Data Assimilation in Multiscale Systems

Humberto C. Godinez, Balu Nadiga, J. David Moulton
Los Alamos National Laboratory

hgodinez@lanl.gov, balu@lanl.gov, moulton@lanl.gov

1. Motivation

Predictive modeling of complex physical systems, such as the climate system, *must capture the coupling of phenomena over an ever increasing range of spatial and temporal scales*. Data assimilation methodologies use information from the model, observational data, and corresponding error statistics to produce an improved model state (the analysis). With *a steadily increasing amount of data available for these systems*, data assimilation is poised to play a significant role in this predictive modeling.

In this work we explore two important aspects of data assimilation for these challenging problems. First, we investigate *the influence of the strength of coupling between processes at different scales:*

- study the dynamics of a model two-scale system,
- evaluate the effectiveness of the data assimilation,
- discuss the impact on data requirements.

Next we consider *the problem of assimilating large amounts of data efficiently*, and propose a new matrix-free implementation of the Ensemble Kalman Filter (EnKF),

- cost of the matrix-free EnKF algorithm scales linearly with the number of observations being assimilated
- numerical experiments show it is significantly faster than the popular EnKF algorithm developed by Evensen [4].

2. Ensemble Data Assimilation for Strongly Coupled Systems

2.1 Two-Scale System

The Lorenz 96 two-scale model ([5]) is defined by the following system of ordinary differential equations

$$\frac{\mathrm{d}X_k}{\mathrm{d}t} = -X_{k-1}(X_{k-2}-X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k} \quad k=1,\dots,K$$

$$\frac{\mathrm{d}Y_{j,k}}{\mathrm{d}t} = -cbY_{j+1,k}(Y_{j+2,k}-Y_{j-1,k}) - cY_{j,k} + \frac{hc}{b}X_k \quad j=1,\dots,J$$

The model is a coupling of two systems, each of which obeys a scaled variant of the Lorenz 96 single system model.

- There are K slow variables X_k, which are defined on a cyclic chain such that $X_{k-K} = X_{k+K} = X_k$.
- For each slow variable X_k there are J fast variables $Y_{j,k}$ associated with X_k. The fast variables $Y_{j,k}, k=1$ are cyclic within k, so that $Y_{j,k-K} = Y_{j,k+K} = Y_{j,k}$, while in j we have $Y_{j-J,k} = Y_{j,k-1}$ and $Y_{j+J,k} = Y_{j,k+1}$.

The systems was designed to represent the dynamics of the coupling between two distinct variables.
The constant F is a forcing term, b determines the ratio of amplitude of slow and fast variables, c determines the decorrelation timescales between fast and slow, and the coupling coefficient h determines how strongly the slow variables drive the fast variables.

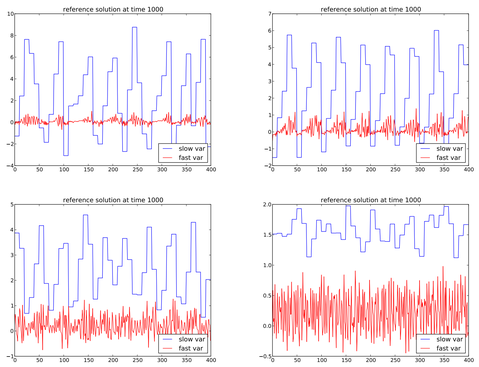

Figure 1: *Solution of Lorenz two-scale model at the final time for $h=1.0$ (top left), $h=2.0$ (top right), $h=3.0$ (bottom left), and $h=4.0$ (bottom right). The blue line for slow variables X_k, red line is fast variables $Y_{j,k}$. The figures show how the correlation of the signal between the slow and fast variables, which is strong when $h=1.0$, degrades as h becomes larger.*

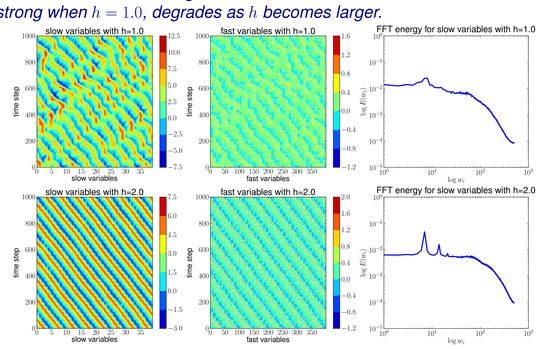

Figure 2: *Time distribution of the slow variables (left figures), fast variables (middle figures), and average FFT energy of the slow variables (left figures) of the Lorenz two-scale model for $h=1.0$ (top) and $h=2.0$ (bottom).*

For our numerical experiments we have the following setup:

- We have $K=40$ slow variables X_k, where each slow variable is linked to $J=10$ fast variables $Y_{j,k}$. This makes a total of 440 variables for the system.
- The forcing term is taken as $F=10$, since this is well into the chaotic regime, $b=10$, and $c=10$.
- The coupling coefficient h is varied from $h=1.0$ to $h=4.0$.
- A 4th order Runge-Kutta method was used for the time stepping, where $\delta t = 0.05$. The time units were defined by Lorenz so that 0.05 units correspond to 6 hours.
- The initial condition is obtained from a 60 day simulation initialized with a random state.

