHF : On unobserved variables



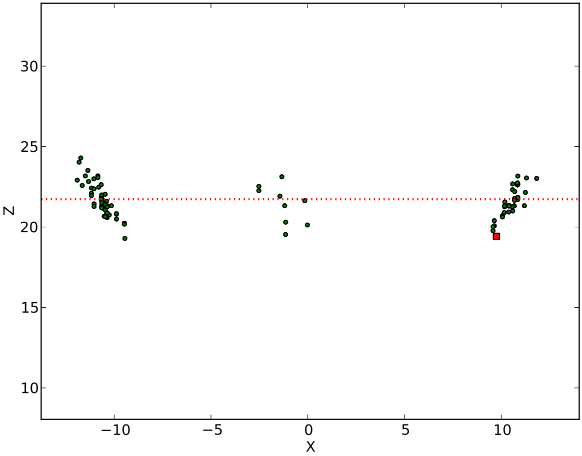
The analysis value for X_i^a is obtained by preserving the particle position in the marginal CDFs.

MRHF : On unobserved variables



This is done for each particle. Can be done in parallel.

MRHF : On unobserved variables

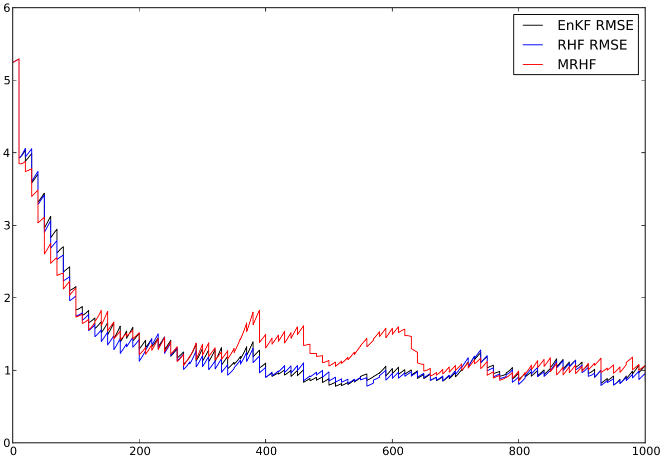


Analysis ensemble.

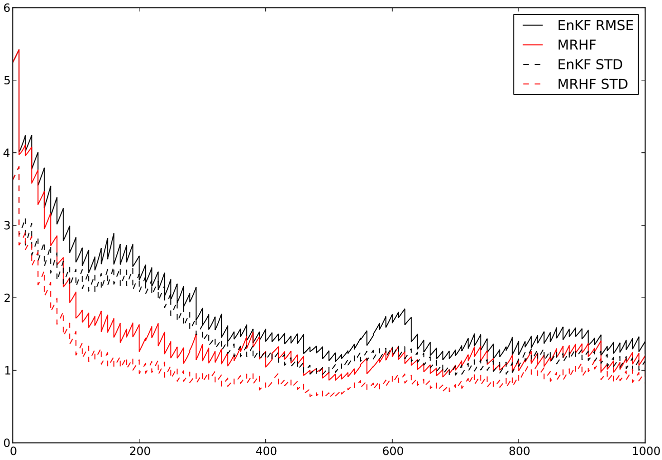
Bimodal PDF represented by 15 particles

Red PDF is sampled, and the 15 particles are used to build a RH PDF (blue).
RH PDF is not 0 between the 2 modes.

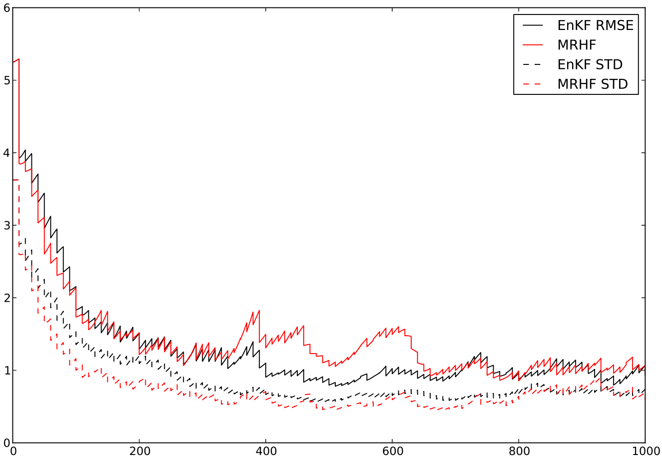
L95, 100 particles, $\alpha = 1$



L95, 100 particles, $\alpha = 1.005$



L95, 100 particles, $\alpha = 1$



L95, 100 particles, $\alpha = 1$

1000th time step