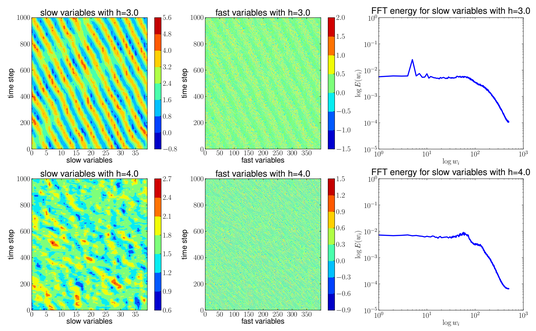

Figure 3: *Same as Figure 2 for $h=3.0$ (top) and $h=4.0$ (bottom).*

Figure 2 and 3 show the time distributed solution of the slow variables (left figures), fast variables (middle figures), and the average FFT energy of the slow variables (left figures). For $h=1.0$ there exist a oscillatory structure in the solution, which will be reflected in the correlation between variables, this structure is enhanced for $h=2.0$. For $h=3.0$ the structure is still clearly visible, and for $h=4.0$ the structure of the solution disappears.

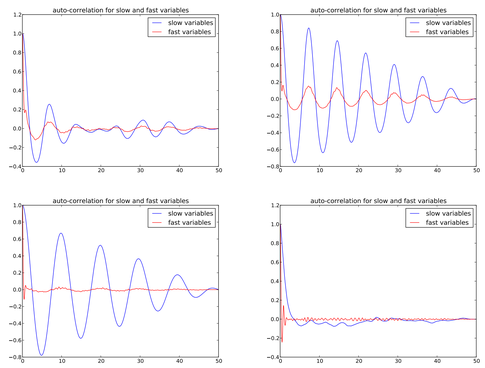

Figure 4: *Auto-correlation of the first slow variable with the rest of the variables for $h=1.0$ (top left), $h=2.0$ (top right), $h=3.0$ (bottom left), and $h=4.0$ (bottom right).*

2.2 Assimilation Experiments with Ensemble Transform Kalman Filter

The ensemble transform Kalman filter (ETKF) was introduced by Bishop *et.al.* [1] as an adaptive sampling method. It computes the correct ensemble perturbations consistent with the error covariance matrix of the analysis. More details can be found in [1].
The following setup is used in the ETKF assimilation experiments:

- The assimilation was done within a twin-experiment framework, where observations are taken from a reference run.
- An ensemble of 500 members was simulated for assimilation
- Each ensemble member is initialized by perturbing the initial condition of the reference run with white Gaussian noise.
- The data assimilation is performed for a time window of 25 days, with assimilation done every 6 hours.
- At every assimilation time 20 observations from the slow variables X_k are considered for each assimilation experiment.

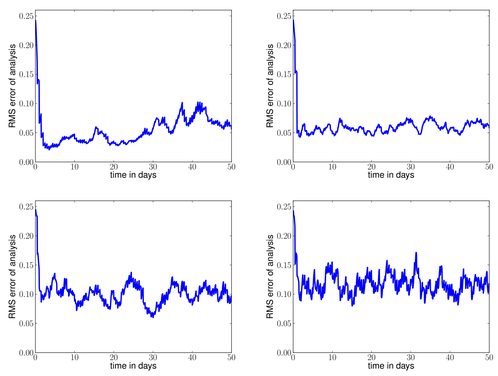

Figure 5: *Error measure of the ETKF assimilation for the slow variables with 20 observations for $h=1.0$ (top left), $h=2.0$ (top right), $h=3.0$ (bottom left), and $h=4.0$ (bottom right).*

The root mean square error measure (reference minus analysis) is shown in Figure 5 for slow variables. The assimilation for $h=1.0$ does a reasonable job, while with $h=2.0$ the assimilation does a very good job at estimating the reference solution. For $h=3.0$ the error measure shows some oscillatory behavior, where the assimilation appears to estimate the reference solution reasonably well over some intervals, but is not always consistent. Finally, for $h=4.0$ we see that the assimilation degrades even further, and the reference solution is approximated with a higher degree of error than in previous cases. The degradation of the assimilation might be caused by the diminishing correlation among the variables of the system, as shown in Figure 4.

2.3 Discussion

- Strong coupling between variables of different time-scales can adversely affect the accuracy of data assimilation methods.
- The Lorenz two-scale model was chosen to study the effects of strong coupling in data assimilation.
- It was observed that the structure of the solution can break down with strong coupling (Figures 2 3). Additionally, the correlation among the variables is diminished, as can be seen in Figure 4.
- An ensemble-based data assimilation was performed for a strongly coupled system, where it was found that the assimilation encounters difficulty in converging to the correct model state (Figure 5).
- The degradation of the assimilation might be attributed to the diminish correlation among the variables of the system.
- In a different experiment (not shown), all slow variables are observed and assimilated. In this case the assimilation is able to recover the reference state, but the uncertainty of the assimilated state remains high.

3. An Efficient Matrix-Free EnKF Algorithm

In the second part of this work we describe an efficient matrix-free implementation of the Ensemble Kalman Filter (EnKF) for the assimilation of large data sets. We propose a matrix-free implementation of the EnKF where a Sherman-Morrison solver that exploits the outer-product form of the matrix is used. The cost of the resulting algorithm scales linearly in the number of observations (data) being assimilated, as shown in the numerical results.

3.1 Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) was introduced by Evensen [3] as a Monte Carlo approximation to Kalman filtering, for non-linear models, and has gained wide acceptance in data assimilation applications. The EnKF uses an ensemble of model states to estimate the model covariance matrix, necessary to compute the analysis in the Kalman filter equations.
The following notation will be used throughout this part:

- n: dimension of model state variable
- m: number of observations
- N: number of ensemble members

For a vector of m measurements $\mathbf{y}^o \in \mathbb{R}^m$ and an ensemble of N forecast $\mathbf{x}_i^f \in \mathbb{R}^n, i=1,\dots,N$ the EnKF analysis equation are given by:

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}\left(\mathbf{y}_i^o - \mathbf{H}\mathbf{x}_i^f\right), \quad i=1,\dots,N \tag{1}$$

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{2}$$

where the matrix $\mathbf{K} \in \mathbb{R}^{n\times m}$ is referred to as the Kalman gain matrix; $\mathbf{P}^f \in \mathbb{R}^{n\times n}$ is the forecast covariance matrix; $\mathbf{R} \in \mathbb{R}^{m\times m}$ is the observations covariance matrix; $\mathbf{H} \in \mathbb{R}^{m\times n}$ is the linear observation operator. The vector $\mathbf{y}_i^o \in \mathbb{R}^m$ is a perturbed observations vector.
In the EnKF the forecast error covariance matrix is obtained through the ensemble of model forecast, using the relation

$$\mathbf{P}^f = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{x}_i^f - \overline{\mathbf{x}}^f\right)\left(\mathbf{x}_i^f - \overline{\mathbf{x}}^f\right)^T, \tag{3}$$

where $\overline{\mathbf{x}}^f$ is the forecast ensemble average.

- Typical techniques use SVD or Cholesky decomposition to deal with the inverse matrix in $\mathbf{K}$.
- Evensen [4] developed an efficient EnKF algorithm based on SVD. Scales linearly with m, nevertheless can become expensive for large m due to the matrix multiplication operations.
- Viable solution is to perform assimilation matrix-free.

3.2 A Matrix-Free Algorithm

Denoting the innovation matrix by

$$\mathbf{C} = \left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right), \tag{4}$$

the analysis step (1)-(2) in EnKF can be rewritten as

$$\mathbf{C}\mathbf{z}_i = \left(\mathbf{y}_i^o - \mathbf{H}\mathbf{x}_i^f\right) \tag{5}$$

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{P}^f\mathbf{H}^T\mathbf{z}_i \tag{6}$$

Notice that using (3), the innovation matrix $\mathbf{C}$ can be expressed as a sum of rank one perturbations to a non-singular matrix. Edigi and Maponni [2] developed an efficient matrix-free linear solver that exploits the particular form of $\mathbf{C}$, based on the Sherman-Morrison identity

$$\left(\mathbf{A} + \mathbf{u}\mathbf{v}^T\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}. \tag{7}$$

Hence the complete Matrix-free assimilation consists of two main steps:

1. Solve the linear system (5) involving the innovation matrix for each ensemble member using the matrix-free Sherman-Morrison linear solver.
2. Update the forecast with equation (6) to obtain the analysis.

The cost of the resulting matrix-free EnKF algorithm is linear in m and more efficient than a matrix decomposition approach since it avoids all matrix operations that are computationally expensive.

3.3 Assimilation Experiments and Results

A global 2D shallow water (SW) model on a sphere is used for the numerical experiments. Equations discretized using finite volume method type method, with a time step $\Delta t = 450$s.

- Model describes hydrodynamic flow on a sphere assuming vertical motion is much smaller than horizontal motion.
- Assume fluid depth is small compared with radius of the sphere (radius of Earth).
- Computations done on a $2.5^\circ \times 2.5^\circ$ grid.
- $\mathbf{x}_0^t$: trajectory produced by SW integration with an initial fluid depth defined by

$$h_0(\lambda,\theta) = \frac{1}{g}\left(\bar{\Phi} + 2\Omega a\vartheta\sin^3(\theta)\cos(\theta)\sin(\lambda)\right),$$

  and initial velocities u_0, v_0 derived from the geostrophic relations.
- Initial condition for the control run $\mathbf{x}_0^b$ is taken from shifting $\mathbf{x}_0^t$ one grid point to the left.

Comparison of timing between our Matrix-Free EnKF and an SVD based implementation from Evensen 2003.

- Ensemble of size $N=100$ model simulations was used, integrated up to 24h.
- Ensemble I.C. generated by perturbing $\mathbf{x}_0^b$ with a random vector sampled from a normal distribution with mean zero.
- A single assimilation is done at 6h or time step 48.
- Varied the number of observations to assimilate from 200 to 3×10^7.
- Overhead of both methods are compared, as well as subsequent operations.

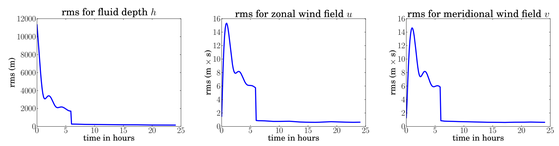

Figure 6: *Time distribution of the root mean square (rms) difference between the reference solution and the ensemble average with assimilation. A single EnKF assimilation of $m=30600$ observations was performed at $t=6$ hours (time step 48). The left figure show the rms error for the fluid depth (h), the middle for the zonal wind velocity field (u), and the left for the meridional wind velocity field (v).*

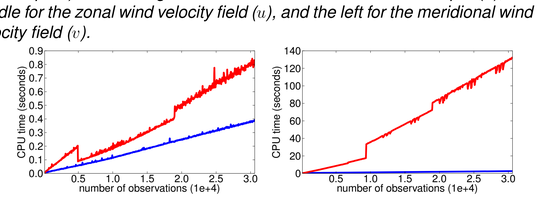

Figure 7: *Left figure: timing performance of the overhead or first part the matrix-free implementation (blue line) and the SVD-based implementation (red line). Right figure: Timing performance of the second part of the Matrix-free implementation (blue line) and the SVD-based implementation (red line). The CPU time is shown as a function of m, the number of observations being assimilated, at $t=6$ hours. Timing is in CPU seconds.*

- The matrix-free EnKF algorithm shows a clear advantage in computational performance against the SVD-based EnKF algorithm.
- The efficiency presented in the experiments is directly related to the number of observations m being assimilated by each algorithm. That is, the matrix-free implementation can efficiently assimilate a large number of observations in a fraction of the time of the SVD-based implementation.

4. Conclusions and Future Work

- The accuracy of data assimilation may be adversely affected by strong coupling in a multi-scale system. This may be due to the loss of structure in the solution, as well as the diminishing correlation among variables of the system.
- An efficient matrix-free EnKF algorithm was presented, where the algorithm compared favorably in computational performance against an SVD-based EnKF algorithm.
- A Parallel version of the matrix-free EnKF algorithm has been implemented and preliminary results indicate a good scaling for the algorithm.
- Future work for the matrix-free algorithm lies in developing an appropriate inflation and localization techniques that take advantage of the computational efficiency of the algorithm.

5. Acknowledgements

This research was supported by the Department of Energy at Los Alamos National Laboratory under contracts DE-AC52-06NA25396 and the DOE Office of Science Advanced Computing Research (ASCR) program in Applied Mathematical Sciences. LA-UR-11-11694.

References

[1] C.H. Bishop, B.J. Etherton, and S.J. Majumdar, *Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects*, Mon. Wea. Rev. **129** (2001), no. 3, 420–436.

[2] N. Egidi and P. Maponi, *A Sherman-Morrison approach to the solution of linear systems*, J. Comput. Appl. Math. **189** (2006), 703–718.

[3] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, J. Geophys. Res. **99** (1994), no. C5, 10143–10162.

[4] _____, *The ensemble Kalman filter: Theoretical formulation and practical implementation*, Ocean dynamics **53** (2003), 343–367.

[5] E.N. Lorenz, *Predictability: A problem partly solved*, Proceedings on predictability, held at ECMWF on 4–8 September 1995.